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»

Мотренко Анастасия Петровна

Оценка объема выборки в задачах прогнозирования

010656 — Математические и информационные технологии

Выпускная квалификационная работа магистра

Научный руководитель:
к. ф.-м. н. Стрижов Вадим Викторович

Москва
2014

Содержание

Аннотация

В работе исследуется применение порождающих и разделяющих моделей классификации в условиях малых выборок. Для оценки необходимого объема выборки предлагается сравнивать эмпирические распределения полученные по подвыборкам, используя расстояние Кульбака-Лейблера между гистограммами, построенными по этим подвыборкам. Вводится критерий принадлежности двух выборок одному распределению, основанный на расстоянии Кульбака-Лейблера между гистограммами. Приводится теоретическое и практическое исследование предложенного критерия. Решается задача оценки объема выборки при порождающем и разделяющем подходах.

Ключевые слова: *плотность совместного распределения дискретной и непрерывной случайных величин, порождающие модели классификации, разделяющие модели классификации, выборка малого объема, эмпирическая функция распределения, расстояние между гистограммами, расстояние Кульбака-Лейблера, задача двух выборок.*

1 Введение

В задачах классификации требуется, по набору известных неслучайных величин, определить метку класса зависящей от них случайной величины. Для этого строится параметрическая модель, параметры которой оцениваются по обучающей выборке прецедентов. В случае, когда процедура оценки параметров включает восстановление функции условного распределения переменной класса при заданном значении независимой переменной, модель называется разделяющей. Зная это распределение, мы можем установить наиболее вероятное значение зависимой переменной. Альтернатива разделяющему подходу — это порождающий подход. При порождающем подходе вначале оценивается функция совместного распределения зависимых и независимых переменных. Затем, с помощью формулы Байеса, выводится условное распределение зависимой переменной. В работах [1, 2] для решения задач классификации используется разделяющий подход. Публикации [4, 5, 6, 7], посвящены сравнению разделяющих и порождающих алгоритмов. Исследование асимптотического поведения функции ошибки прогнозирования при использовании каждого из подходов, приведено в [4]. В частности, показано, что разделяющий подход доставляет большую точность прогнозирования при неограниченном увеличении объема выборки. В [5] авторы впервые представили исследование этих алгоритмов в неасимптотическом случае, получив теоретически и экспериментально следующий результат: хотя асимптотическая ошибка порождающих алготритмов классификации больше, чем разделяющих, порождающие алгоритмы быстрее приближаются к асимптоте. Этот результат означает, что при небольших объемах выборки следует применять порождающий подход. Позднее, были опубликованы другие теоретические исследования на эту тему [6, 7], а также идеи по комбинированию порождающего и разделяющего подходов [9, 8, 10, 11]. Основная идея комбинированного подхода такова: поскольку не всегда можно выделить однозначно лучший подход, попробуем использовать оба подхода для получения наибольшей обобщающей способности.

Поскольку строгого критерия выбора между порождающим и разделяющим подходами не найдено, в данной работе также рассмотрена комбинация разделяющего и порождающего подходов. Комбинированный подход заключается в использовании выпуклой линейной комбинации функций условного и совместного правдоподобия как функционала качества. Ставятся и решаются задачи оценки параметров условного и совместного распределений, а также коэффициентов линейной комбина-

ции в зависимости от объема выборки.

Для сравнения подходов ставится задача оценки необходимого объема выборки в рамках каждого их подходов. Предлагается считать объем выборки достаточным, если выборки данного объема из рассматриваемого распределения будут отнсены некоторым критерием к одному распределению. В данной работе для оценки близости эмпирических распределений двух выборок используется расстояние Кульбака-Лейблера между гистограммами этих выборок.

В литературе по математической статистике вводится множество коэффициентов, показывающих что некоторые два распределения P и Q близки друг к другу. Такие коэффициенты в различных источниках называются расстоянием между распределениями [14], мерами разделяющей информации [15], мерами статистического расстояния [16]. В работе [17] описан метод порождения коэффициентов $d(P, Q)$ «непохожести» двух распределений, обладающих некоторыми стандартными свойствами, например:

1) коэффициент $d(P, Q)$ должен быть определен на всех парах распределений с одним носителем;

2) значение $d(P, Q)$ должно быть минимально при $P = Q$;

3) при любом измеримом преобразовании носителя распределений P и Q расстояние между ними не увеличивается.

Идея метода [17] заключается в том, чтобы рассмотреть различные выпуклые функции случайной величины Q/P. С точки зрения распределения P, матожидание Q/P независимо от Q, а дисперсия стремится к нулю при $Q \to P$. Так же в [17] показано, что многие известные функции расстояния могут быть получены этим методом. В частности, им могут быть порождены все f-дивергенции [18] и в том числе расстояние Кульбака-Лейблера [14]. В работе [19] приведено сравнение многих известных расстояний с точки зрения скорости сходимости эмпирического распределения к истинному, а также качественного поведения функции расстояния при сходимости. При решении задачи кластеризации в обработке изображений, были введены меры [20], [21], основанные на метрике

В данной работе решается задача выбора между порождающей и разделяющей гипотезами порождения данных при оценке необходимого объема выборки. Предложен метод выбора модели классификации, основанный на сравнении параметров

функций условного и совместного распределения переменной класса и независимых переменных. Решается задача оценки необходимого объема выборки при порождающем и разделяющем подходах с использованием расстояния Кульбака-Лейблера между гистограммами подвыборок в качестве показателя близости их распределений. Показано, что распределение расстояние Кульбака-Лейблера между гистограммами из одного распределения в пределе ограничено сверху распределением χ^2. Предложен критерий для решения задачи двух выборок, основанный на расстоянии Кульбака-Лейблера между их гистограммами. Продемонстрировано применение критерия к решению задачи двух выборок для различных пар распределений и показана его состоятельность.

В качестве примера рассмотрим задачу прогнозирования инфаркта. Заболевания сердечно-сосудистой системы могут протекать, не проявляясь клинически. Тем не менее, обнаружение нарушений, связанных с работой сердца, по косвенным признакам вполне возможно [12]. В данной работе в качестве признаков (биомаркеров) используются концентрации белков и их соединений, абсорбированные на поверхности кровяных телец. Разделение пациентов на две группы по состоянию здоровья приводит к задаче классификации. Сложность решаемой задачи состоит в малом объеме выборки, приводящем к переобучению модели.

2 Постановка задачи

Рассмотрим выборку вида $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, где $\mathbf{x}_i \in \mathbb{R}^n$ — описание i-го элемента выборки, а $y_i \in \{0, 1\}$ — его метка класса. Введем обозначение $D = \{\mathbf{X}, \mathbf{y}\}$, где $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$, $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal$ — матрица плана. Будем называть *гипотезой порождения данных* предположения о виде распределения элементов выборки. Определим также модель классификации. *Моделью* f назовем параметрическое множество функций $f : (\boldsymbol{\beta} \times \mathbf{x}) \mapsto [0, 1]$, $\boldsymbol{\beta} \sim \mathbb{R}^{n+1}$. При фиксированном векторе параметров $\boldsymbol{\beta}$ модель f порождает некоторый классификатор:

$$a_f(\mathbf{x}) = [f(\mathbf{x}, \boldsymbol{\beta}) > 0] \in \{0, 1\}. \tag{2.1}$$

Согласно разделяющей модели, выбираются параметры $\boldsymbol{\beta}_{\text{D}}$, максимизирующие условное правдопобие данных

$$\boldsymbol{\beta}_{\text{D}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^{n+1}} L_{\text{D}}(\boldsymbol{\beta}), \text{ где } \; L_{\text{D}} = \ln p(\mathbf{y}|\mathbf{X}). \tag{2.2}$$

Распредение свободных переменных $\mathbf{X}$ при этом не учитывается. В качестве разделяющей модели будем рассматривать логистическую регрессию, в рамках которой условное распределение переменной метки класса y считается бернуллевским $y|\mathbf{x} \sim \mathcal{B}(\theta)$. Гипотеза порождения данных в логистической регрессии имеет следующий вид:

$$y|\mathbf{x} \sim \mathcal{B}(\theta_i), \quad \text{где } \theta = p(y|\mathbf{x}). \tag{2.3}$$

Параметры $\boldsymbol{\beta}_{\text{G}}$ порождающей модели максимизируют совместное правдоподобие свободной переменной $\mathbf{x}$ и метки класса y

$$\boldsymbol{\beta}_{\text{G}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^{n+1}} L_{\text{G}}(\boldsymbol{\beta}), \text{ где } \quad L_{\text{G}} = \ln p(\mathbf{y}, \mathbf{X}). \tag{2.4}$$

В качестве порождающей модели рассмотрим линейный дискриминантный анализ. Согласно линейному дискриминантному анализу, предполагается, что переменная метки класса y имеет распределение Бернулли, а вектор $\mathbf{x}$ распределен нормально, с матожиданием $\boldsymbol{\mu}_y$, зависящим от метки класса y и ковариационной матрицей Σ.

$$y \sim \mathcal{B}(P), \quad \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma). \tag{2.5}$$

Обозначим построенные по некоторой выборке D классификаторы, соответствующие логистической регресии и линейному дискриминантному анализу как a_{D} и a_{G}. Ниже показано, как оцениваются параметры каждой из моделей, а также установлена связь между ними.

