МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (государственный университет)
ФАКУЛЬТЕТ УПРАВЛЕНИЯ И ПРИКЛАДНОЙ МАТЕМАТИКИ
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А.А. ДОРОДНИЦЫНА РАН
КАФЕДРА «ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ»

Левдик Павел Владимирович

Неявная кросс-валидация для выбора подмножества информативных признаков в задаче обучения распознаванию образов по методу опорных векторов.

511656 — Математические и информационные технологии

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА

Научный руководитель:
ведущий научный сотрудник ВЦ РАН
д.т.н., профессор В. В. Моттль

Москва 2015 г.

Аннотация

Широким классом задач машинного обучения является восстановление зависимости по эмпирическим данным, среди которых наиболее типичной является задача классификации. Наиболее популярным способом обучения распознаванию образов является метод опорных векторов (SVM). Как правило, для упрощения модели требуется провести отбор признаков с целью оставить только самые информативные. Этого добиваются, в частности, путем включения в обучающий критерий SVM, наряду с традиционным квадратичным регуляризующим членом, еще и модульного регуляризатора с некоторым дополнительным коэффициентом. Для того, чтобы подобрать нужные значения теперь уже двух структурных параметров, отвечающих за соответствие рассматриваемой модели данных предположениям наблюдателя, обычно используют многократное повторение процесса обучения посредством кросс-валидации. Чтобы избежать перебора, в данной диссертации мы предлагаем метод математической имитации кросс-валидации, используя естественные допущения о природе неизвестного вероятностного распределения, породившего обучающую выборку. Мы назвали этот подход неявной кросс-валидацией. Таким образом, многократное повторение процесса обучения заменяется вычислением единственного критерия, зависящего только от исходных данных и параметров регуляризации.

В отличие от обычной кросс-валидации, требующей многократного повторения процесса обучения при разных разбиениях обучающей совокупности, метод неявной кросс-валидации предполагает однократное вычисление критерия, зависящего только от исходных данных и варьируемых параметров регуляризации.

В данной диссертации экспериментально сравниваются методы обычной и неявной кросс-валидации в задаче распознавания образов на основе представления объекта вектором вторичных признаков, образованных сравнением его сходства с каждым из объектов обучающей совокупности. Отбор подмножества релевантных обучающих объектов, с которыми достаточно сравнивать предъявленный объект, обеспечивает критерий обучения Elastic Net SVM, содержащий структурный параметр, определяющий степень селективности отбора. В качестве источника экспериментальных данных использован известный массив «Chicken Pieces Silhouettes Database», содержащий 446 силуэтных изображений частей курицы (крылышко, спинка, ножка, бедрышко, грудка), представленных 44 разными функциями парного сравнения изображений. Оказалось, что неявная кросс-валидация выделяет всего несколько релевантных объектов, достаточных для решающего правила, в то время как обычный скользящий контроль в данной задаче оказался нечувствительным к селективности отбора.

Содержание

1 Введение

Задача восстановления объективно существующей зависимости между наблюдаемыми свойствами объектов реального мира и их некоторой скрытой характеристикой, доступной для наблюдения лишь в пределах конечной обучающей совокупности, является центральной задачей интеллектуального анализа данных. В частности, если скрытая характеристика принимает значения из конечного неупорядоченного множества, то такую задачу принято называть задачей обучения распознаванию образов, а в случае числовой скрытой характеристики говорят о задаче восстановления регрессионной зависимости. В настоящей диссертации рассматривается первая из этих задач.

Наиболее общее понимание задачи обучения основано на предположении, что с каждым объектом из некоторого множества объектов реального мира, привлекающего внимание наблюдателя, объективно связаны значения двух его характеристик, одна из которых доступна для непосредственного наблюдения, а другая скрыта. В теории обучения обычно предполагается, что природа, случайным образом выбирая один объект, генерирует, тем самым, случайную пару значений его наблюдаемой и скрытой характеристик, причем соответствующее совместное распределение вероятностей объективно существует, но наблюдателю неизвестно. Наблюдатель всякий раз видит значение лишь наблюдаемой характеристики, в то время как природа требует, чтобы он «угадывал» значение скрытой характеристики объекта, штрафуя неправильное оценивание в соответствии с известной функцией потерь. Наблюдатель вынужден выработать решающее правило, связывающее с каждым наблюденным значением доступной характеристики объекта предполагаемое значение его скрытой характеристики. Как правило, наблюдатель формирует свое решающее правило на основе предположения о некотором параметрическом классе зависимостей, так что выбор наблюдателем конкретного решающего правила полностью задается выбором параметра. Естественной объективной оценкой «качества» решающего правила является математическое ожидание потерь, которое в теории обучения принято называть средним риском ошибки. Очевидно, что наблюдатель, выбирая вариант решающего правила, а именно значение определяющего его параметра, не может вычислить средний риск ошибки, поскольку совместное распределение наблюдаемой и скрытой характеристик случайно появляющегося объекта ему неизвестно.

Единственную объективную информацию о свойствах природы, доступную наблюдателю, несет обучающая совокупность, под которой понимается конечное множество пар значений как наблюдаемой, так и скрытой характеристики объектов, случайно выбранных природой в соответствии с объективно существующим распределением вероятностей. Выбирая решающее правило, наблюдатель может лишь вычислить для всякого его варианта среднее арифметическое значение функции потерь, называемое эмпирическим риском ошибки. При выборе решающего правила общепринятым соображением, основанным на данных, является минимизация эмпирического риска в некотором классе решающих правил.

Однако параметрический класс решающих правил, изначально принятый наблюдателем, может оказаться слишком широким для ограниченного объема обучающей совокупности, и средний риск ошибки результата обучения по критерию минимума эмпирического риска может оказаться неприемлемо большим. Такое явление принято называть переобучением. Другим общепринятым соображением, направленным на уменьшение опасности переобучения, является использование априорной (регуляризующей) информации об «ожидаемом» решающем правиле восстановления зависимости. Другими словами, наблюдатель пытается сузить параметрический класс зависимостей, накладывая на параметр априорные регуляризирующие требования, в свою очередь контролируемые структурным параметром. Как правило, априорная информация выражена в виде некоторого функционала на классе решающих правил, подлежащего минимизации, причем обычно такой функционал содержит параметр, контролирующий отклонение решающего правила от некоторого подмножества наиболее «простых» правил, и называемый структурным параметром «сложности» класса решающих правил.

В современной теории обучения эти два соображения объединяются в единый критерий обучения, получая тем самым регуляризованный критерий минимизации эмпирического риска. Естественно, что результат обучения, т.е. решающее правило, получаемое в качестве решения задачи минимизации, зависит от структурного параметра, отвечающего за сложность зависимости между ненаблюдаемой и наблюдаемой компонентами объекта.

Очевидным показателем «качества» выбора структурных параметров и, следовательно, получаемого решающего правила, является средний риск ошибки оценивания скрытой характеристики нового случайного объекта, не входящего в обучающую совокупность. Однако, вычисление среднего риска принципиально невозможно, поскольку наблюдателю неизвестно совместное распределение вероятностей на множестве пар значений наблюдаемой и скрытой характеристик объектов в генеральной совокупности. В качестве общепринятого компромисса на практике обычно заменяют критерий минимума среднего риска ошибки при выборе структурного параметра на его суррогат, вычисленный путем кросс-валидации единственной обучающей совокупности, доступной наблюдателю. Метод кросс-валидации (Cross-Validation) [1] заключается в том, что обучающая совокупность многократно разбивается на две части, по одной из которых определяется решающее правило для каждого пробного значения структурного параметра, а по другой оценивается среднее значение ошибки.

Проблемная ситуация заключается в том, что методы кросс-валидации требуют многократного повторения обучения при разных разбиениях обучающей совокупности, что определяет их чрезвычайно высокую вычислительную сложность. В частности, наиболее популярными видами кросс-валидации являются блочная кросс-валидация, заключающаяся в разбиении обучающей совокупности на достаточно большое число частей и поочередном использовании каждой части в качестве контрольной при обучении по остальным частям (K-fold Cross-Validation), и скользящий контроль [2], в котором поочередно выделяется один объект в качестве контрольного, а обучение проводится по оставшимся объектам (Leave-one-out Cross-Validation). При этом число повторений обучения равно кратности разбиения обучающей совокупности на блоки, а в методе скользящего контроля совпадает с числом объектов в обучающей совокупности.

Для разрешения этой проблемной ситуации в кандидатской диссертации Е.О. Черноусовой [3] был предложен общий метод, основанный на некотором предположении наблюдателя о возможном параметрическом классе совместных распределений наблюдаемой и скрытой характеристик случайно появляющегося объекта, и назван в диссертации методом неявной кросс-валидации. Метод основан на мысленном эксперименте, заключающемся в получении двух независимых выборок, по первой из которых находится решающее правило как по обучающей совокупности, а на второй измеряется эмпирический риск ошибки восстановления скрытой характеристики объекта. В качестве критерия выбора значений структурных параметров предложено использовать математическое ожидание эмпирического риска ошибки.

В отличие от обычной кросс-валидации, требующей многократного повторения процесса обучения при разных разбиениях обучающей совокупности, метод неявной кросс-валидации предполагает однократное вычисление критерия, зависящего только от исходных данных и варьируемых параметров регуляризации.

В диссертации Е.О. Черноусовой [3] рассмотрена только задача оценивания числовой (регрессионной) зависимости. Доказано, что в случае квадратичной функции потерь, адекватной широкому классу задач построения регрессионных моделей данных, и квадратичного регуляризующего штрафа, налагаемого на вектор искомых коэффициентов регрессии, несмещенная оценка математического ожидания эмпирического риска ошибки выражается через элементы обучающей совокупности в виде простой формулы. Показано, что частным случаем такого критерия выбора структурных параметров при некоторых специальных предположениях о модели данных является известный информационный критерий Акаике [4].

Целью настоящей диссертации является обобщение принципа неявной кросс валидации, разработанного Е.О. Черноусовой для регрессионных моделей данных, на класс задач обучения распознаванию образов по методу опорных векторов (SVM – Support Vector Machine) [5,6].

В основе метода опорных векторов, разработанного В.Н. Вапником и А.Я. Червоненкисом [6], в основе которого лежит ими же ранее предложенный метод обобщенного портрета [7]. Предполагается, что всякий объект воспринимается через вектор его числовых признаков как точка в линейном пространстве $\mathbf{x}(\omega) \in \mathbb{R}^n$, размерность которого определяется числом признаков n. Предполагается также, что для суждения о принадлежности объекта к одному из двух классов $y = \pm 1$ достаточно вычислить значение линейной решающей функции (decision function) $d(\mathbf{x} \mid \mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b\colon \mathbb{R}^n \to \mathbb{R}$, знак которой непосредственно укажет класс объекта $\mathbf{a}^T\mathbf{x} + b \gtrless 0$. Очевидно, что линейная функция $d(\mathbf{x} \mid \mathbf{a}, b)$ определяет дискриминантную гиперплоскость в $\mathbb{R}^n$, а решение о классе объекта определяется тем, по какую сторону от нее окажется вектор признаков объекта $\mathbf{x}$. Обучение линейного классификатора сводится к формированию значений параметров $(\mathbf{a}, b)$ на основе анализа конечной обучающей совокупности $\{(\mathbf{x}_j, y_j),\ j = 1, \ldots, N\}$. В основе метода лежит естественная идея выбирать ту гиперплоскость, которая в обучающей совокупности разделяет векторы признаков объектов разных классов с наибольшим зазором, дополнительно штрафуя возможные нарушения некоторыми объектами этого общего «идеального» требования. В данной диссертации исследуется и развивается именно метод опорных векторов.