2.1 Оценка параметров разделяющей модели

В логистической регрессии условная вероятность принадлежности объекта классу $y_i = 1$ определяется логистической функцией:

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^{\intercal}\mathbf{w} - c)} = \theta_i, \tag{2.6}$$

где $\boldsymbol{\beta} = [\mathbf{w}^{\intercal}, c]^{\intercal}$, $\mathbf{w} \in \mathbb{R}^n$ — вектор параметров модели, полученный решением задачи (2.2).

Будем считать, что объекты выборки порождаются независимо друг от друга. Тогда правдоподобие L_{D} данных представимо в виде произведения

$$p(\mathbf{y}|X) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i) = \prod_{i=1}^{m} \theta_i^{y_i}(1 - \theta_i)^{(1-y_i)},$$

что позволяет выписать разделяющую функцию правдоподобия

$$L_{\text{D}} = \sum_{i=1}^{m} y_i \ln \theta_i + (1 - y_i) \ln(1 - \theta_i).$$

2.2 Оценка параметров порождающей модели

Оценка параметров $\boldsymbol{\beta}_{\text{G}}$ порождающей модели получается путем максимизации функции (2.4) совместного правдоподобия данных $L_{\text{G}} = \ln p(\mathbf{x}, y)$. Чтобы задать функцию плотности совместного распределения $p(\mathbf{x}, y)$ признаков объекта и его метки класса, воспользуемся определением условной вероятности и представим функцию $p(\mathbf{x}, y)$ в виде произведения

$$p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x}) = p(y|\mathbf{x})(P(1)p(\mathbf{x}|1) + P(0)p(\mathbf{x}|0)), \tag{2.7}$$

где $P(1)$ и $P(0)$ — априорные верояности классов, и выполнятеся $P(0) + P(1) = 1$. Пусть $P(1) = P$, тогда $P(0) = 1 - P$. Согласно гипотезе (2.5), условная плотность $p(\mathbf{x}|y)$ вектора свободных переменных при фиксированной метке класса нормальна

$$p(\mathbf{x}|y) = \frac{1}{\sqrt{(2\pi)^n \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^{\intercal} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_y)\right). \tag{2.8}$$

Чтобы вывести функцию $p(y|\mathbf{x})$, рассмотрим отношение функций апостериорных вероятностей метки класса:

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{p(\mathbf{x}, 1)}{p(\mathbf{x})} \frac{p(\mathbf{x})}{p(\mathbf{x}, 0)} = \frac{P p(\mathbf{x}|1)}{(1 - P) p(\mathbf{x}|0)}.$$

Учитывая вид (2.8) функции плотности $p(\mathbf{x}|y)$, получим

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{P}{1 - P} \exp\left(\mathbf{w}^{\intercal}\mathbf{x} + \tilde{c}\right) = \exp\left(\mathbf{w}^{\intercal}\mathbf{x} + c\right),$$

где параметры $c = \tilde{c} + \log(P/1 - P)$ и $\mathbf{w}$ выражаются через параметры нормального распределения следующим образом:

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{2.9}$$

$$c = \log \frac{P}{1 - P} + \tilde{c} = -\frac{1}{2}\left(\boldsymbol{\mu}_0^{\intercal} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^{\intercal} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1\right) = \log \frac{P}{1 - P} - \frac{1}{2} \mathbf{w}^{\intercal}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0). \tag{2.10}$$

Учитывая тождество $p(1|\mathbf{x}) + p(0|\mathbf{x}) = 1$, получим функцию $p(y|\mathbf{x})$ условного распределения метки класса y:

$$p(y|\mathbf{x}) = \left(1 + \left(\frac{1 - P}{P}\right)^{2y-1} \exp\left(-(2y - 1)(\mathbf{w}^{\intercal}\mathbf{x} + \tilde{c})\right)\right)^{-1}, \; y \in \{0, 1\}.$$

В частности,

$$p(y=1|\mathbf{x}) = \left(1 + \left(\frac{1-P}{P}\right)\exp\left(-\mathbf{w}^{\mathsf{T}}\mathbf{x} - \tilde{c}\right)\right)^{-1} = \frac{1}{1+\exp\left(-\mathbf{w}^{\mathsf{T}}\mathbf{x} - \tilde{c} + \ln\left(\frac{1-P}{P}\right)\right)},$$

где вектор $[\mathbf{w}_{\text{G}}^{\mathsf{T}}, c_{\text{G}}]^{\mathsf{T}} = [\mathbf{w}_{\text{G}}^{\mathsf{T}}, \tilde{c} - \ln\left(\frac{1-P}{P}\right)]^{\mathsf{T}}$ вычислен в соответствии с (2.9), (2.10), а параметры $(\boldsymbol{\mu}, \Sigma) = \boldsymbol{\beta}$ получены решением задачи (2.4).

3 Ошибка прогнозирования

Для сравнения моделей (2.2) и (2.4) будем рассматривать вероятность $\varepsilon(a)$ ошибочного прогноза классификатора (2.1) в рамках каждой из моделей

$$\varepsilon(a) = \text{P}(a(\mathbf{x}) \neq y), \tag{3.1}$$

где вектор $\mathbf{x}$ распределен согласно (2.8), причем матрица $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2)$. Заметим, что классификаторы a_{D} и a_{G}, порожденные обеими моделями, могут быть представлены в виде:

$$a(y) = [\mathbf{w}^{\mathsf{T}}\mathbf{x} + c \geqslant 0].$$

Тогда вероятность $\varepsilon(a_{\text{D}})$ ошибки разделяющей модели выражается следующим образом:

$$\varepsilon(a_{\text{D}}) = P \cdot \text{P}(\mathbf{w}_{\text{D}}^{\mathsf{T}}\mathbf{x} + c_{\text{D}} \geqslant 0 | y=0) + (1-P)\text{P}(\mathbf{w}_{\text{D}}^{\mathsf{T}}\mathbf{x} + c_{\text{D}} < 0 | y=1) =$$

$$= 1 + Z \int\limits_{\mathbf{w}_{\text{D}}^{\mathsf{T}}\mathbf{x} + c_{\text{D}} < 0} \left(P\exp\left(-\sum_{j=1}^{n}\frac{(\mathbf{x}-\boldsymbol{\mu}_1)_j^2}{2\sigma_j^2}\right) - (1-P)\exp\left(-\sum_{j=1}^{n}\frac{(\mathbf{x}-\boldsymbol{\mu}_0)_j^2}{2\sigma^2}\right)\right) dx = \tag{3.2}$$

$$= 1 + Z \int\limits_{\mathbf{w}_{\text{D}}^{\mathsf{T}}\mathbf{x} + c_{\text{D}} < 0} E(\mathbf{x})\mathbf{dx}. \tag{3.3}$$

Аналогично, для вероятности $\varepsilon(a_{\text{G}})$ ошибки линейного дискриминантного анализа получим

$$\varepsilon(a_{\text{G}}) = 1 + Z \int\limits_{\mathbf{w}_{\text{G}}^{\mathsf{T}}\mathbf{x} + c_{\text{G}} < 0} E(\mathbf{x})\mathbf{dx}. \tag{3.4}$$

Определим области знакопостоянства функции $E(\mathbf{x})$. Учитывая тождество

$$E(x) > 0 \Leftrightarrow \frac{P}{1-P}\exp\left(\sum_{j=1}^{n}\frac{(\mathbf{x}-\boldsymbol{\mu}_0)_j^2 - (\mathbf{x}-\boldsymbol{\mu}_1)_j^2}{2\sigma_j^2}\right) > 1,$$

получаем условие

$$E(\mathbf{x}) > 0 \Leftrightarrow \sum_{j=1}^{n}\frac{x_j(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_j}{\sigma_j^2} > \frac{1}{2}\sum_{j=1}^{n}\frac{(\boldsymbol{\mu}_1)_j^2 - (\boldsymbol{\mu}_0)_j^2}{2\sigma_j^2} + \ln\frac{1-P}{P} \Leftrightarrow \mathbf{w}_{\text{G}}^{\mathsf{T}}\mathbf{x} + c_{\text{G}} > 0, \tag{3.5}$$

где параметры $\mathbf{w}_{\text{G}}$ и c_{G} получены с помощью (2.9), (2.10) по истинным значениям параметров нормального распределения.