При всей эффективности метод опорных векторов (SVM) остается эвристическим по своей конструкции. С момента его создания в мировой литературе был предпринят ряд попыток снабдить его некоторой вероятностной интерпретацией [8,9]. Однако эти интерпретации оставались неполными, не позволяющими в полной мере использовать вероятностный аппарат для наделения чрезвычайно популярного метода опорных векторов принципиально новыми свойствами. Чтобы сугубо вероятностный принцип неявной кросс валидации Е.О. Черноусовой был применим к методу SVM, мы используем его вероятностную интерпретацию, разработанную А.И. Татарчуком [10].

В диссертации А.И. Татарчука предложена специальная байесовская постановка обучения распознаванию двух классов объектов в линейном признаковом пространстве, приводящая к обобщению метода опорных векторов и являющаяся теоретической основой для создания новых селективных методов обучения. Основная идея байесовской постановки заключается в построении системы вероятностных предположений о паре плотностей распределения объектов двух классов $\varphi(\mathbf{x} | y = \pm 1, \mathbf{a}, b)$, определяемой объективно существующей, но неизвестной гиперплоскостью $(\mathbf{a}, b)$ в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе $\Psi(\mathbf{a}, b)$. При этом именно структура первого семейства распределений определяет характерный принцип обучения по методу опорных векторов.

Что же касается априорного распределения параметров дискриминантной гиперплоскости, а именно ее направляющего вектора, то одна из основных идей диссертации заключается в том, что характер этого распределения определяет априорную склонность к увеличению одних компонент направляющего вектора и уменьшению других, предопределяя тем самым скрытое разделение признаков объектов на информативные и неинформативные. Выбор этого распределения играет роль регуляризации при формировании формальной байесовской постановки задачи обучения.

В диссертации экспериментально сравниваются методы обычной и неявной кросс-валидации в задаче распознавания образов на основе представления объекта вектором вторичных признаков, образованных сравнением его сходства с каждым из объектов обучающей совокупности. Отбор подмножества релевантных обучающих объектов, с которыми достаточно сравнивать предъявленный объект, обеспечивает критерий обучения Elastic Net SVM, содержащий структурный параметр, определяющий степень селективности отбора. В качестве источника экспериментальных данных использован известный массив «Chicken Pieces Silhouettes Database», содержащий 446 силуэтных изображений частей курицы (крылышко, спинка, ножка, бедрышко, грудка), представленных 44 разными функциями парного сравнения изображений. Оказалось, что неявная кросс-валидация выделяет всего несколько

релевантных объектов, достаточных для решающего правила, в то время как обычный скользящий контроль в данной задаче оказался нечувствительным к селективности отбора.

2 Метод опорных векторов и его обобщения

Если задана конечная обучающая выборка объектов реального мира $\omega \in \Omega$, представленных векторами числовых признаков

$$\boldsymbol{x}(\omega)=\left(x_i(\omega), i\in\mathbb{I}\right)\in\mathbb{R}^n,\ n=|\mathbb{I}|,\ \mathbb{I}=\{1,...,n\} \tag{1}$$

вместе со значениями их скрытой числовой характеристики

$$(\boldsymbol{X},\boldsymbol{y})=\left\{\left(\boldsymbol{x}(\omega_j)=\boldsymbol{x}_j, y(\omega_j)=y_j\right), y_j\in\mathbb{R}\ \text{ или }\ y_j=\pm1,\ j\in\mathbb{J}\right\}, \mathbb{J}=\{1,...,N\}, \tag{2}$$

то линейные методы оценки зависимости порождают линейное решающее правило $\left(\boldsymbol{a}=(a_i, i=1,...,n)\in\mathbb{R}^n, b\in\mathbb{R}\right)$, которое применимо к любому новому объекту в пространстве признаков $(\boldsymbol{x},y)\notin\left\{(\boldsymbol{x}_j,y_j),\ j\in\mathbb{J}\right\}$:

$$d(\boldsymbol{x})=\boldsymbol{a}^T\boldsymbol{x}+b\begin{cases}=\hat{y}(\boldsymbol{x})-\text{оценивание числовой регрессии},\\ \gtrless 0,\ \hat{y}(\boldsymbol{x})=\pm1-\text{обучение распознаванию образов}.\end{cases} \tag{3}$$

В данной диссертации мы будем рассматривать главным образом задачу распознавания образов, а именно один из нескольких вариантов общепринятого метода опорных векторов (SVM) [11,12]:

$$\begin{cases}\sum_{i\in\hat{\mathbb{I}}} a_i^2+C\sum_{j\in\mathbb{J}}\xi_j^2\to min(a_i, i\in\hat{\mathbb{I}}, b, \xi_j,\ j\in\mathbb{J}),\\ y_j\left(\sum_{i\in\hat{\mathbb{I}}} a_i x_{ij}+b\right)\geq 1-\xi_j, \xi_j\geq 0,\ j\in\mathbb{J}=\{1,...,N\}.\end{cases} \tag{4}$$

Здесь $\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,...,n\}$ – подмножество “активных” признаков, $\hat{n}=|\hat{\mathbb{I}}|\leq n$. Если $\hat{\mathbb{I}}$ зафиксировано, то основное преимущество метода SVM заключается в том, что он отбирает относительно небольшое подмножество векторов обучающей выборки $\boldsymbol{x}_j\in\mathbb{R}^{\hat{n}}$, $j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\subset\mathbb{J}$, называемых опорными векторами, которые полностью определяют оценку направляющего вектора разделяющей гиперплоскости $\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}\in\mathbb{R}^{\hat{n}}$. Легко показать [11,12], что направляющий вектор, получаемый в результате обучения, есть линейная комбинация только *опорных векторов*

$$\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}=\sum_{j\in J} y_j\hat{\lambda}_{\hat{\mathbb{I}},C,j}\boldsymbol{x}_j\in\mathbb{R}^{\hat{n}}. \tag{5}$$

Здесь переменные $(\hat{\lambda}_{\hat{\mathbb{I}},C,j},\ j\in\mathbb{J})$ – неотрицательные множители Лагранжа при ограничениях типа неравенств $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)\geq 1-\xi_j$ в (4). Если $\lambda_j=0$, то соответствующее ограничение неактивно $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)>1-\xi_j$ и $\xi_j=0$; наоборот, когда $\lambda_j>0$, это означает, что ограничение активно $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)=1-\xi_j$, и тогда говорят, что соответствующие объекты обучающей совокупности являются опорными векторами.

В результате итоговое решающее правило получается обычно намного более простым, чем выражение (3)

$$d(\boldsymbol{x})=\sum_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} y_j\hat{\lambda}_{\hat{\mathbb{I}},C,j}\boldsymbol{x}_j^T\boldsymbol{x}+\hat{b}_{\hat{\mathbb{I}},C}\gtrless 0,\ \hat{\mathbb{J}}_{\hat{\mathbb{I}},C}=\{j:\hat{\lambda}_{\hat{\mathbb{I}},C,j}>0\}\subseteq\mathbb{J}, \tag{6}$$

и полностью определяется подмножеством опорных векторов $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}$ вместе с их принадлежностями к классам $\left\{(\boldsymbol{x}_j,y_j),\ j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\right\}$ и вместе с положительными множителями Лагранжа, связанные с ними $\{\hat{\lambda}_{\hat{\mathbb{I}},C,j}>0,\ j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\}$.

Выбор подмножества активных признаков $\hat{\mathbb{I}}\subseteq\mathbb{I}$ вместе со значением структурного параметра $C>0$ в (4) полностью определяет количество опорных векторов $\hat{N}_{\hat{\mathbb{I}},C}=|\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|$, и определяет, таким образом, меру сложности решающего правила (6). Как следствие, пара структур-

ных параметров $(\hat{\mathbb{I}}, C)$ оказывает критическое влияние на обобщающую способность получаемой SVM.

Если отсутствует отдельная тестовая выборка, то единственным способом подбора $(\hat{\mathbb{I}}, C)$ является кросс-валидация на обучающей совокупности (2). Однако, традиционная *явная* кросс-валидация требует многократных повторений обучения для оценки структурных параметров $(\hat{\mathbb{I}}, C)$.

В данной работе мы даем математическое обоснование предположения, сформулированного в [13], что информационный критерий Акаике (AIC) [14], изначально разработанный для оценки регрессионной зависимости, может рассматриваться как *гипотетическая* кросс-валидация. Его преимущество заключается в том, что он дает аналитический инструмент для сравнения разных моделей.

Мы исходим из более общего взгляда на машинное обучение и математически имитируем процесс кросс-валидации, опираясь на некоторую естественную эвристику о виде неизвестного распределения вероятностей, породившего обучающую совокупность. Такой принцип мы называем «гипотетической бесперебoрной кросс-валидацией», и показываем, что классический AIC является его частным случаем при некоторых особо простых предположениях. Применение этого принципа к SVM явно использует зависимость результата обучения только от опорных векторов.

Мы применяем принцип бесперебoрной гипотетической кросс-валидации не только к классическому методу SVM с заданным набором индивидуальных признаков объектов $\hat{\mathbb{I}} \subseteq \mathbb{I}$ (1), но также и к методу релевантных векторов (RVM) [15,16,18], основанному на предположении о том, что объекты реального мира $\omega \in \Omega$ воспринимаются только с помощью произвольной функции $S(\omega', \omega'')$ их попарных сходств или различий. Идея заключается в интерпретации совокупности значения этой функции между случайным объектом $\omega \in \Omega$ и объектами обучающей выборки $\{\omega_j, j=1,\ldots,N\}$ как вектор вторичных признаков этого объекта

$$\boldsymbol{x}(\omega) = \left(x_i(\omega) = S(\omega_i, \omega), i \in \mathbb{I} = \mathbb{J}\right), \ \mathbb{I} = \{1,\ldots,n\} = \{1,\ldots,N\}, \tag{7}$$

с последующим применением стандартного метода SVM в $\mathbb{R}^n = \mathbb{R}^N$. Мы рассматриваем обобщенный метод SVM с отбором признаков

$$\begin{cases} \sum_{i\in\mathbb{I}} a_i^2 + \mu \sum_{i\in\mathbb{I}} |a_i| + C \sum_{j\in J} \xi_j^2 \to \min(a_i, i \in \mathbb{I} = \mathbb{J}, b, \delta_j, j \in \mathbb{J}), \\ y_j \left(\sum_{i\in\mathbb{I}} a_i x_{ij} + b\right) \geq 1 - \xi_j, \xi_j \geq 0, \ j \in \mathbb{J} = \{1,\ldots,N\}, \end{cases} \tag{8}$$

который отличается от предложенного в [17] под названием *Doubly Regularized SVM* или *Elastic Net SVM* только квадратичным штрафом ξ_j^2 (вместо ξ_j) за нарушение основного требования SVM, заключающегося в обеспечении положительной величины зазора между двумя классами объектов обучающей выборки $y_j \left(\sum_{i\in\mathbb{J}} a_i x_{ij} + b\right) \geq 1$.