В этом случае $\mathbf{w} = \mathbf{w}_{\text{D}} = \mathbf{w}_{\text{G}}$, $\tilde{c} = c_{\text{D}} = c_{\text{G}} - \ln\frac{1-P}{P}$, и области интегрирования (3.3) и (3.4) совпадают.

Generative VS Discriminative. Рассмотрим вероятность ошибки $\varepsilon(a)$ при каждом из подходов, когда объем выборки стремится к бесконечности. В этом случае настроенные параметры близки к истинным и, согласно (3.5) область интегрирования в (3.4) практически совпадает с областью значений $\mathbf{x}$, для которых $E(\mathbf{x}) < 0$. Кроме того, учитывая постановку задач (2.2), (2.4) и вид оптимизируемых функционалов L_{D}, L_{G}, имеем на обучающей выборке:

$$\prod_{i=1}^{m} \frac{1}{1+\exp\bigl(-(2y_i-1)(\mathbf{w}_{\text{D}}^{\intercal}\mathbf{x}_i + c_{\text{D}})\bigr)} > \prod_{i=1}^{m} \frac{1}{1+\exp\bigl(-(2y_i-1)(\mathbf{w}_{\text{G}}^{\intercal}\mathbf{x}_i + c_{\text{G}})\bigr)},$$

что означает, что с достаточно большой вероятностью на обучающей выборке выполнено

$$\mathbf{w}_{\text{D}}^{\intercal}\mathbf{x}_i + c_{\text{D}} < \mathbf{w}_{\text{G}}^{\intercal}\mathbf{x}_i + c_{\text{G}} \quad \text{для } y_i = 1,$$

$$\mathbf{w}_{\text{D}}^{\intercal}\mathbf{x}_i + c_{\text{D}} > \mathbf{w}_{\text{G}}^{\intercal}\mathbf{x}_i + c_{\text{G}} \quad \text{для } y_i = 0,$$

то есть при всех правильных решениях порождающего классификатора, разделяющий классификатор так же даст правильный ответ. Следовательно, при $m \to \infty$ имеем $\varepsilon_{\text{D}} < \varepsilon_{\text{G}}$.

Рассмотрим другой предельный случай: выборка состоит из одного объекта (пусть это объект класса “1”). В этом случае параметры w_{D} разделяющего классификатора будут неограниченно велики в соответствии с решением задачи $\mathbf{w}^{\intercal}\mathbf{x} + c \to \infty$, эквивалентной в данном случае задаче (2.2). В то же время, множитель $(Pp(\mathbf{x}|1) + (1-P)p(\mathbf{x}|0))$ в определении функционала L_{G} выступает в роли регуляризатора и не дает параметрам уйти на бесконечность, поэтому при $m \approx 1$ имеем $\varepsilon_{\text{G}} < \varepsilon_{\text{D}}$.

4 Оценка объема выборки с использованием расстояния Кульбака-Лейблера

Предлагаемый подход основан на наблюдении за изменением параметров $\mathbf{w}$ регрессионной модели при изменении состава выборки. Рассмотрим некоторое множе-

ство индексов объектов $\mathcal{B}_1 \in \mathcal{J}$, а также множество $\mathcal{B}_2 \in \mathcal{J}$, такое что:

$$|\mathcal{B}_1 \setminus \mathcal{B}_2 \cup \mathcal{B}_2 \setminus \mathcal{B}_1| \leqslant 2.$$

Таким образом, множество $\mathcal{B}_2$ может быть получено из $\mathcal{B}_1$ путем удаления, добавления или замены одного элемента. Оценивая параметры на различных подвыборках, будем получать различные результаты. На рисунке 1 продемонстрировано, как меняется положение разделяющей гиперплоскости, определяемой выражением

$$\mathbf{w}^\intercal \mathbf{x} + c \geqslant \lambda \ln \frac{1-P}{P}$$

при добавлении в выборку двух элементов. Если объем выборки $D_{\mathcal{B}_1}$ достаточно ве-

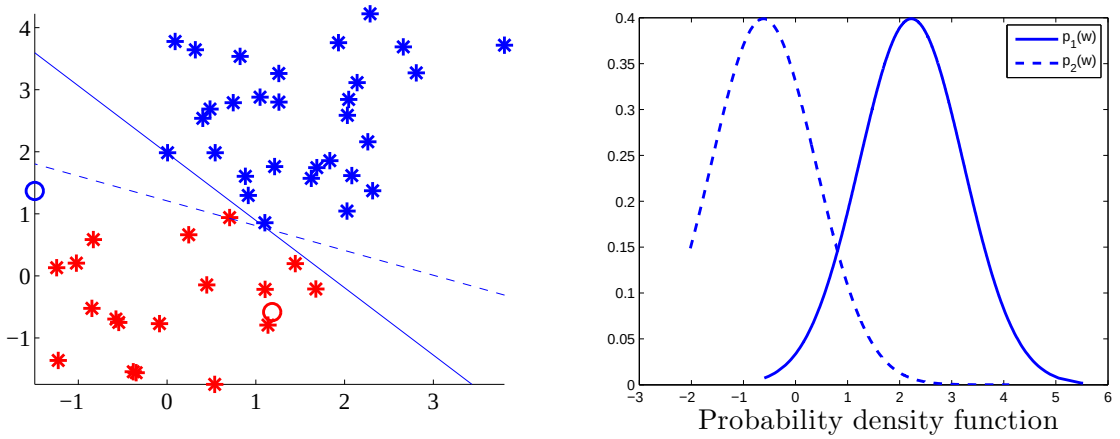


Рис. 1: Два класса, разделенные гиперплоскостью. Пунктирной линией обозначено положение гиперплоскости после того как два новых объекта (выделенных окружностями), были добавлены в выборку.

лик, небольшое изменение ее состава $D_{\mathcal{B}_2}$ не должно приводить к существенному изменению параметров модели. Простейший способ сравнивать параметры на различных подвыборках — с помощью

$$||\mathbf{w}_1 - \mathbf{w}_2|| = \sqrt{\sum_{i=1}^{|\mathcal{A}|} \left(w_i^1 - w_i^2\right)^2}.$$

Предлагается сравнивать функции распределения параметров модели на подвыборках $D_{\mathcal{B}_1}$ и $D_{\mathcal{B}_2}$ с помощью расстояния Кульбака-Лейблера. Опишем процедуру оценки объема выборки $D = (X, \mathbf{y}) = \{z_i\}$.

Пусть объем подвыборок Z и Z' выборки D равен m и m' соответственно. Разобьем область значений случайной величины из P на N промежутков $[a_i, a_{i+1}]$, и обозначим $p_i = P(a_i < z \leqslant a_{i+1})$ вероятность случайной величине с распределением P

принять значение из i-го промежутка; n_i и n'_i — количество объектов выборок Z и Z', попавших в i-тый промежуток. Обозначим $\hat{P}_m$ гистограмму, построенную по выборке Z объема m из распределения P. Гистограмма $\hat{P}_m$ задается набором оценок:

$$\hat{P}_m(a_i < z \leqslant a_{i+1}) = \frac{n_i}{m} = \hat{p}_i, \quad i = 1, \ldots, N-1. \tag{4.1}$$

вероятности p_i. Для выборки D будем рассматривать среднее расстояние Кульбака-Лейблера $D_{\mathrm{KL}}(P_m||P'_m)$ между подвыборками Z и Z' одинакового объема. Минимальный объем выборки m^*, при котором расстояние $D_{\mathrm{KL}}(P_m||P'_m)$ становится меньше некоторого порога $\bar{t}(m)$, будем считать искомой оценкой необходимого объема выборки. О задании порога $\bar{t}(m)$ будет рассказано в разделе 6.