Наличие L_1 регуляризации $\sum_{i\in\mathbb{I}} |a_i|$ с взвешивающим параметром $\mu \geq 0$, в отличие от (4), обеспечивает важнейшее свойство теперь уже дважды регуляризованного метода SVM приписывать лишним компонентам направляющего вектора a_i строго нулевые значения, автоматически находя, таким образом, подмножество «информативных» признаков $\hat{\mathbb{I}}_{C\mu} = \{\hat{a}_{C\mu,i} \neq 0\} \subset \mathbb{I} = \mathbb{J}$, то есть, «информативных» объектов обучающей выборки.

Эти объекты называют *релевантными объектами*, или *релевантными векторами* в случае, если функция сравнения $S(\omega', \omega'')$ обладает свойствами кернела, т.е. погружает объекты в линейное пространство [16]. Если $\mu = 0$, метод превращается в классический SVM с сохранением всех признаков $\hat{\mathbb{I}}_{C\mu} = \mathbb{I} = \mathbb{J}$. Напротив, если $\mu \to \infty$, критерий становится слишком селективным $\hat{\mathbb{I}}_{C\mu} \to \varnothing$. Таким образом, с ростом структурного параметра μ характер обучения изменяется от отсутствия селективности (полного сохранения всех вторичных признаков) до избыточной селективности (полного подавления всех признаков).

Этот факт существенно использует наш принцип беспереборной гипотетической кросс-валидации, позволяя при оценке регуляризующих параметров (C,μ), в основном, селективного параметра μ, избежать явной кросс-валидации, чрезвычайно затратной с вычислительной точки зрения, и обеспечить тем самым наилучшую обобщающую способность дважды регуляризованного метода SVM.

3 Принцип гипотетической кросс-валидации

Наш принцип беспереборной гипотетической кросс-валидации основан на двух эвристиках, которые будут сформулированы в подсекциях 3.2, 3.3 и 3.4. Чтобы пояснить терминологию, мы сначала кратко опишем общую задачу восстановления зависимости по эмпирическим данным в подсекции 3.1.

3.1 Общая задача восстановления зависимости

Пусть Ω – набор объектов реального мира $\omega\in\Omega$, каждый из которых связан с двумя характеристиками, измеряемыми в произвольных шкалах $\boldsymbol{x}(\omega)\in\mathbb{X}$ и $y(\omega)\in\mathbb{Y}$, первая из которых всегда доступна для непосредственного наблюдения, как на этапе «обучения», так и на этапе «экзамена», а вторая доступна только при обучении, но скрыта при «экзамене». Предполагается, что объекты $\omega\in\Omega$ многократно и независимо выбираются природой согласно объективно существующего распределения вероятностей как пара $(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}$. Плотность распределения

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} f^*(\boldsymbol{x},y)d\boldsymbol{x}dy=1,\ \ f^*(x,y)=\mathrm{g}^*(\boldsymbol{x})\varphi^*(y|\boldsymbol{x}), \tag{9}$$

неизвестна наблюдателю, желание которого заключается в решении задачи оценивания ненаблюдаемой характеристики случайно появляющихся объектов $\hat{y}(x):\mathbb{X}\to\mathbb{Y}$, т.е., в восстановлении скрытой зависимости.

В частности, в задаче распознавания образов $\mathbb{Y}=\{-1,1\}$, понимание условного распределения вероятностей $\varphi^*(y|x)$ в (9) как плотности вполне согласуется с бинарным характером индекса класса объекта $y=\pm1$, поскольку в случае двух классов всегда можно трактовать этот индекс как действительнозначную случайную величину, сконцентрированную в двух точках, и задаваемую сингулярной плотностью

$$\varphi^*(y|\boldsymbol{x})=(1-p^*)\delta\big(y-(-1)\big)+p^*\delta(y-1), \tag{10}$$

где $\delta(z)$ – дельта-функция Дирака, $p^*=P^*(y=1|\boldsymbol{x})$.

Предположим, что наблюдатель получил конечную обучающую совокупность $(\boldsymbol{X},\boldsymbol{y})$ вида (2). Предположим далее, что наблюдатель намерен использовать параметрическое семейство решающих правил $\hat{y}(\boldsymbol{x},\boldsymbol{a})$, $\boldsymbol{a}\in\mathbb{R}^n$, и функцию штрафа

$$q(y,\boldsymbol{x},\boldsymbol{a}),\ \text{в простейшем случае,}\ q(y,\boldsymbol{x},\boldsymbol{a})=Loss\big(y,\hat{y}(\boldsymbol{x},\boldsymbol{a})\big). \tag{11}$$

Оптимальной стратегией для выбора параметра $\boldsymbol{a}$ была бы минимизация среднего риска $\int_{\mathbb{X}}\int_{\mathbb{Y}} q(y,\boldsymbol{x},\boldsymbol{a})f^*(\boldsymbol{x},y)dyd\boldsymbol{x}\to\ \min(\boldsymbol{a})$, однако это невозможно, потому что распределение неизвестно. Общепринятый компромисс – минимизировать эмпирический риск, вычислимый с помощью доступной обучающей выборки

$$Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})=\frac{1}{N}\sum\nolimits_{j=1}^{N}q(y_j,\boldsymbol{x}_j,\boldsymbol{a})\propto\sum\nolimits_{j=1}^{N}q(y_j,\boldsymbol{x}_j,\boldsymbol{a})\to\min(\boldsymbol{a}). \tag{12}$$

Допустим, далее, что наблюдатель имеет некоторое интуитивное априорное представление о значении главного параметра $\boldsymbol{a}\in\mathbb{R}^n$, сформулированное им в виде функции, зависящей от параметра, которую, как ему представляется, нужно минимизировать $V(\boldsymbol{a},C)\to\min(\boldsymbol{a})$. Здесь C – дополнительный скалярный или векторный параметр, называемый *структурным параметром*, служащий для того, чтобы контролировать степень нежелательности выхода $\boldsymbol{a}$

от некоторого подмножества $\mathbb{A} \subset \mathbb{R}^n$, связываемого наблюдателем с “особенно простыми” и поэтому предпочтительными решающими правилами.

Общепринято практикой является использование регуляризованной версии критерия обучения в виде компромисса

$$\hat{\boldsymbol{a}}_C(\boldsymbol{X}, \boldsymbol{y}) = \arg\min_{\boldsymbol{a}} \{V(\boldsymbol{a}, C) + Q(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{a})\}. \tag{13}$$

Очевидно, что результат обучения будет самым существенным образом зависеть от величины структурного параметра C, и его выбор является, пожалуй, главной задачей теории машинного обучения.

3.2 Эвристика наблюдателя о незлонамеренности природы

Всегда возможно представить неизвестную совместную плотность вероятности скрытой и наблюдаемой характеристик произвольного объекта реального мира $f^*(\boldsymbol{x}, y)$ как произведение маргинальной плотности одной переменной и условной плотности другой (9). Таким образом, совместная плотность распределения обучающей выборки в целом может быть представлена как произведение

$$F^*(\boldsymbol{X}, \boldsymbol{y}) = G^*(\boldsymbol{X})\Phi^*(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{j=1}^{N} g^*(\boldsymbol{x}_j)\varphi^*(y_j \mid \boldsymbol{x}_j). \tag{14}$$

Естественно, обе плотности остаются здесь неизвестными, но допустим наблюдатель пытается несколько «смягчить» свое полное незнание, мысленно вводя в неизвестное распределение $\varphi^*(y \mid \boldsymbol{x})$ параметр $\boldsymbol{a}$, который существует только в его воображении $\varphi^*(y \mid \boldsymbol{x}) = \int_{\mathbb{R}^n} \varphi(y \mid \boldsymbol{x}, \boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}$. В терминах предполагаемого механизма формирования обучающей совокупности, это означает равенство

$$\Phi^*(\boldsymbol{y} \mid \boldsymbol{X}) = \int_{\mathbb{R}^n} \Phi(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}, \text{ где}$$

$$\Phi(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{a}) = \prod_{j=1}^{N} \varphi(y_j \mid \boldsymbol{x}_j, \boldsymbol{a}) \tag{15}$$

рассматривается как полностью известное семейство условных распределений, тогда как $\Psi^*(\boldsymbol{a})$, напротив, предполагается абсолютно неизвестным. Другими словами, наблюдатель рассматривает $\Phi^*(\boldsymbol{y} \mid \boldsymbol{X})$ в (1444) как неизвестную смесь известных условных распределений.

Рассмотрение параметрического семейства $\varphi(y \mid \boldsymbol{x}, \boldsymbol{a})$ как экспоненты от функции потерь

$$\varphi(y \mid \boldsymbol{x}, \boldsymbol{a}) \propto \exp\big(-q(y, \boldsymbol{x}, \boldsymbol{a})\big), \Phi(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{a}) \propto \exp\big(-Q(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{a})\big), \tag{16}$$

где нормализующий коэффициент, обозначаемый символом пропорциональности $\propto$, не зависит от $\boldsymbol{a}$, эквивалентно предположению о *незлонамеренности природы*, т.е. неявно предполагается, что для каждого $\boldsymbol{a} \in \mathbb{R}^n$ пары $(\boldsymbol{x}, y)$, соответствующие малым величинам принятой функции потерь $q(y, \boldsymbol{x}, \boldsymbol{a})$, генерируются чаще, чем при больших величинах потерь.

3.3 Мысленный эксперимент

Рассматриваемый в диссертации принцип неявной кросс-валидации основан на следующем мысленном эксперименте наблюдателя. Пусть природа разыграла конкретное значение случайного параметра $\boldsymbol{a} \in \mathbb{R}^n$ в соответствии с неизвестной плотностью $\Psi^*(\boldsymbol{a})$, а также выборку наблюдаемых характеристик объектов $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ согласно опять же неизвестной плотности $G^*(\boldsymbol{X})$ (14).

Предположим, далее, что природа, дважды независимо применив условное распределение $\Phi(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{a})$, получила две реализации характеристик объектов $\boldsymbol{y} = (y_1, ..., y_N)$ и $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_N)$. Таким образом, можно мысленно рассматривать две гипотетических выборки $(\boldsymbol{X}, \boldsymbol{y})$ and $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$ с одинаковыми величинами $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$. Будем использовать их как обучающую выборку $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$, которая дает некоторую оценку целевого параметра $\hat{\boldsymbol{a}}_C(\boldsymbol{X}, \tilde{\boldsymbol{y}})$

(1313), и тестовую выборку $(\boldsymbol{X},\boldsymbol{y})$, по которой вычислим функциию потерь $Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)$. Суть гипотетической кросс-валидации заключается в минимизации математического ожидания потерь:

$$\iiint Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}\times \\ \times G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X}\to\min(C). \tag{17}$$

В реальности у наблюдателя есть только одна обучающая выборка $(\boldsymbol{X},\boldsymbol{y})$, и он можем вычислить потери только на той же выборке, которая использовалась для обучения.

Насколько испортится критерий, подлежащий максимизации, по сравнению с идеей (17)? Каким должен быть штраф за использование оценки $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})$, вычисленной по той же выборке, вместо независимой оценки $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$?

Теорема 1. *Эквивалентная форма критерия (17) имеет вид:*

$$C^*=\arg\min_C\Big\{\int_{\mathbb{X}\times\mathbb{Y}} Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{y},\boldsymbol{X})\big)F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{X}d\boldsymbol{y}+ \\ \int_{\mathbb{X}\times\mathbb{R}^m}\Delta(C,\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})G^*(\boldsymbol{X})d\boldsymbol{X}d\boldsymbol{a}\Big\}, \tag{18}$$

$$\text{где } \Delta(C,\boldsymbol{X},\boldsymbol{a})=\int_{\mathbb{Y}\times\mathbb{Y}} Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\times \\ \times\Phi(\tilde{\boldsymbol{y}}\,|\,\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}. \tag{19}$$

Доказательство теоремы приведено в приложении 8.1.

Для многих типичных функций потерь $Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})$ и регуляризирующих функций $V(\boldsymbol{a},C)$ (13), применимых к широкому классу практических задач, штраф (*19*) не зависит от параметра $\boldsymbol{a}$:

$$\int_{\mathbb{Y}\times\mathbb{Y}} Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\times \\ \times\Phi(\tilde{\boldsymbol{y}}\,|\,\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}=\Delta(C,\boldsymbol{X}). \tag{20}$$

Теорема 2. В случае независимого от параметра штрафа (*19*), идея гипотетической кросс-валидации (17) преобразуется к простому выражению:

$$C^*=\arg\min_C\left\{\iint\Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)+\Delta\big(C,\boldsymbol{X}\big)\Big]F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{y}d\boldsymbol{X}\right\}. \tag{21}$$

Доказательство теоремы приведено в приложении 8.2.

3.4 Критерий гипотетической кросс-валидации

Однако, критерий (21) все еще не подходит для практического использования, потому что плотность совместного распределения неизвестна наблюдателю.

Вторая эвристика заключается в замене математического ожидания (21) его несмещенной оценкой:

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_C\underbrace{\Big\{\underbrace{Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)}_{\textit{empirical risk}}+\underbrace{\Delta(C,\boldsymbol{X})}_{\textit{penalty}}\Big\}}_{\textit{structural risk}}. \tag{22}$$

Это и есть критерий гипотетической бесперебoрной кросс-валидации, который мы рассматриваем в этой работе.

Его структура аналогична критерию Вапника-Червоненкиса, который состоит в минимизации структурного риска [11], но отличается от него интерпретацией штрафа $\Delta(C,\boldsymbol{X})$ (20). В теории Вапника-Червоненкиса штраф характеризует верхнюю грань неизвестного среднего риска, который получен из общих неравенств теории вероятностей и параметризован размерностью Вапника-Червоненкиса (VC-dimension). Следует помнить, что понятие VC-размерности было сформулировано только для простейшей бинарной функции потерь в распознавании образов и неприменимо, в частности, к SVM. В отличие от этого, штраф $\Delta(C,\boldsymbol{X})$ (20) применим для более широкого класса функций потерь, но базируется на более ограни-

чивающих эвристических предположениях о данных. Как мы увидим далее в разделе 5, это совместимо со структурой SVM.

В соответствии с первым эвристическим предположением (16), мы имеем $Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})=-\ln\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})+const$. Гипотетическая кросс-валидация

$$\int\left[\ln\Phi\left(\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\right)\right]\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}\to\max$$

аналогична, таким образом, максимизации информации по Кульбаку о неизвестном распределении $\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$, содержащейся в его оценке, полученной по другой выборке $\Phi\left(\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\right)$. Поэтому уместно интерпретировать наш критерий гипотетической кросс-валидации как неявно основанный на информационно-теоретических соображениях и рассматривать его как обобщение классической идеи Хиротуги Акаике, изложенной в [4].

4 Простейший частный случай: Линейная регрессия и исходный информационный критерий Акаике

4.1 Модель линейной регрессии.

Пусть скрытая характеристика принимает действительные значения $y\in\mathbb{Y}=\mathbb{R}$, а наблюдаемая является действительным вектором, т.е., $\boldsymbol{x}\in\mathbb{X}=\mathbb{R}^n$. Мы будем предполагать, что функция потерь (11)-(12) линейно-квадратична

$$\begin{aligned}&q(y,\boldsymbol{x},\boldsymbol{a})=(y-\boldsymbol{x}^T\boldsymbol{a})^2,\ \ \boldsymbol{X}=(\boldsymbol{x}_1\cdots\boldsymbol{x}_N)\ (n\times N),\\&Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})=\left\|\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right\|^2=(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}),\end{aligned}\tag{23}$$

и, таким образом, предполагаемое условное распределение (16) является нормальным $\varphi(y|\boldsymbol{x},\boldsymbol{a})=\mathcal{N}(y|\boldsymbol{x}^T\boldsymbol{a},\sigma^2)$ с заданной дисперсией $\sigma^2=1/2$. Таким образом, это задача оценивания линейной регрессии.

В случае простейшей квадратичной функции регуляризации $V(\boldsymbol{a},C)=\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a}$, где симметричная положительно полуопределенная матрица $\boldsymbol{B}_C$ $(n\times n)$ зависит от структурного параметра C, комбинированный критерий обучения (13) даст оценку вектора коэффициентов регрессии

$$\begin{aligned}&\hat{\boldsymbol{a}}_C(\mathbf{X},\mathbf{y})=\operatorname{argmin}_{\boldsymbol{a}}\left\{\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a}+\sum_{j\in\hat{\mathbb{J}}_\beta}(y_j-\boldsymbol{x}_j^T\boldsymbol{a})^2\right\}=\\&\arg\min_{\boldsymbol{a}}\left\{\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a}+\left\|\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right\|^2\right\}=(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{y}.\end{aligned}\tag{24}$$

Теорема 3. *Для линейно-квадратичной функции потерь (23) и, таким образом, нормальной условной плотности скрытой переменной* $\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$ *(16), штраф (19) за использование оценки* $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})$ *(24), вычисленной по той же выборке, вместо независимой оценки* $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$ *не зависит от неизвестного параметра* $\boldsymbol{a}$ *(20):*

$$\Delta(C,\boldsymbol{X})=Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]\tag{25}$$

Доказательство теоремы приведено в приложении 8.3.

Таким образом, критерий гипотетической кросс-валидации (22) для модели линейной регрессии имеет вид

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_C\left\{\left\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\right\|^2+Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]\right\}.\tag{26}$$

4.2 Исходный информационный критерий Акаике

Покажем, что исходный информационный критерий Акаике [4] является частным случаем критерия гипотетической кросс-валидации для модели линейной регрессии (26). Для этого достаточно дополнительно предположить, что компоненты вектора коэффициентов ре-

грессии априори упорядочены $\boldsymbol{a}=(a_1\cdots a_n)$, и что целочисленный структурный параметр $0\le C\le n$ соответствует числу ненулевых регрессионных коэффициентов:

$$\boldsymbol{a}=(\boldsymbol{a}_C,\boldsymbol{a}_{n-C})=(a_1,...,a_C,a_{C+1}=0,...,a_n=0)$$

Это предположение может быть представлено простейшей квадратичной функцией регуляризации

$$V(\boldsymbol{a},C)=\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a},\ \ \boldsymbol{B}_C=Diag\Big(\underbrace{\frac{1}{\rho}\cdots\frac{1}{\rho}}_{C}\underbrace{\rho\cdots\rho}_{n-C}\Big),\ \rho\to\infty. \tag{27}$$

Теорема 4. *При предположении (27)*

$$\lim_{\rho\to\infty}\Delta(C,\boldsymbol{X})=\lim_{\rho\to\infty}Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]=C\,.$$

Доказательство теоремы приведено в приложении 8.4.

Таким образом, критерий гипотетической кросс-валидации (26) сокращается до особенно простого вида:

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_C\left\{\left\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\right\|^2+C\right\}. \tag{28}$$

Это как раз и есть идея информационного критерия Акаике [4].

Сравнение (28) и (22) позволяет нам интерпретировать штрафной член $\Delta(C,\mathbf{X})$ в критерии гипотетической кросс-валидации как *обобщенную вещественную размерность модели данных*.

5 Второй частный случай: метод опорных векторов

5.1 Подмножество опорных векторов как независимый нечисловой структурный параметр

Рассмотрим параметрическое семейство дискриминантных гиперплоскостей $\boldsymbol{a}^T\boldsymbol{x}+b\gtrless 0$ в $\hat{n}$-мерном пространстве признаков (4)

$$\boldsymbol{x}=(x_i,i\in\hat{\mathbb{I}})\in\mathbb{R}^{\hat{n}},\boldsymbol{a}=(a_i,i\in\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,...,n\})\in\mathbb{R}^{\hat{n}},b\in\mathbb{R},|\hat{\mathbb{I}}|=\hat{n}, \tag{29}$$

которые разделяют два класса объектов $y=\pm 1$, представленных векторами признаков $\boldsymbol{x}\in\mathbb{R}^{\hat{n}}$. Цель классификации заключается в том, чтобы выбрать такую гиперплоскость, чтобы признаковые векторы объектов разных классов попали главным образом в разные полупространства. Пусть функция потерь $q(y,\boldsymbol{x},\boldsymbol{a},b)$ (11), применимая к любым объектам $(\boldsymbol{x},y)$, будет выбрана в виде

$$q(y,\boldsymbol{x},\boldsymbol{a},b)=\begin{cases}0, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b)>1,\\ \left(1-y(\boldsymbol{a}^T\boldsymbol{x}+b)\right)^2, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b)<1,\end{cases} \tag{30}$$

что штрафует вектор признаков $\boldsymbol{x}$, если он находится на неправильной стороне от гиперплоскости, но также штрафует расстояние от точки $\boldsymbol{x}$ до гиперплоскости по евклидовой метрике на правильной стороне, меньшее, чем 1. Соответствующий эмпирический риск обучающей совокупности (2) будет выражаться суммой:

$$Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a},b)=\sum_{j\in\mathbb{J}}q(y_j,\boldsymbol{x}_j,\boldsymbol{a},b)=\sum_{j:y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1}\left(y_j-(\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2. \tag{31}$$

Если мы примем квадратичную функцию регуляризации $V(\boldsymbol{a},b)=\boldsymbol{a}^T\boldsymbol{a}$, то обучающий критерий (13) примет вид

$$\begin{pmatrix}\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})\\ \hat{b}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})\end{pmatrix}=\arg\min_{\boldsymbol{a},b}\left\{\boldsymbol{a}^T\boldsymbol{a}+C\sum_{j:y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1}\left(y_j-(\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2\right\}, \tag{32}$$

что эквивалентно стандартному критерию SVM (4).