5 Статистическая значимость расстояния Кульбака-Лейблера

Чтобы оценка объема выборки на основе расстояния Кульбака-Лейблера обладает статистической значимостью, необходимо исследовать распределение расстояния Кульбака-Лейблера между гистограммами, построенными по выборкам X и X' из одного распределения P. В данном разделе будет показано, что хотя расстояние Кульбака-Лейблера не имеет предельного распределения, для него можно получить предельные оценки сверху.

Пусть пока выборки X и X' имеют одинаковый объем m. Рассмотрим расстояние $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ между гистограммами $\hat{P}_m$ и $\hat{P}'_m$. По определению, расстояние Кульбака-Лейблера $D_{\mathrm{KL}}(Q||P)$ между распределениями Q и P равно

$$D_{\mathrm{KL}}(Q||P) = \int P \cdot f\left(\frac{Q}{P}\right), \tag{5.1}$$

где $f(t) = t \ln t$. Функция f строго выпукла и дважды дифференцируема в единице, и, повторяя рассуждения из [18], разложим подынтегральное выражение из правой части (5.1) по f в окрестности единицы:

$$P(x) f\left(\frac{Q(x)}{P(x)}\right) = f(1) + f'(1)(Q(x) - P(x)) + \frac{f''(1)}{2}\frac{(Q(x)-P(x))^2}{P(x)} + P(x) o\left(\left(\frac{Q(x)}{P(x)} - 1\right)^3\right),$$

где $f(1) = 0,\ f''(1) = 1$. Подставив вместо Q распределение $\hat{P}_m$, определяемое (4.1), и просуммировав по i, получим соотношение

$$D_{\mathrm{KL}}(\hat{P}_m||P) = \sum_{i=1}^{N} p_i f\left(\frac{\hat{p}_i}{p_i}\right) =$$

$$= \frac{1}{2}\sum_{i}^{N}\frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{i=1}^{N} p_i \cdot \varepsilon\left(\left(\frac{\hat{p}_i}{p_i} - 1\right)^3\right) \sim \frac{1}{2m}\sum_{i}^{N}\frac{(n_i - mp_i)^2}{mp_i}.$$

и следующий предельный переход:

$$2m \cdot D_{\mathrm{KL}}(\hat{P}_m||Q) \sim m\sum_{i=1}^{N}\frac{(\hat{p}_i - p_i)^2}{p_i)} = \sum_{i=1}^{N}\frac{(n_i - mp_i)^2}{mp_i} \to \chi^2_N \ \text{ при } m \to \infty. \qquad (5.2)$$

Докажем следующую теорему:

Теорема 5.1. *Случайная величина* $2m \cdot D_{\mathrm{KL}}(Q||\hat{P}_m) \to \chi^2_N$ *по распределению при* $m \to \infty$.

Доказательство.

Аналогично доказательству предельного перехода (5.2), разложим $D_{\mathrm{KL}}(Q||\hat{P}_m)$ по степеням $f(t)$ вблизи единицы и получим

$$D_{\mathrm{KL}}(Q||\hat{P}_m) \sim \frac{1}{2}\sum_{i=1}^{N}\frac{(\hat{P}_m(\xi_i) - Q(\xi_i))^2}{\hat{P}_m(\xi)} = \frac{1}{2m}\sum_{i=1}^{N}\frac{(n_i - mp_i)^2}{n_i}.$$

Пусть $G_m(x)$ — функция распределения случайной величины $\sum_{i=1}^{N}\frac{(n_i-mp_i)^2}{mp_i}$, $F_m(x)$ — случайной величины $\sum_{i=1}^{N}\frac{(n_i-mp_i)^2}{n_i}$. Так как $G_m(x)$ сходится поточечно к $F_{\chi^2_{N-1}}$ при $m \to \infty$, имеем

$$|G_m(x) - F_{\chi^2_{N-1}}| < \frac{\epsilon}{2} \quad \forall m > m'.$$

Покажем, $|G_m(x) - F_m(x)| \to 0$ при $m \to \infty$. Для этого покажем, что $\forall \varepsilon > 0$ найдется объем выборки m_0 такой, что для всех $m > m_0$ выполняется

$$\mathrm{P}\left(\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leqslant \frac{\varepsilon}{N}\right) > 1 - \varepsilon. \qquad (5.3)$$

Согласно центральной предельной теореме,

$$\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}} \to \mathcal{N}(0,1) \ \text{ по распределению при } m \to \infty,$$

причем для скорости сходимости имеет место неравенство Берри-Эссена:

$$|Q_m(x) - \Phi(x)| \leqslant \frac{A}{\sqrt{m}},$$

где $Q_m(x)$ — функция распределения величины $\frac{n_i-mp_i}{p_i(1-p_i)\sqrt{m}}$, $\Phi(x)$ — функция стандартного нормального распределения, A — некоторая константа. Тогда вероятность

$$\mathrm{P}\left(\left|\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}\right| < C\right) = Q_m(C) - Q_m(-C) \geqslant 2\Phi(C) - 1 - \frac{2A}{\sqrt{m}}. \qquad (5.4)$$

Пусть, кроме того, выполняется $0 < 1 - p \leqslant p_i \leqslant p < 1$. Тогда, с вероятностью $\mathrm{P}_C \geqslant 2\Phi(C) - 1$ выполняется

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| = \frac{|n_i - mp_i|^3}{mn_ip_i} \leqslant \frac{C^3(1-p_i)^3p_i^2}{n_i}\sqrt{m} \leqslant \frac{C^3(1-p)^3p}{\sqrt{m} - C(1-p)}.$$

Обозначим $m_1 = [4C^2(1-p)^2]$, тогда при $m > m_1$ имеет место $\sqrt{m} - C(1-p_i) > \frac{1}{2}\sqrt{m}$ и

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leqslant \frac{2C^3(1-p_i)^3p_i}{\sqrt{m}}.$$

Тогда, для фиксированного ε определим

$$C_\varepsilon = \frac{\epsilon^{1/3}m^{1/6}}{(1-p_i)(2p_iN)^{1/3}},\ \mathrm{P}_m(\varepsilon) = 2\Phi(C_\varepsilon) - 1 - \frac{2A}{\sqrt{m}}.$$

При заданном ε вероятность $\mathrm{P}_m(\varepsilon) \to 1$ при $m \to \infty$, поэтому найдется m_2 такое что для любого $m > m_2$ выполнено $\mathrm{P}_m(\varepsilon) > 1 - \varepsilon$. Выбрав $m_0 = \max(m_1, m_2)$, получим утверждение (5.3). Тогда

$$\left|\sum_{i=1}^{N}\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| \leqslant \sum_{i=1}^{N}\left|\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| < \varepsilon \quad \text{при } m > m_0.$$

Из только что доказанного следует, что $|F_m(x) - G_m(x)| \to 0$ при $m \to 0$. Тогда, $\forall \epsilon > 0\ \exists m'' :$ при $m > m''$ выполняется

$$|F_m(x) - F_{\chi^2_{N-1}}| < |F_m(x) - G_m(x)| + |G_m(x) - F_{\chi^2_{N-1}}| < \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

■

Доказанная теорема и утверждение (5.2) позволяют получить оценки распределения случайных величин $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||Q)$ и $2m \cdot D_{\mathrm{KL}}(Q||\hat{P}_m)$ при больших m. Для решения задачи кластеризации (**??**) нам потребуется также исследовать поведение расстояния Кульбака-Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l)$ между гистограммами, построенными по выборкам X, X' различных длин m и l. Воспользовавшись неравенством треугольника

$$D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leqslant D_{\mathrm{KL}}(\hat{P}_m||Q) + D_{\mathrm{KL}}(Q||\hat{P}_l),$$

получим следствия из теоремы 5.1:

Следствие 1: $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m) \leqslant \chi^2_{2N}$ в пределе при $m \to \infty$.

Следствие 2: Пусть выборки X, X' растут таким образом, что $m/l \to \rho$, $0 < \rho < \infty$. Тогда

$$2\frac{ml}{m+l} \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leqslant \chi^2_{2N}$$

в пределе при $m, l \to \infty$.

Доказательство.