Нам будет удобно заменить парную запись параметров дискриминантной гиперплоскости $(\boldsymbol{a},b)$ на единый вектор $\mathbf{a}=(\boldsymbol{a},b)\in\mathbb{R}^{\hat{n}+1}$, добавив еще один элемент b к направляющему вектору, а также элемент 1 к вектору признаков, так что $\mathbf{x}=(\boldsymbol{x},1)\in\mathbb{R}^{\hat{n}+1}$. В таком случае, задача SVM (4) примет вид

$$\begin{cases}\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\sum_{j\in\mathbb{J}}\xi_j^2\to min(\mathbf{a},\xi_j,\, j\in\mathbb{J}),\\ y_j\mathbf{a}^T\mathbf{x}_j\geq 1-\xi_j,\,\xi_j\geq 0,\, j\in\mathbb{J}=\{1,...,N\},\end{cases}\quad \underset{((\hat{n}+1)\times(\hat{n}+1))}{\mathbf{B}_{\hat{\mathbb{I}},C}}=\begin{pmatrix}(1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0}\\ \mathbf{0}^T & 0\end{pmatrix}. \tag{33}$$

Решение соответствующей двойственной задачи даст нам оптимальную дискриминантную гиперплоскость $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})$, т.е., $(\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C},\hat{b}_{\hat{\mathbb{I}},C})$ и решающее правило классификации (3), которое применимо к векторам новых объектов, а также подмножество опорных объектов обучающей совокупности $\hat{\mathbb{J}}_C\subseteq\mathbb{J}$:

$$\begin{aligned}&\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}=\left\{j\in\mathbb{J}: y_j\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})=1-\hat{\xi}_j,\,\hat{\xi}_j>0\right\},\\ &\hat{\xi}_j^2=(1-y_j\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2=y_jy_j(1-y_j\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2=(y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2.\end{aligned} \tag{34}$$

Традиционно, для выбора подходящих величин структурных параметров $(\hat{\mathbb{I}},C)$, пользователь должен повторять обучение (33) для достаточно большого набора пробных значений, и принять тот результат $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})$, который доставляет минимальное значение ошибке кросс-валидации.

Чтобы сделать понятие гипотетической бесперебoрной кросс-валидации, сформулированное в 4.1 для модели линейной регрессии, применимым также и к SVM, мы будем полностью отождествлять значение комбинированного структурного параметра $(\hat{\mathbb{I}},C)$ с итоговым подмножеством опорных векторов $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\subseteq\mathbb{J}$. Пусть задача обучения (33) решена для некоторой комбинации $(\hat{\mathbb{I}},C)$, т.е. найдено подмножество опорных объеков. Заметим, что только векторы признаков и индексы классов опорных объектов

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}=\begin{array}{c}\hat{N}_{\hat{\mathbb{I}},C}\\ \hat{n}\\ 1\end{array}\begin{pmatrix}\boldsymbol{x}_{j_1}\cdots\boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},}}}\\ 1\cdots 1\end{pmatrix}=\begin{pmatrix}\hat{\boldsymbol{X}}_C\\ 1\cdots 1\end{pmatrix},\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}=(y_{j_1}\cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T,\,\hat{N}_{\hat{\mathbb{I}},C}=|\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|, \tag{35}$$

влияют на результат обучения. Действительно, в соответствии с (33), (34) и (35), подмножество опорных объектов полностью определяет направляющий вектор оптимальной разделяющей гиперплоскости:

$$\begin{aligned}&\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})=\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y};\hat{\mathbb{J}}_{\hat{\mathbb{I}},C})=\operatorname{argmin}\left\{\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\sum_{j\in\hat{\mathbb{J}}_\beta}(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}=\\ &\operatorname{argmin}\left(\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\left\|\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}-\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T\mathbf{a}\right\|^2\right)=(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}.\end{aligned} \tag{36}$$

Мы рассматриваем подмножество опорных векторов как независимый структурный параметр решающего правила. Суть нашего подхода заключается в том, чтобы рассматривать подмножество $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}$ как структурный параметр, подлежащий кросс-валидации в последовательности пробных значений $[\hat{\mathbb{J}}_{\hat{\mathbb{I}}_1,C_1},...,\hat{\mathbb{J}}_{\hat{\mathbb{I}}_m,C_m}]$ вместо последовательности пробных значений исходного структурного параметра $[(\hat{\mathbb{I}}_1,C_1),...,(\hat{\mathbb{I}}_m,C_m)]$.

5.2 Применение эвристики наблюдателя о незлонамеренности природы

По своей математической формулировке задача (36) совпадает с задачей для регрессионного оценивания (24). Однако, фундаментальное различие состоит в том, что переменные y_j принимают в (36) только два значения $y_j=\pm 1$, тогда как в (24) они являются вещественными числами $y_j\in\mathbb{R}$. Как следствие, попытка применить предположение (16) приводит к неизбежной зависимости нормализующего коэффициента $\propto$ от неизвестного вектора $\boldsymbol{a}$, и последующий математическая структура в секциях 3.2 и 3.3 становится неприменимой.

Эвристический выход, который мы здесь излагаем, состоит в том, чтобы максимально продолжить формальную аналогию (36) и (24), и буквально рассматривать задачу SVM с заданным подмножеством опорных объектов, как если бы это было для оценивания регрессии. Такая замена приведет к использованию выражения (*25*) как штрафа за использование оценки $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})$ (36), вычисленной по той же выборке, вместо независимой оценки $\hat{\mathbf{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$:

$$\Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C}(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C}+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\right]. \tag{37}$$

Таким образом, наш эвристический критерий гипотетической кросс-валидации для SVM по своей структуре аналогичен соответствующему критерию для модели линейной регрессии (26):

$$(\hat{\mathbb{I}},C)(\boldsymbol{y},\boldsymbol{X}) = \arg\min_{\hat{\mathbb{I}},C}\left\{\sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}}\hat{\xi}^2_{\hat{\mathbb{I}},C,j}+Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C}(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}^T_{\hat{\mathbb{I}},C}+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\right]\right\}. \tag{38}$$

Критерий (38) является универсальным в том смысле, что он применим к любой последовательности пар $[(\hat{\mathbb{I}}_1,C_1),...,(\hat{\mathbb{I}}_m,C_m)]$. Однако опыт показывает, что выбор коэффициента C существенно менее проблематичен, нежели выбор подмножества признаков $\hat{\mathbb{I}}$. Многочисленные эксперименты показали, что при обучении по методу опорных векторов всегда целесообразно принимать как можно бóльшие значения C, в то время как выбор подмножества признаков неизбежно должен контролироваться той или иной версией кросс-валидации.

Следующая теорема показывает, что если значение коэффициента C достаточно велико и фиксировано, критерий гипотетической кросс-валидации (38) может использоваться в существенно более простом виде.

Теорема 5. $\lim_{C\to\infty}\Delta(\hat{\mathbb{I}},C,\boldsymbol{X}) = \min\{\hat{n},\hat{N}_{\hat{\mathbb{I}},\mathrm{C}}-1\}+1$, *где* $\hat{n}=|\hat{\mathbb{I}}|$, $\hat{N}_{\hat{\mathbb{I}},\mathrm{C}}=|\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|$ *– число опорных объектов в соответствии с (29) и (35).*

Доказательство теоремы приведено в приложении 8.5.

Последняя теорема показывает, что штраф (37), по крайней мере, для больших значений коэффициента C, может пониматься как эффективная размерность параметров гиперплоскости $(\boldsymbol{a}\in\mathbb{R}^{\hat{n}},b\in\mathbb{R})$, которая обычно меньше, чем $\hat{n}+1=|\hat{\mathbb{I}}|+1$.

Применительно к выбору признакового подмножества $\hat{\mathbb{I}}\subseteq\mathbb{I}$ с достаточно большим значением C, наш критерий гипотетической кросс-валидации для SVM (4), (33) и (38), в соответствии с записью (35), принимает форму:

$$\hat{\mathbb{I}}_C(\boldsymbol{y},\boldsymbol{X}) = \arg\min_{\hat{\mathbb{I}}\subseteq\mathbb{I}}\left\{\sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}}\hat{\xi}^2_{\hat{\mathbb{I}},C,j}+\left(\min\{\hat{n},\hat{N}_{\hat{\mathbb{I}},\mathrm{C}}\}+1\right)\right\}. \tag{39}$$

Однако, огромное число 2^n всех подмножеств признаков $\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,...,n\}$ препятствует прямому использованию этого “наивного” правила. В следующем разделе мы покажем, как можно быстро выбрать пробную последовательность подмножеств $[\hat{\mathbb{I}}_1,...,\hat{\mathbb{I}}_m]$ в исходном множестве признаков $\hat{\mathbb{I}}_k\subset\mathbb{I}$.

5.3 Метод релевантных векторов (объектов)

Дополнительный регуляризующий член $\mu\sum_{i\in\mathbb{I}}|a_i|$ в (8) добавляет в критерий обучения новые характеристики, которые существенно отличаются от стандартного SVM. Этот член служит для автоматического отбора наиболее информативного подмножества вторичных признаков, *релевантных объектов* $\hat{\mathbb{I}}_{C\mu}\subset\mathbb{I}=\mathbb{J}$, роль которых имеет много общего с опорными объектами (векторами) в классическом SVM, только теперь *релевантные объекты* соответствуют ненулевым коэффициентам решения выпуклой задачи обучения. Так как дважды регуляризованный критерий остается выпуклым, то не имеет значения, какой алгоритм используется для численного решения. В частности, алгоритм, предложенный в [18] эффективно выделяет оптимальное подмножество вторичных признаков (объектов $\hat{\mathbb{I}}_{C\mu}$).

Экспериментальные исследования показали, что целесообразно зафиксировать достаточно большое значение параметра C и варьировать только один параметр селективности μ.