Действительно, при выполнении условия $m/l \to \rho$, $0 < \rho < \infty$, имеем

$$\frac{l}{m+l} \to \frac{1}{1+\rho}, \quad \frac{m}{m+l} \to \frac{\rho}{1+\rho}$$

и

$$2\frac{ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}_l) \leqslant \frac{l}{m+l} 2m D_{\mathrm{KL}}(\hat{P}_m || Q) + \frac{m}{m+l} 2l D_{\mathrm{KL}}(Q || \hat{P}_l) \to \chi^2_{2N}.$$

■

Обозначим величину $\frac{2ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m)$ через $\xi_{m,l}$. Следствие 2 дает верхнюю оценку поведения случайно величины $\xi_{m,l}$ при больших m и l, а именно: пусть $\eta \sim \chi^2_{2N}$, тогда при достаточно больших m и l выполнено для любого элементарного исхода w из вероятностного пространства Ω выполнено $\xi_{m,l}(w) < \eta(w)$. Следовательно, для любого $x \in \mathbb{R}$ верно

$$\mathrm{P}(\xi_{m,l} < x) \geqslant \mathrm{P}(\eta < x). \tag{5.5}$$

В следующем разделе покажем, как этот факт будет использоваться для проверки принадлежности временных рядов одному распределению.

6 Проверка принадлежности двух выборок к одному распределению

Для решения задачи агрегирования временных рядов $\mathbf{x}$ и $\mathbf{x}'$ необходимо уметь принимать решение о принадлежности временных рядов одному распределению. Опишем процедуру проверки гипотезы о принадлежности выборок X и X', составленных (**??**) из временных рядов $\mathbf{x}$ и $\mathbf{x}'$. Пусть нулевая гипотеза H_0 состоит в принадлежности выборок X и X' к одному распределению:

$$H_0 : P(x) = P'(x).$$

Сформулируем критерий проверки гипотезы H_0 при альтернативе $H_1 : P(x) \neq P'(x)$. Для этого определим критическую область $U(\alpha)$ для статистики $t_{m,l}$ с уровнем значимости α:

$$U(\alpha) = \{t : \bar{t}_{1-\alpha} > t \text{ или } t > \bar{t}_\alpha\},$$

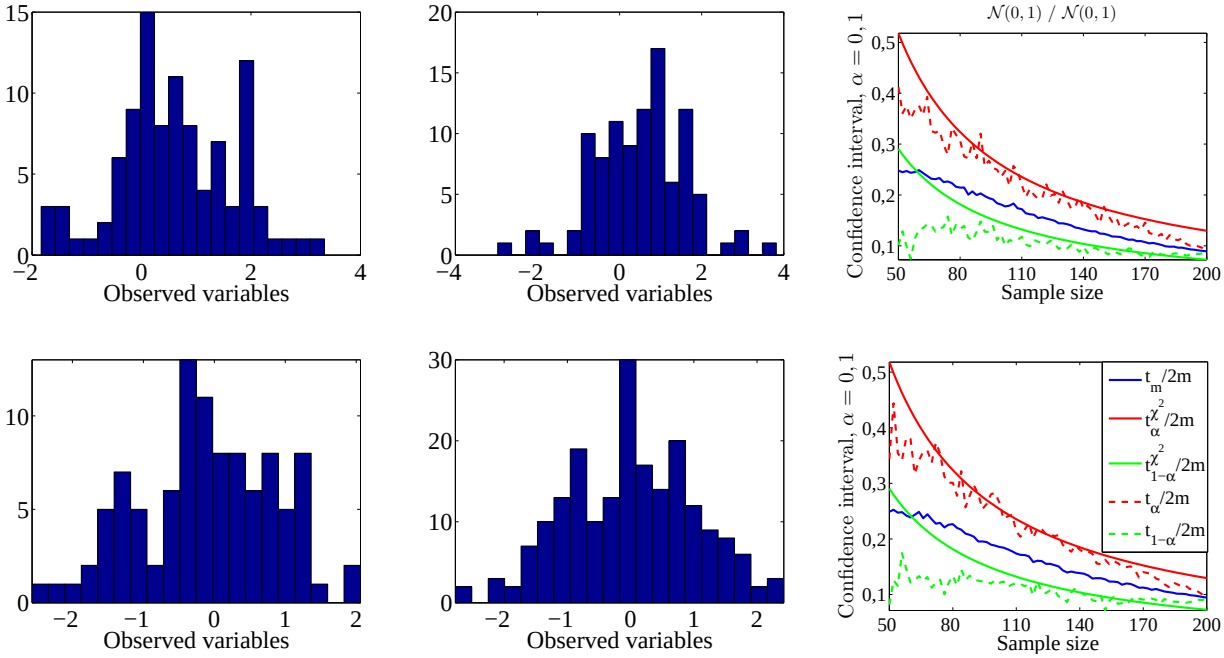


Рис. 2: Гистограммы, построенные по двум выборкам из нормального распределения и зависимость статистики t_m от объема выборки. Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0.1$. На рисунках снизу выборки были зашумлены.

где критическое значение $\bar{t}_\alpha$ определяется соотношением

$$\mathrm{P}(t > \bar{t}_\alpha | H_0) = \alpha. \tag{6.1}$$

Так как предельное распределение величины $\xi_{m,l}$ неизвестно, будем использовать критическую область, задаваемую распределением χ^2_{2N}. Будем говорить, что данные отвергают гипотезу H_0 в случае, если статистика $t_{m,l}$ принадлежит критической области U^{χ^2}

$$U^{\chi^2}(\alpha) = \{t : \bar{t}^{\chi^2}_{1-\alpha} > t \text{ или } t > \bar{t}^{\chi^2}_{\alpha}\}, \tag{6.2}$$

где $\bar{t}^{\chi^2}$ — критическое значение величины χ^2_{2N}:

$$\mathrm{P}(t > \bar{t}^{\chi^2}_{\alpha} | t \sim \chi^2_{2N}) = \alpha.$$

Из неравенства (5.5) и определения (6.1) критических значений следует, что критические области U и U^{χ^2} несравнимы, то есть

$$\bar{t}_{1-\alpha} < \bar{t}^{\chi^2}_{1-\alpha}, \quad \bar{t}_\alpha < \bar{t}^{\chi^2}_{\alpha}.$$

Это означает, что возможны следующие ситуации:

1. Случай $\bar{t}^{\chi^2}_{1-\alpha} < t_{m,l} < \bar{t}_{\alpha}$, когда статитика $t_{m,l}$ одновременно принадлежит истиной, но неизвестной критической области U, и вычислимой критической области U^{χ^2}.

2. Случай $\bar{t}_{1-\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{1-\alpha}$, когда статистика $t_{m,l}$ принадлежит истиной, но неизвестной критической области U, и не принадлежит U^{χ^2}. Так как $t_{m,l} \in U$, то с высокой вероятностью гипотеза H_0 неверна, и мы рискуем принять неверное решение об истинности гипотезы H_0. То есть зазор между $\bar{t}_{1-\alpha}$ и $\bar{t}^{\chi^2}_{1-\alpha}$ повышает вероятность ошибки второго рода.

3. Случай $\bar{t}_{\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{\alpha}$, когда статистика $t_{m,l}$ попадает в U^{χ^2}, хотя на самом деле $t_{m,l}$ не принадлежит U. В этом случае велика вероятность, что H_0 верна, но решение будет принято в пользу H_1. Таким образом, зазор между $\bar{t}_{\alpha}$ и $\bar{t}^{\chi^2}_{\alpha}$ повышает вероятность ошибки первого рода.

Второй случай разрешается следующим образом: использование симметризованного расстояния позволяет перейти от двусторонних критериев U и U^{χ^2} вида (6.2) к односторонним критериям

$$U_1(\alpha) = \{t : t > \bar{t}_{\alpha}\}, \quad U_1^{\chi^2}(\alpha) = \{t : t > \bar{t}^{\chi^2}_{\alpha}\}.$$

В этом случае $U_1^{\chi^2} \subseteq U_1$ и справедливо следствие $t_{m,l} \in U_1^{\chi^2} \Rightarrow t_{m,l} \in U_1$. Кроме того, далее будет показано (теорема 6), что при увеличении объема выборки m вероятность отклонить гипотезу H_0 с помощью критерия (6.2) в случае, если гипотеза H_0 неверна, стремится к единице. Влияние третьего случая на возможность применения критерия (6.2) для принятия нулевой гипотезы исследуется экспериментально. Эксперименты, приведенные ниже и в разделе «Вычислительный эксперимент» показывают, что при истинности нулевой гипотезы области U и U^{χ^2} достаточно близки для принятия верного решения.