После того, как подмножество вторичных признаков выбрано

$$\hat{\mathbb{I}}_{C\mu} \subset \mathbb{I},\ \hat{n}_{C\mu} = |\hat{\mathbb{I}}_{C\mu}|, \tag{40}$$

становится целесообразным применить обычный метод SVM (4) для этого подмножества:

$$\begin{cases} \sum_{i \in \hat{\mathbb{I}}_{C\mu}} a_i^2 + C \sum_{j \in \mathbb{J}} \xi_j^2 \to min(a_i, i \in \hat{\mathbb{I}}_{C\mu}, b, \xi_j,\ j \in \mathbb{J}), \\ y_j \left(\sum_{i \in \hat{\mathbb{I}}_{C\mu}} a_i x_{ij} + b \right) \geq 1 - \xi_j, \xi_j \geq 0,\ j \in \mathbb{J} = \{1, ..., N\}. \end{cases} \tag{41}$$

Применение любого стандартного алгоритма дает подмножество опорных объектов

$$\hat{\mathbb{J}}_{C\mu} = \{ j : \hat{\xi}_{C\mu, j} > 0 \},\ \hat{N}_{C\mu} = |\hat{\mathbb{J}}_{C\mu}|, \tag{42}$$

т.е., таких объектов, что $y_j (\sum_{i \in \hat{\mathbb{I}}_{C\mu}} \hat{a}_{C\mu, i} x_{ij} + \hat{b}_{C\mu}) = 1 - \hat{\xi}_{C\mu, j}$. В отличие от (6), подмножество опорных объектов, во-первых, будет зависеть от обоих структурных параметров (C, μ), и, во-вторых, определяет решающее правило, которое принимает во внимание только релевантные вторичные признаки любого нового объекта $i \in \hat{\mathbb{I}}_{C\mu}$ (7):

$$\begin{gathered} d\left(\boldsymbol{x}(\omega) | C, \mu\right) = d\left(x_1(\omega), ..., x_N(\omega) | C, \mu\right) = \\ \sum_{j \in \hat{\mathbb{J}}_{C\mu}} y_j \hat{\lambda}_{C,j} \sum_{i \in \hat{\mathbb{I}}_{C\mu}} x_{ij} x_i + \hat{b}_C \gtrless 0,\ \ x_i = x_i(\omega) = S(\omega_i, \omega),\ i \in \mathbb{I} = \mathbb{J}. \end{gathered} \tag{43}$$

Обучение в любой из точек $(\mu_1 < ... < \mu_m)$ дает соответствующую последовательность релевантных подмножеств $(\hat{\mathbb{I}}_{C\mu_1}, ..., \hat{\mathbb{I}}_{C\mu_m})$ имеющих размеры $(\hat{n}_{C\mu_1}, ..., \hat{n}_{C\mu_m})$, которые, в основном, имеют тенденцию формировать вложенные подмножества $(\hat{\mathbb{I}}_{C\mu_1} \supset \hat{\mathbb{I}}_{C\mu_2} \supset ... \supset \hat{\mathbb{I}}_{C\mu_m})$, однако, последнее свойство не всегда строго выполняется. Нам необходимо найти «наиболее подходящий» уровень селективности, используя принцип гипотетической кросс-валидации.

К версии дважды регуляризованного SVM (8), которую мы называем методом релевантных векторов с управляемой селективностью, вполне применима эвристическая аналогия с задачей оценивания регрессионной зависимости, которую мы применили к обычной SVM в секции 5.2. Штраф (37) за некорректную оценку потерь полностью применим для формулировки SVM (41) с зафиксированным подмножеством вторичных признаков $\hat{\mathbb{I}}_{C\mu} \subset \mathbb{I} = \mathbb{J}$ ранее оцененных дважды регуляризованным SVM (8). Единственное различие состсоит в том, что матрицы $\mathbf{B}_{C\mu}$ и $\hat{\mathbf{X}}_{C\mu}$ имеют теперь меньшие размерности, чем (33) и (35) в соответствии с (40) и (42):

$$\hat{\mathbf{B}}_{C\mu} = \begin{pmatrix} (1/C) \boldsymbol{I}_{\hat{n}_{C\mu}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} \left((\hat{n}_{C\mu} + 1) \times (\hat{n}_{C\mu} + 1) \right),$$

$$\hat{\mathbf{X}}_{C\mu} = \begin{pmatrix} \hat{\boldsymbol{x}}_{C\mu, j_1} ... \hat{\boldsymbol{x}}_{C\mu, j_{\hat{N}_{C\mu}}} \\ 1 \quad ... \ 1 \end{pmatrix} \left((\hat{n}_{C\mu} + 1) \times \hat{N}_{C\mu} \right),\ \hat{\boldsymbol{x}}_{C\mu, j} = (x_{i_1, j} ... x_{i_{\hat{n}_{C\mu}}, j})^T \in \mathbb{R}^{\hat{n}_{C\mu}}.$$

Соответственно, критерий гипотетической кросс-валидации (39), применительно к выбору μ с заданным C, примет вид

$$\begin{gathered} \hat{\mu}(\boldsymbol{y}, \boldsymbol{X}) = \arg\min_{\mu} \left\{ \sum_{j \in \hat{\mathbb{J}}_{C\mu}} \hat{\xi}_{C\mu, j}^2 + Tr\left[\hat{\mathbf{X}}_{C\mu} \hat{\mathbf{X}}_{C\mu}^T (\hat{\mathbf{X}}_{C\mu} \hat{\mathbf{X}}_{C\mu}^T + \mathbf{B}_{C\mu})^{-1} \right] \right\} \cong \\ \arg\min_{\mu} \left\{ \sum_{j \in \hat{\mathbb{J}}_{C\mu}} \hat{\xi}_{C\mu, j}^2 + \left(\min\{\hat{n}_{C\mu}, \hat{N}_{C\mu}\} + 1 \right) \right\}. \end{gathered} \tag{44}$$

Эксперименты показали, что этот критерий очень хорошо работает на практике.

6 Экспериментальная иллюстрация

6.1 Массив экспериментальных данных

Массив данных Chicken Pieces Silhouettes Database [19] состоит из 446 контурных изображений частей тушки курицы. Каждая часть относится к одной из пяти категорий, которые отражают конкретные части тушки (рис. 1).

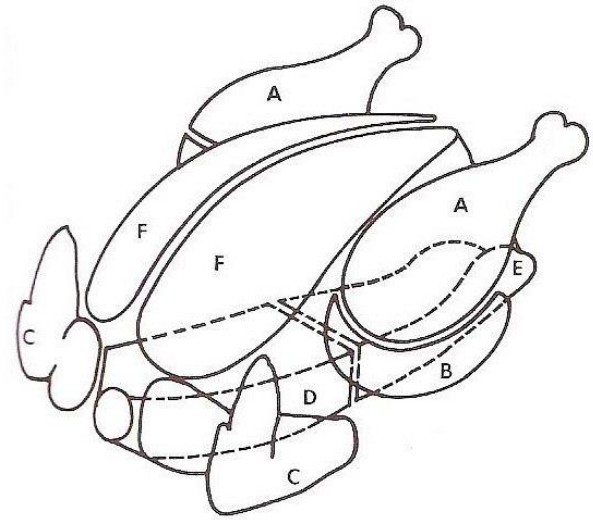

Рисунок 1. Трехмерная разметка частей тушки курицы.

Каждое изображение в бинарном формате представляет собой силуэт конкретного кусочка. Размер и ориентация изображений безразличны. Примеры контурных изображений частей приведены на рис. 2.

(1) Крылышко 117

(2) Спинка 76

(3) Ножка 96

(4) Бедрышко 61

(5) Грудка 96

Рисунок 2. Примеры контурных изображений частей тушки.

На множестве изображений авторами массива данных определены 44 разные числовые функции парного сравнения изображений, представляющие их в компьютере $S_i(\omega',\omega''):\Omega\times\Omega\to\mathbb{R}$, $i=1,...,44$. Каждая функция выражает некоторую меру сходства контуров сравниваемых изображений. Строго говоря, эти функции по своим свойствам являются кернелами, т.е. образуют положительно определенные функции двух переменных. Однако в нашем исследовании это не играет роли, поскольку в версии метода релевантных векторов (7)-(8) в качестве функций парного сравнения могут выступать произвольные функции.

В наших экспериментах мы использовали только два класса изображений, а именно, «спинка» (2) и «ножка» (3) (рис. 2). Таким образом, мы рассматриваем двухклассовую задачу с 172 объектами: $\Omega=\{\omega_j,\ j=1,...,N=172\}$, $y_j=\pm1$. Из всех функций парного сравнения мы исследовали только одну $S(\omega',\omega''):\Omega\times\Omega\to\mathbb{R}$.

6.2 Вторичные признаки объектов и процесс обучения

Мы представили каждый объект ω_j как N-размерный вектор его вторичных признаков, т.е. значений функции его сравнения со всеми элементами обучающей выборки

$$\boldsymbol{x}_j = (x_{j1},\ldots,x_{jN}) = \left(S(\omega_j,\omega_1),\ldots,S(\omega_j,\omega_N)\right) \in \mathbb{R}^N,$$

и решили задачу RVM (8) с большим значением параметра $C>0$ и возрастающими величинами селективности параметра $\mu \geq 0$.

Для каждого μ решение этой задачи $(\hat{a}_{C\mu,1},\ldots,\hat{a}_{C\mu,N},\hat{b}_{C\mu},\hat{\xi}_{C\mu,1},\ldots,\hat{\xi}_{C\mu,N})$ определяет дискриминантную гиперплоскость в соответствующем N-размерном пространстве признаков $\sum_{i=1}^{N}\hat{a}_{C\mu,i}x_i + \hat{b}_{C\mu} \gtrless 0$.

6.3 Экспериментальное сравнение с классическим методом скользящего контроля

Для каждого пробного значения μ процедура выделяет подмножество релевантных вторичных признаков $\hat{\mathbb{J}}_{C\mu}=\{\hat{a}_{C\mu,i}\neq 0\}\subset\mathbb{J}$, а именно, релевантных объектов обучающей выборки.

Чтобы обеспечить некоторую базу для сравнения предложенного метода гипотетической кросс-валидации, мы применили к каждому значению селективности отбора признаков μ также стандартный метод скользящего контроля. Результат показан на рис. 3.

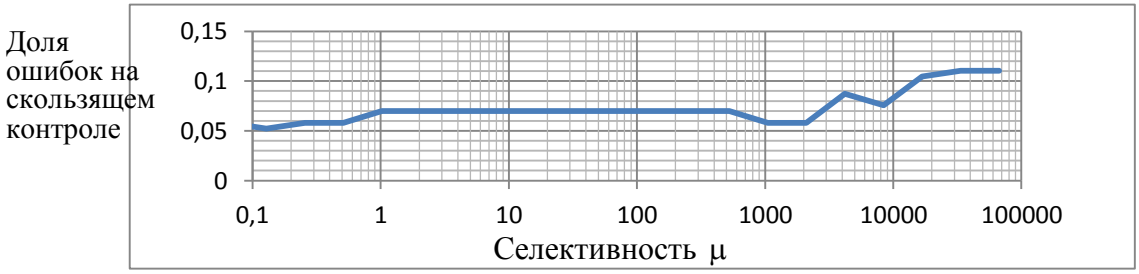

Рисунок 3. Ошибка кросс-валидации по методу скользящего контроля при возрастании селективности.

Как видим, оценка обобщающей способности по методу традиционного скользящего контроля практически не зависит от уровня селективности. Это означает, что стандартный метод SVM без отбора признаков ($\mu=0$) не склонен к переобучению, даже если использовать все огромное множество вторичных признаков. Это происходит потому, что разные вторичные признаки не несут почти никакой дополнительной информации о классе объекта. Однако такое решающее правило крайне неудобно, поскольку новый объект приходится сравнивать со всеми объектами обучающей совокупности.

Далее, мы применили метод беспереборной кросс-валидации в том виде в котором он представлен в разделе 5.3. График критерия $\sum_{j\in\hat{\mathbb{J}}_{C\mu}}\hat{\xi}^2_{C\mu,j}+\left(\min\{\hat{n}_{C\mu},\hat{N}_{C\mu}\}+1\right)$ (44) показан на рис 4. Теперь наблюдается отчетливый минимум структурного риска при $\mu=500$. В этой точке неявная кросс-валидация выделяет подмножество из всего лишь 5-ти релевантных объектов, достаточное для решающего правила классификации новых объектов.

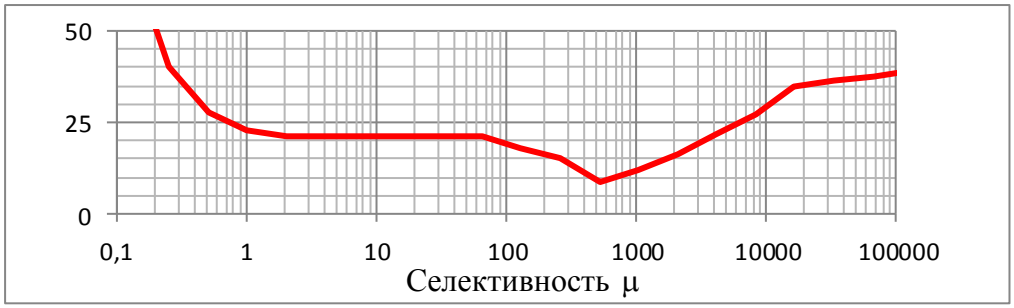

Рисунок 4. Критерий гипотетической кросс-валидации для выбора уровня селективности (44).

7 Заключение

Чтобы избежать многократных повторений, требуемых для традиционной кросс-валидации при варьировании параметров, определяющих структурный риск, мы предложили новый подход, заключающийся в мысленной эмуляции процесса кросс-валидации на основе некоторых естественных предположений о неизвестном распределении вероятностей, порождающем наблюдаемые данные, в частности, обучающую совокупность. Мы назвали этот подход методом бесперебoрной гипотетической кросс-валидации, и показали, что классический информационный критерий Акаике является его частным случаем при некоторых наиболее простых и даже «наивных» предположениях о механизме порождения данных. В приложении к задаче обучения распознаванию образов по методу опорных векторов, известному под аббревиатурой SVM – Support vector Machine, наш подход заключается в добавлении к минимальному значению эмпирического риска, достигнутого в результате обучения, дополнительного штрафа на расширение класса модели данных, величина которого зависит от размерности Вапника-Червоненкиса (VC dimension).

На примере известного массива данных Chicken-pieces Data Set мы экспериментально продемонстрировали эффективность нашего подхода при выборе уровня селективности отбора признаков в дважды регуляризованной версии SVM.

8 Приложения

8.1 Доказательство теоремы 1

Добавим и вычтем значение функции штрафа с оценкой вектора параметров по той же выборке $Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)$. Группируя слагаемые, имеем:

$$\begin{aligned}
&\iiint Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X} = \\
&\iiint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)+Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\times \\
&\qquad\times\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X} = \\
&\iiint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X} + \\
&\iiint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X} = \\
&\iint Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\underbrace{\Big\{\int \Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})}_{F^*(\boldsymbol{X},\boldsymbol{y})}d\boldsymbol{y}\,d\boldsymbol{X} + \\
&\iiint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Big\{\int \Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Big\}G^*(\boldsymbol{X})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\,d\boldsymbol{X} = \\
&\int\limits_{\mathbb{X}\times\mathbb{Y}} Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{y},\boldsymbol{X})\big)F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{X}d\boldsymbol{y} + \\
&\iint \underbrace{\Big\{\iint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}\Big\}}_{\Delta(C,\boldsymbol{X},\boldsymbol{a})}\Psi^*(\boldsymbol{a})d\boldsymbol{a}G^*(\boldsymbol{X})d\boldsymbol{X} = \\
&\int\limits_{\mathbb{X}\times\mathbb{Y}} Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{y},\boldsymbol{X})\big)F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{X}d\boldsymbol{y} + \int\limits_{\mathbb{X}\times\mathbb{R}^m} \Delta(C,\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})G^*(\boldsymbol{X})d\boldsymbol{X}d\boldsymbol{a}.
\end{aligned}$$

Теорема доказана.

8.2 Доказательство теоремы 2

Если значение

$$\Delta(C,\boldsymbol{X},\boldsymbol{a}) = \iint \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}\,d\boldsymbol{y}$$

не зависит от вектора $\boldsymbol{a}$, то есть $\Delta(C,\boldsymbol{X},\boldsymbol{a}) = \Delta(C,\boldsymbol{X})$, то

$$\int\limits_{\mathbb{X}\times\mathbb{R}^m} \Delta(C,\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})G^*(\boldsymbol{X})d\boldsymbol{X}d\boldsymbol{a} = \int\limits_{\mathbb{X}} \Delta(C,\boldsymbol{X})G^*(\boldsymbol{X})d\boldsymbol{X} = \int\limits_{\mathbb{X}\times\mathbb{Y}} \Delta(C,\boldsymbol{X})F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{X}d\boldsymbol{y}.$$

Теорема доказана.

8.3 Доказательство теоремы 3

Еще раз запишем выражение (19) для штрафа за использование оценки $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})$, вычисленной по той же выборке, вместо независимой оценки $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$:

$$\Delta(C,\boldsymbol{X},\boldsymbol{a}) = \int\limits_{\mathbb{Y}\times\mathbb{Y}} \Big[Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big)\Big]\Phi(\tilde{\boldsymbol{y}}\mid\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}.$$

С учетом квадратичности штрафной функции перепишем подынтегральное выражение

$$Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big) = \big\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big\|^2 - \big\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big\|^2.$$

Раскроем квадраты норм:

$$\begin{aligned}
&Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big)-Q\big(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\big) = \big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big\|^2 - \big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})\big\|^2 \\
&-2\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}}) + 2\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y}).
\end{aligned}$$

Заметим, что усреднение по $\tilde{\boldsymbol{y}}$ выражения $\big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big\|^2$ и усреднение по $\boldsymbol{y}$ выражения $\big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})\big\|^2$ совпадают (замена переменной интегрирования):

$$\int\limits_{\mathbb{Y}} \big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\big\|^2\Phi(\tilde{\boldsymbol{y}}\mid\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}} = \int\limits_{\mathbb{Y}} \big\|\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})\big\|^2\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}\ .$$

Значит, вычисление $\Delta(C,\boldsymbol{X},\boldsymbol{a})$ сводится к вычислению интегралов:

$$\Delta(C,\boldsymbol{X},\boldsymbol{a})=2\int_{\mathbb{Y}}\int_{\mathbb{Y}}\left[\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})-\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\right]\Phi(\tilde{\boldsymbol{y}}\mid\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}=$$

$$2\int_{\mathbb{Y}}\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}-2\int_{\mathbb{Y}}\int_{\mathbb{Y}}\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\tilde{\boldsymbol{y}})\Phi(\tilde{\boldsymbol{y}}\mid\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}.$$

Вычислим отдельно каждое слагаемое, учитывая, что $\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})=(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{y}$. Тогда подынтегральное выражение во втором слагаемом имеет вид: $\boldsymbol{y}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\tilde{\boldsymbol{y}}$, так что усреднение по независимым переменным $\boldsymbol{y}$ и $\tilde{\boldsymbol{y}}$ дает:

$$2\left(\int_{\mathbb{Y}}\boldsymbol{y}^T\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}\right)\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\int_{\mathbb{Y}}\tilde{\boldsymbol{y}}^T\Phi(\tilde{\boldsymbol{y}}\mid\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}\right)=$$

$$2\left(\boldsymbol{X}^T\boldsymbol{a}\right)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{a}\right)=2\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a}.$$

Теперь выпишем первое слагаемое:

$$2\int_{\mathbb{Y}}\boldsymbol{y}^T\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\mathbf{X},\boldsymbol{y})\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}=2\int_{\mathbb{Y}}\boldsymbol{y}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{y}\,\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}.$$

Воспользуемся формулой, связывающей второй момент случайных величин и их дисперсию через квадрат математического ожидания:

$$\int_{\mathbb{Y}}\boldsymbol{y}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{y}\,\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}=$$

$$\int_{\mathbb{Y}}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}+\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a}.$$

С учетом второго слагаемого, получаем, что итоговое выражение записывается через математическое ожидание квадратичной формы от центрированного вектора наблюдений:

$$\Delta(C,\boldsymbol{X},\boldsymbol{a})=2\int_{\mathbb{Y}}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}=$$

$$2\int_{\mathbb{Y}}Tr\left[\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)\right]\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}=$$

$$2\int_{\mathbb{Y}}Tr\left[\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)^T\right]\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}=$$

$$2Tr\left\{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\left(\int_{\mathbb{Y}}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right)^T\Phi(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}\right)\right\}=$$

$$2\sigma^2Tr\left\{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\right\}=Tr\left\{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\right\}.$$

Теорема доказана.

8.4 Доказательство теоремы 4

Во-первых, покажем, что для любой блочной симметрической матрицы с квадратными диагональными блоками

$$\left(\begin{array}{c|c}\mathbf{D}_{11}&\mathbf{D}_{12}\\\hline\mathbf{D}_{21}&\mathbf{D}_{22}\end{array}\right)=\left(\begin{array}{c|c}\mathbf{D}_{11}&\mathbf{D}_{12}\\\hline\mathbf{D}_{12}^T&\mathbf{D}_{22}\end{array}\right)$$

справедливо выражение

$$\left(\begin{array}{c|c}\mathbf{D}_{11}&\mathbf{D}_{12}\\\hline\mathbf{D}_{12}^T&\mathbf{D}_{22}\end{array}\right)^{-1}=\left(\begin{array}{c|c}(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}&-(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\\\hline-\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}&\mathbf{D}_{22}^{-1}\left[\mathbf{I}+\mathbf{D}_{12}^T(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\right]\end{array}\right).$$

Действительно, пусть $\left(\begin{array}{c|c}\mathbf{D}_{11}&\mathbf{D}_{12}\\\hline\mathbf{D}_{12}^T&\mathbf{D}_{22}\end{array}\right)^{-1}=\left(\begin{array}{c|c}\mathbf{C}_{11}&\mathbf{C}_{12}\\\hline\mathbf{C}_{12}^T&\mathbf{C}_{22}\end{array}\right)$.

Тогда

$$\left(\begin{array}{c|c}\mathbf{D}_{11} & \mathbf{D}_{12}\\ \hline \mathbf{D}_{12}^T & \mathbf{D}_{22}\end{array}\right)\left(\begin{array}{c|c}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \hline \mathbf{C}_{12}^T & \mathbf{C}_{22}\end{array}\right)=\left(\begin{array}{c|c}\mathbf{I} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{I}\end{array}\right),\ \text{т.е.}\ \begin{cases}\mathbf{D}_{11}\mathbf{C}_{11}+\mathbf{D}_{12}\mathbf{C}_{12}^T=\mathbf{I},\\ \mathbf{D}_{12}^T\mathbf{C}_{11}+\mathbf{D}_{22}\mathbf{C}_{12}^T=\mathbf{0},\end{cases}\ \begin{cases}\mathbf{D}_{11}\mathbf{C}_{12}+\mathbf{D}_{12}\mathbf{C}_{22}=\mathbf{0},\\ \mathbf{D}_{12}^T\mathbf{C}_{12}+\mathbf{D}_{22}\mathbf{C}_{22}=\mathbf{I}.\end{cases}$$

Из второго уравнения левой системы матричных уравнений следует, что $\mathbf{C}_{12}^T=\mathbf{C}_{21}=-\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T\mathbf{C}_{11}$. Подставляя это выражение в первое уравнение системы, получим равенство $\mathbf{C}_{11}=(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}$. Отсюда следует, что $\mathbf{C}_{12}=-(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}$. Из второго уравнения правой системы получим $\mathbf{C}_{22}=\mathbf{D}_{22}^{-1}\left[\mathbf{I}+\mathbf{D}_{12}^T(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\right]$, что и требовалось доказать.