Пример применения критерия (6.2) **при истинности** H_0**.** На рисунке 5 изображены гистограммы из для двух выборок из

- стандартного нормального распределения 5 (сверху)

- и стандартного нормального распределения с шумом $\varepsilon \sim 0.1 \cdot R[0,1]$ 5 (снизу),

а также зависимость расстояния Кульбака-Лейблера между гистограммами между выборками одинакового объема от объема выборки m и область допустимых значений с точки зрения критерия (6.2). Здесь вместо критических значений $\bar{t}^{\chi^2}_{1-\alpha}$, $\bar{t}^{\chi^2}_{\alpha}$ и $t_m = 2m \cdot D_{\mathrm{KL}}$ отложены величины $\bar{t}^{\chi^2}_{1-\alpha}/2m$, $\bar{t}^{\chi^2}_{\alpha}/2m$ и $t_m/2m$, чтобы продемонстрировать масштаб расстония Кульбака-Лейблера и наличие сходимости. Рисунки показывают, что в данном случае использование распределения χ^2_{2N} в качестве оценки предельного распределения статистики t_m позволяет принять верное решение о принадлежности рядов к одному распределению. Пунктирная линия показывает границу области, в которую вошли $1 - \alpha = 90\%$ выборки и задает оценку критической области для $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$. Более подробно результаты описаны в разделе «Вычислительный эксперимент».

Покажем теперь, что критерий (6.2) так же можно использовать для отвержения гипотезы H_0.

Теорема 6.1. *Критерий* (6.2) *состоятелен:*

$$\lim_{m\to\infty} \mathrm{P}(t_m \in U|H_1) = 1,$$

то есть вероятность отвергнуть гипотезу H_0, если распределения временных рядов X и X' не совпадают, с увеличением выборки стремится к единице.

Доказательство.

Пусть функции распределения P и P' временных рядов не совпадают. Тогда найдется $x^* \in \mathbb{R}$, в котором значения этих функций различны $P(x^*) \neq P'(x^*)$. Тогда найдется такой способ разбиения пространства $\mathbb{R}$, что для некоторого i вероятность попадания в i-тый промежуток не одинакова для рассматриваемых случайнах величин:

$$P(a_i < x \leqslant a_{i+1}) = p_i \neq p'_i = P'(a_i < x \leqslant a_{i+1}).$$

Пусть $p_i > p'_i$. Согласно (5.4), при больших m с вероятностью $\mathrm{P} > (2\Phi(C_1) - 1)(2\Phi(C_2) - 1)$ выполнено

$$|n_i - mp_i| < C_1\sqrt{m}, \quad |n'_i - mp'_i| < C_2\sqrt{m}.$$

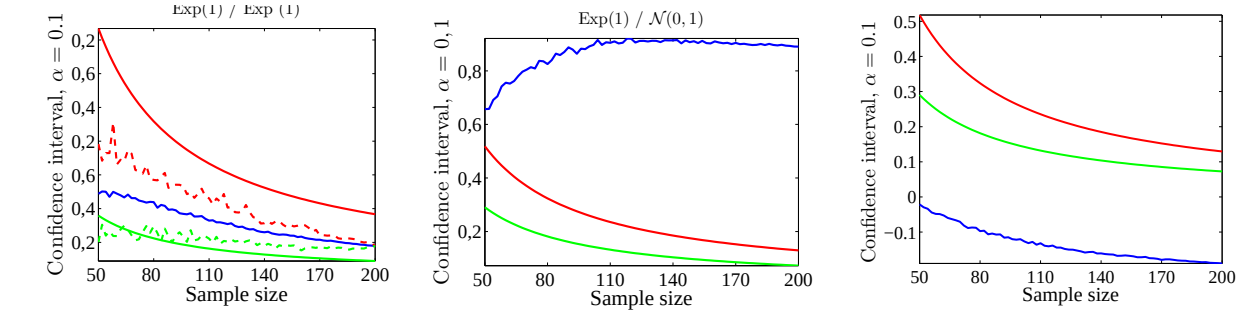
Для любого $\varepsilon > 0$ найдется константа C_ε : $\mathrm{P} > (2\Phi(C_\varepsilon) - 1)^2 > 1 - \varepsilon$. Выберем $C_1 = C_2 = C_\varepsilon$. Тогда $(n_i - n'_i) > m(p_i - p'_i) + O(\sqrt{m})$ и

$$\frac{(n_i - n'_i)^2}{n_i} > m\frac{(p_i - p'_i)^2}{p_i} + O(\sqrt{m}) > Cm. \tag{6.3}$$

Следовательно, для любого $\alpha \in (0,1)$ при достаточно больших m

$$t_m = 2mD_{\mathrm{KL}}(\hat{P}_m^1 || \hat{P}_m^2) \sim \sum_{i=1} \frac{(n_i - n_i')^2}{n_i} > Cm > \bar{t}_\alpha$$

с вероятностью $\mathrm{P} > 1 - \epsilon$, то есть вероятность $\mathrm{P}(t_m > \bar{t}_\alpha) \to 1$ при $m \to \infty$. ■



(a) $Exp(1)||Exp(1)$ (b) $Exp(1)||\mathcal{N}(0,1)$ (c) $Exp(1)||R0,1$

(d) $\mathcal{N}(0,1)||Exp(1)$ (e) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (f) $\mathcal{N}(0,1)||R0,1$

(g) $R0,1||Exp(1)$ (h) $R0,1||\mathcal{N}(0,1)$ (i) $R0,1||R0,1$

Рис. 3: Зависимость статистики t_m от объема выборки для различных пар распределений. Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0.1$.

7 Вычислительный эксперимент

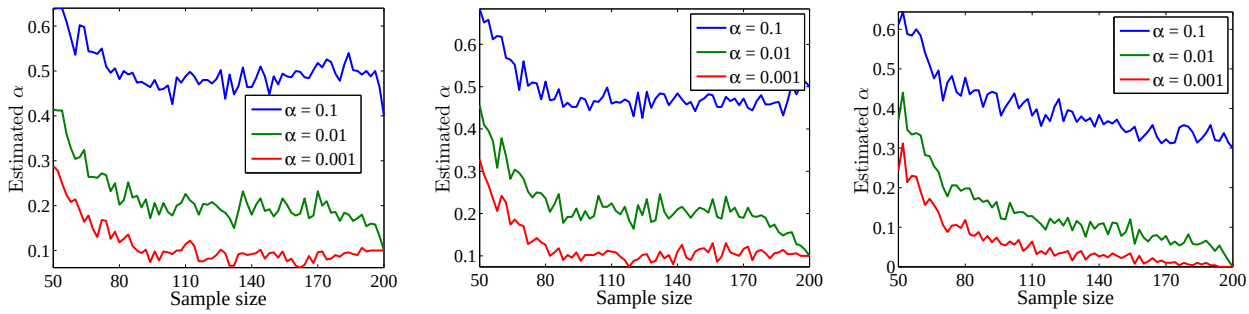
Работа критерия была рассмотрена на различных парах распределений. Для выбранной пары распределений повторяется следующая процедура:

1) генерируются выборки X и X' одинакового объема m;

2) по выборкам строятся гистограммы $\hat{P}_m$ и $\hat{P'}_m$ с фиксированным числом разбиений $N = 20$ и вычисляется расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P'}_m)$;

3) расстояния усредняются по 1000 генерациям выборок

4) объем m выборки увеличивается.



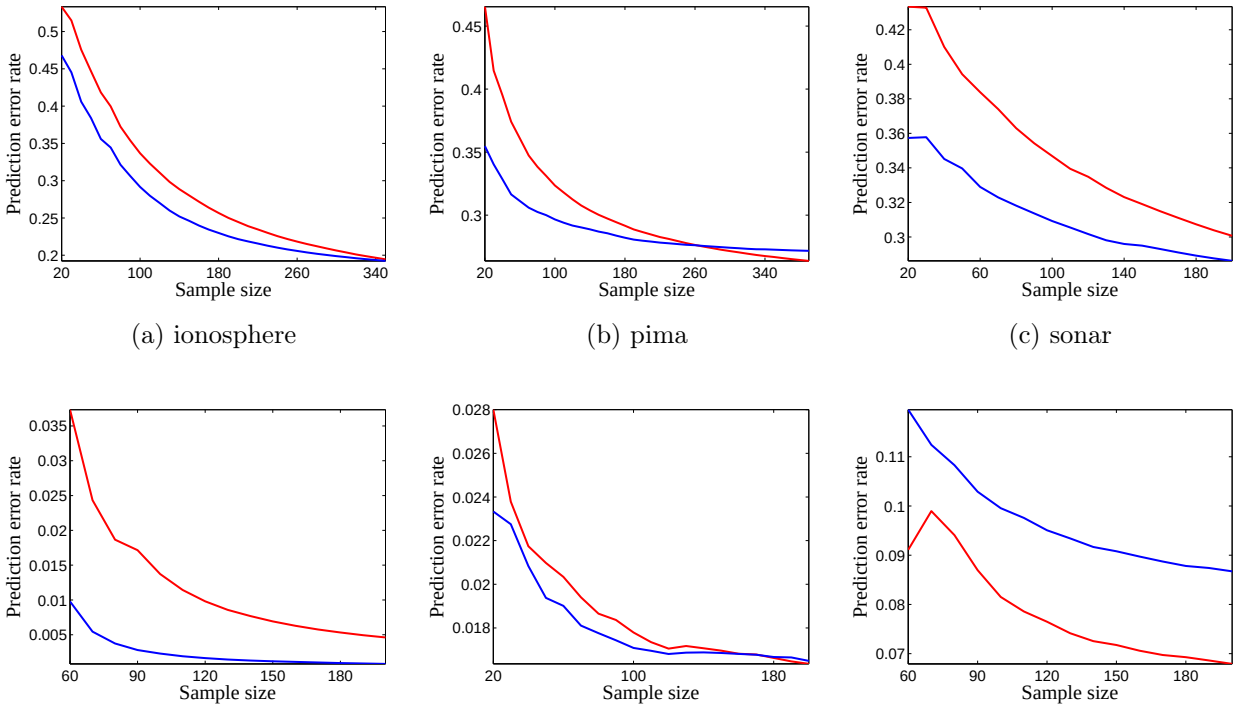
(a) $Exp(1)||Exp(1)$ (b) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (c) $R(0,1)||R(0,1)$

Рис. 4: Зависимость фактического уровня значимости от объема выборки различных уровнях значимости критерия χ^2_{2N}.

На графике 6 отложены расстояния $D_{\mathrm{KL}}(\hat{P}_m|\hat{P'}_m)$ в зависимости от объема выборки и критические значения $\bar{t}^{\chi^2}/2m$ при заданном уровне значимости α. Заметим, что в случае различных распределений расстояние $D_{\mathrm{KL}}(\hat{P}_m|\hat{P'}_m)$ быстро попадает в критическую область, и характер его зависимости от m согласуется с оценкой (6.3). Отрицательные значения, нехарактерные для расстояния Кульбака–Лейблера, возникают при численном приближении интеграла (5.1), когда распределение в знаменателе под знаком логарифма иеет большую область определения. Именно из-за отрицательных значений был использован двусторонний критерий. Для симметризованного расстояния Кульбака–Лейблера используется односторонний критерий. В случае с выборками из одного распределения, на графике также отложены пунктиром оценки критических значений $\bar{t}$, полученные экспериментально. Видно, что,

несмотря на то критические области U^{χ^2} и U не совпадают, при истинности гипотезы H_0 зависимость не попадает ни в U^{χ^2}, ни в U.

Оценка фактического значения α. Чтобы оценить реальный уровень значимости решения о принятии или отвержении гипотезы H_0, подсчитаем среднюю долю объектов выборки, попавших в U^{χ^2} при заданном α. Результаты отражены на рисунке 4. Из рисунков следует, что для достижения уровня значимости $\alpha = 0.1$ нужно использовать в качестве оценки U критическую область U^{χ^2} с уровнем значимости 0.001.



(a) ionosphere (b) pima (c) sonar

(d) $\sigma = 0.1$ (e) $\sigma = 0.2$ (f) $\sigma = 0.3$

Рис. 5: Красный — логистическая регрессия, синий — наивный Байес; синтетические данные

7.1 Сравнение порождающего и разделяющего подходов

Сравнение подходов проводится в случае логистической регрессии (разделяющий подход) и наивного байесовского классификатора (порождающий подход). В качестве приближения вероятности ошибки (3.1) рассмотрим частоту ошибки каждого

классификатора a на выборке D:

$$\hat{\varepsilon}_m(a) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) \neq y_i].$$

На рисунке 7 (сверху) изображены зависимости ошибки $\hat{\varepsilon}_m(a)$ от объема выборки. Для каждого объема m выборки параметры настраивались по 0.25% выборки, а значение ошибки вычислялось по оставшимся элементам выборки. Разбиение проводилось 100 раз для каждого m, после чего значение $\hat{\varepsilon}_m(a)$ усреднялось. Красным цветом изображена зависимость для логисической регрессии, синим — для навного байесовского классификатора.

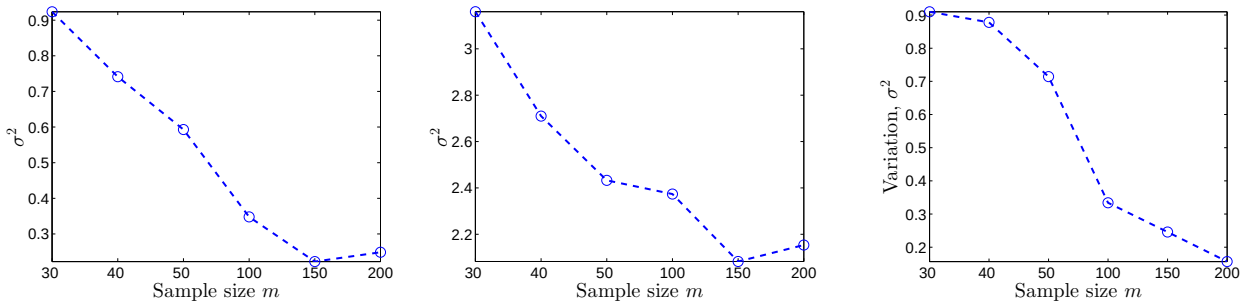
Тот же эксперимент был проведен на синтетических данных, сгенерированных в соответствии с предположением (2.5). Каждый класс описывается только одним признаком:

$$x_i \sim \mathcal{N}(\mu_1, \sigma^2), \quad \text{если } y_i = 1,$$

$$x_i \sim \mathcal{N}(\mu_0, \sigma^2), \quad \text{если } y_i = 0,$$

где $\mu_1 - \mu_0 = 1$, а σ^2 различно в разных экспериментах. На рисунке 7 изображены зависимости $\hat{\varepsilon}_m(a)$ от объема выборки для значений $\sigma^2 = 0.1,\ 0.2,\ 0.3$. В первом случае, когда дисперсия признаков очень мала, наивный байесовкий классификатор имеет лучшую обобщающую способность независимо от объема выборки. При $\sigma^2 = 0.3$ логистическая регрессия дает меньшую ошибку прогнозирования при всех исследуемых значениях объема выборки m. При $\sigma^2 = 0.2$, логистическая регрессия дает меньшую ошибку прогнозирования лишь начиная с значения $m \approx 170$. Таким образом, решение о выборе классификатора a_{D} или a_{G} зависит не только от объема выборки, но и от параметров распределения данных.

Зависимость средней ошибки $\hat{\varepsilon}$ от параметров распределения $\mu_1,\ \mu_0, \sigma^2$ и объема выборки m продемонстрируем экспериментом на синтетических данных, сгенерированных в соответствии с принятыми предположениями (2.5) о распределении данных. В данном эксперименте фиксировались математические ожидания μ_1 и μ_0, параметр σ^2 варьировался. Параметры настраивались по выборкам размеров $m = 50,\ 100,\ 150$ при увеличении параметра σ^2. Зависимость ошибки классификации $\hat{\varepsilon}$ для каждого из классификаторов a_{D} и a_{G} от параметра σ^2 изображена на рисунке С увеличением σ^2 ошибки a_{D} и a_{G} ожидаемо возрастают и во всех случаях наблюдается пересечение: при малых значениях σ^2 имеем $\hat{\varepsilon}(a_{\text{D}}) > \hat{\varepsilon}(a_{\text{G}})$, а начиная с некоторого момента



(a) $P = 0.3$ (b) $P = 0.6$ (c) $P = 0.7$

Рис. 6: Зависимость объема выборки, при котором происходит пересечение кривых $\varepsilon_{\text{D}}(m)$ и $\varepsilon_{\text{G}}(m)$ от дисперсии σ^2 для синтетических данных с фиксированными матожиданиями $\mu_1 = 1$, $\mu_0 = 0$.