Теперь воспользуемся этой формулой в предположении, что

$$V(\boldsymbol{a},C)=\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a},\ \ \boldsymbol{B}_C=Diag\Big(\underbrace{\frac{1}{\rho}\cdots\frac{1}{\rho}}_{C}\ \underbrace{\rho\cdots\rho}_{n-C}\Big),\ \rho\to\infty.$$

То есть для правого нижнего блока матрицы $\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C$ при $\rho\to\infty$ диагональные элементы стремятся к бесконечности, а значит $\left(\mathbf{D}_{22}\right)^{-1}\to\mathbf{0}$ покомпонентно. Следовательно,

$$\left(\begin{array}{c|c}\mathbf{D}_{11} & \mathbf{D}_{12}\\ \hline \mathbf{D}_{12}^T & \mathbf{D}_{22}\end{array}\right)^{-1}=\left(\begin{array}{c|c}(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1} & -(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\\ \hline -\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1} & \mathbf{D}_{22}^{-1}\left[\mathbf{I}+\mathbf{D}_{12}^T(\mathbf{D}_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}^T)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\right]\end{array}\right)$$

$$\to\left(\begin{array}{c|c}\left(\mathbf{D}_{11}\right)^{-1} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right).$$

Значит, $\lim\limits_{\rho\to\infty}\Delta(C,\boldsymbol{X})=\lim\limits_{\rho\to\infty}Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]=Rank\left[\mathbf{D}_{11}\right]=C$.

Теорема доказана.

8.5 Доказательство теоремы 5

Согласно (37) штраф $\Delta(\hat{\mathbb{I}},C,\boldsymbol{X})$ определяется выражением $\Delta(\hat{\mathbb{I}},C,\boldsymbol{X})=Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\right]$.

Для нахождения обратной матрицы $(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}$ воспользуемся формулой:

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1}=\begin{pmatrix}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\end{pmatrix}.$$

Переписав $\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}$, используя (35) , получим:

$$(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}=\left(\begin{array}{c|c}\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\boldsymbol{I}_n & \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\\ \hline \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T & \hat{N}_C\end{array}\right)^{-1}=$$

$$\begin{pmatrix}\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\mathbf{I}_n-\frac{1}{\hat{N}_C}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1} & -\dfrac{\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\mathbf{I}_n\right]^{-1}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j}{\hat{N}_C-\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\mathbf{I}_n\right]^{-1}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j}\\ -\frac{1}{\hat{N}_C}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\mathbf{I}_n-\frac{1}{\hat{N}_C}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1} & \dfrac{1}{\hat{N}_C-\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T+(1/C)\mathbf{I}_n\right]^{-1}\sum\limits_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j}\end{pmatrix}.$$

Далее умножим эту матрицу слева на матрицу $\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T$, имея в виду, что согласно принятым обозначениям (35)

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T = \left(\begin{array}{c|c} \hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T & \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j \\ \hline \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T & \hat{N}_C \end{array}\right).$$

Заметим, что поскольку нам потребуется взять оператор следа, нам не нужно знать элементы вне диагонали. С учетом этого замечания

$$\begin{array}{c} \\ \hat{n} \\ 1 \end{array}\begin{array}{c} \begin{array}{cc} \hat{n} & 1 \end{array} \\ \left(\begin{array}{c|c} \hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T & \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j \\ \hline \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T & \hat{N}_C \end{array}\right) \end{array}\times$$

$$\left(\begin{array}{c|c} \left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1} & -\frac{\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j}{\hat{N}_C - \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j} \\ \hline -\frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1} & \frac{1}{\hat{N}_C - \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j} \end{array}\right)$$

$$= \left(\begin{array}{c|c} \mathbf{G} & * \\ \hline * & h \end{array}\right),$$

где

$$\mathbf{G}(\hat{n}\times\hat{n}) = \hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1} -$$

$$\frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1},$$

$$h = -\frac{\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j}{\hat{N}_C - \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j} + \frac{\hat{N}_C}{\hat{N}_C - \sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n\right]^{-1}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j} = 1.$$

Таким образом,

$$\Delta(C,\boldsymbol{X}) = Tr\left\{\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1}\right\} -$$

$$\frac{1}{\hat{N}_C}Tr\left\{\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T + (1/C)\mathbf{I}_n - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]^{-1}\right\} + 1 =$$

$$= Tr\left\{\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T + (1/C)\mathbf{I}_n\right]^{-1}\right\} + 1.$$

Итак $\Delta(C,\boldsymbol{X}) = Tr\left\{\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T\right]\left[\hat{\boldsymbol{X}}_C\hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C}\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j\sum_{j\in\hat{\mathbb{J}}_C}\boldsymbol{x}_j^T + (1/C)\mathbf{I}_n\right]^{-1}\right\} + 1.$

Заметим, что

$$\hat{\boldsymbol{X}}_C = (\boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}_C}})(n \times \hat{N}_C),\ \hat{\boldsymbol{X}}_C^T = \begin{pmatrix} \boldsymbol{x}_{j_1}^T \\ \vdots \\ \boldsymbol{x}_{j_{\hat{N}_C}}^T \end{pmatrix} (\hat{N}_C \times n),\ \ \hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T = (\boldsymbol{x}_{j_1} \cdots \boldsymbol{x}_{j_{\hat{N}}}) \begin{pmatrix} \boldsymbol{x}_{j_1}^T \\ \vdots \\ \boldsymbol{x}_{j_{\hat{N}_C}}^T \end{pmatrix} = \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_j^T,$$

$$\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T = \hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \sum_{l \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_l^T = \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_j^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \sum_{l \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_l^T .$$

Тогда

$$Tr\left\{ \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right] \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T + \frac{1}{C} \mathbf{I}_n \right]^{-1} \right\} =$$

$$Tr\left\{ \left[\frac{1}{\hat{N}_C} \hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right] \left[\frac{1}{\hat{N}_C} \hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T + \frac{1}{\hat{N}_C} \frac{1}{C} \mathbf{I}_n \right]^{-1} \right\} =$$

$$Tr\left\{ \underbrace{\left[\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_j^T - \left(\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \right) \left(\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \right)^T \right]}_{\substack{\textit{оценка ковариационной матрицы} \\ \textit{векторов обучающей совокупности} \\ \textit{в пределах опорного подмножества}}} \underbrace{\left[\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \boldsymbol{x}_j^T - \left(\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \right) \left(\frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \right)^T + \frac{1}{\hat{N}_C} \frac{1}{C} \mathbf{I}_n \right]^{-1}}_{\substack{\textit{оценка ковариационной матрицы} \\ \textit{векторов обучающей совокупности} \\ \textit{в пределах опорного подмножества}}} \right\}.$$

Далее,

$$Tr\left\{ \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right] \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T + (1/C) \mathbf{I}_n \right]^{-1} \right\}$$

$$\underset{C \to \infty}{\to} Tr\left\{ \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right] \left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right]^{-1} \right\} =$$

$$Rank\left[\hat{\boldsymbol{X}}_C \hat{\boldsymbol{X}}_C^T - \frac{1}{\hat{N}_C} \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j \sum_{j \in \hat{\mathbb{I}}_C} \boldsymbol{x}_j^T \right] = \min\{\hat{n}, \hat{N}_{\hat{\mathbb{I}},\mathrm{C}} - 1\}.$$

Таким образом,

$$\lim_{C \to \infty} \Delta(\hat{\mathbb{I}}, C, \boldsymbol{X}) = \min\{\hat{n}, \hat{N}_{\hat{\mathbb{I}},\mathrm{C}} - 1\} + 1.$$

Теорема доказана.

Литература

1. P.A. Devijver, J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, GB, 1982.
2. Бонгард М.М., Вайнцвайг М.Н. Об оценках ожидаемого качества признаков. *Проблемы кибернетики*, 1968, вып. 20, с. 151-157.
3. Черноусова Е.О. *Бесперебoрные методы кросс-валидации для оценивания обобщающей способности регрессионных моделей*. Диссертация к.ф.-м. н. Вычислительный центр РАН, 2013.
4. Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716-723.
5. C. Cortes, V. Vapnik. Support-Vector Networks. *Machine Learning*, 1995, 20, pp. 273-297.
6. V. Vapnik. *Statistical Learning Theory*. John-Wiley & Sons, Inc., 1998, 736 p.
7. Вапник В.Н., Червоненкис А.Я. *Теория распознавания образов (статистические проблемы обучения)*. М.: Наука, 1974.
8. John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in large margine classifiers*, 1999, pp. 61-74, MIT Press.
9. Peter Sollich. Probabilistic methods for Support Vector Machines. *Advances in Neural Information Processing Systems*, 12, 2000, pp. 349-355, MIT Press.
10. Татарчук А.И. *Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков*. Диссертация к.ф.-м. н. Вычислительный центр РАН, 2014.
11 C. Cortes, V. Vapnik. Support-Vector Networks. *Machine Learning*, 1995, 20, pp. 273-297.
12 V. Vapnik. *Statistical Learning Theory*. John-Wiley & Sons, Inc., 1998, 736 p.
13 A. Bab-Hadiashar, D. Suter. *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*. Springer Verlag, New York, Inc., 2000.
14 H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716-723.
15 R. Duin, E. Pekalska, D. de Ridder. Relational discriminant analysis. *Pattern Recognition Letters*, Vol. 20, 1999, pp. 1175-1181.
16 C. Bishop, M. Tipping. Variational Relevance Vector Machines. Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pp. 46-53. Morgan Kaufmann, 2000.
17 L. Wang, J. Zhu, H. Zou. The doubly regularized support vector machine. Statistica Sinica, Vol. 16, 2006, pp. 589-615.
18 O. Seredin, V. Mottl, A. Tatarchuk, N. Razin, D. Windridge. Convex Support and Relevance Vector Machines for selective multimodal pattern recognition. *Proceedings of the 21th International Conference on Pattern Recognition*, Tsukuba, Japan, November 11-15, 2012. IAPR, 2012, ISSN 978-4-9906441-1-6, 2012, pp. 1647-1650.
19 G. Andreu, A. Crespo, J.M. Valiente. Selecting the Toroidal Self-Organizing Feature Maps (TSOFM) Best Organized to Object Recognition. *Proceedings of ICNN´97*, vol. 2, pp. 1341-1346, Houston, Texas (USA). IEEE. June, 1997.