$\hat{\varepsilon}(a_{\text{D}}) < \hat{\varepsilon}(a_{\text{G}})$. Причем значение σ^2 в точке пересечения тем меньше, чем больше объем выборки m.

Для различных значений априорной вероятности P класса “1” была подсчитана зависимость объема выборки m при котором происходит пересечение кривых $a_{\text{D}}(m)$ и $a_{\text{G}}(m)$ от дисперсии σ^2 распределения (2.5). Результаты изображены на рисунке 7.1. Попадание пары $(m(\sigma^2), \sigma^2)$ на график для некоторого значения P означает, что для выбранных P и σ^2 выполняется

$$\varepsilon_m(a_{\text{G}}) < \varepsilon_m(a_{\text{D}}) \quad \text{при } m < m(\sigma^2),$$

$$\varepsilon_m(a_{\text{G}}) = \varepsilon_m(a_{\text{D}}) \quad \text{при } m = m(\sigma^2),$$

$$\varepsilon_m(a_{\text{G}}) > \varepsilon_m(a_{\text{D}}) \quad \text{при } m > m(\sigma^2).$$

Заметим, что чем больше величина $|P - 1/2|$, характеризующая ассиметричность выборки, тем более плотными должны быть классы, чтобы порождающий подход давал лучший результат, чем разделяющий.

7.2 Определение необходимого объема выборки при порождающем и разделяющем подходах

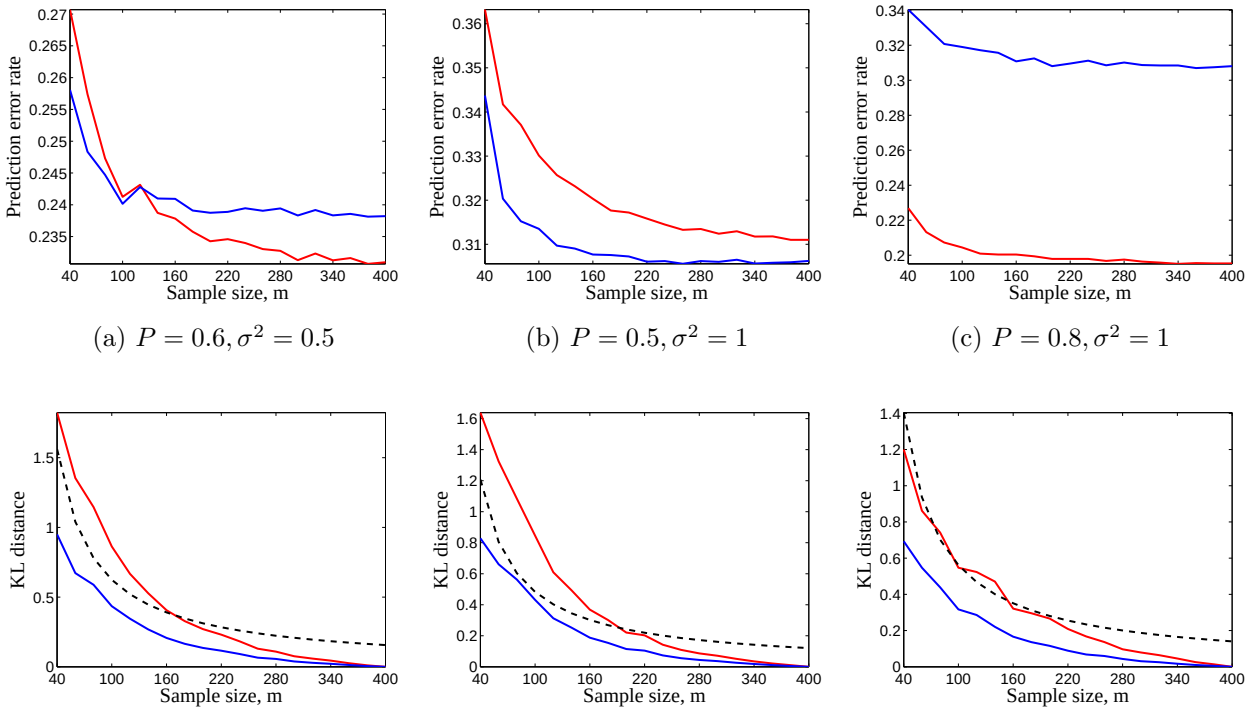
Предполагается, что эффективное применение классификатора возможно только в том случае, если параметры настраивались на выборке объема $m \geqslant m^*$, где m^* — необходимый объем выборки, определяемый в разделе 4 «Оценка объема выборки с

использованием расстояния Кульбака-Лейблера». Так как оцениваемый объем выборки зависит от принятой модели распределения данных, принятое предположение приводит к задаче оценке значений m^*_{D} и m^*_{G} необходимого объема выборки при разделяющем и порождающем подходе. Выпишем расстояние Кульбака-Лейблера между гистограммами $\hat{p}$ выборок X и X' объема m в каждом случае:

$$\text{KL}_{\text{D}} = D(\hat{p}(y|X')||\hat{p}(y|X)) \sim \sum_{j=0,1}\sum_{i=1}^{N} \frac{(n_{ij}/n_j - n'_{ij}/n'_j)^2}{2n_{ij}/n_j}, \quad n_j = \sum_{i=1}^{m} n_{ij},$$

$$\text{KL}_{\text{G}} = D(\hat{p}(y, X'_m)||\hat{p}(y, X)) \sim \sum_{j=0,1}\sum_{i=1}^{N} \frac{(n_{ij} - n'_{ij})^2}{2mn_{ij}}.$$

На рисунке 7 изображена зависимость расстояний Кульбака-Лейблера KL_{G} и KL_{D} между выборками объема m от величины m при порождающем и разделяющем подходах. Видно, что расстояния сходятся к одной величине, причем KL_{G} сходится быстрее. Зависимости были построены на синтетических данных, сгенерированных в соответствии с (2.5) с параметрами $\mu_1 = 1$, $\mu_0 = 0$, $\sigma^2 = 1$. На рисунках сверху изоб-



(a) $P = 0.6, \sigma^2 = 0.5$ (b) $P = 0.5, \sigma^2 = 1$ (c) $P = 0.8, \sigma^2 = 1$

Рис. 7: Зависимость расстояния Кульбака-Лейблера от объема выборки при порождающем и разделяющем подходах для различных случаев взаимного расположения кривых ε_{D} и ε_{G}.

ражены зависимости ошибки классификации при порождающем и разделяющем подходах от объема выборки для различных параметров распределения данных. Рисунки демонстрируют различное относительное положение кривых ε_{D} и ε_{G}. На рисунках снизу красным и синим отложены зависимости расстояния Кульбака-Лейблера между гистограммами, построенными на подвыборках, от их объема m. Пунктиром обозначена граница критической области. Для каждого из подходов достаточный объем m^*_{D} и m^*_{G} соответствует пересечению кривых ε_{D} и ε_{G} границей критической области. В соответствии с рисунками, предлагается отдавать предпочтение порождающему классификатору, если объем выборки $m < m^*_{\text{D}}$ и разделяющему, если $m \geqslant m^*_{\text{D}}$. Ситуация $m < m^*_{\text{G}}$ в вычислительных экспериментах реализована не была. Это может объясняться тем, что подсчитать расстояние Кульбака-Лейблера при малых значениях объема выборки становится невозможным: слишком много эмпирических вероятностей $\hat{p}_{ij}$ обнуляется. Для устранения этого эффекта планируется исследование более устойчивых модификаций расстояния Кульбака-Лейблера [23, 24].

8 Заключение

В работе рассмотрена задача определения объема выборки при решении задач классификации и прогнозирования. Предложен метод, основанный на определении устойчивости эмпирических распределений, оцениваемых на подвыборках заданного объема. Такой подход, в отличие от классических статистических методов позволяет позволяет учесть постановку задачи классификации. Для сравнения распределений используется расстояние Кульбака-Лейблера. Получены предельные оценки сверху распределения расстояния Кульбака-Лейблера между гистограммами из одного распределения, предложен критерий решения задачи двух выборок на основе расстояния Кульбака-Лейблера. Продемонстрирована возможность использования предложенного метода оценки объема выборки для выбора между порождающим и разделяющим подходами к классификации.

Список литературы

[1] McCullagh P., Nelder J. A. *Generalized Linear Models.* 2nd edition. Chapman and Hall, London. 1989.

[2] Hosmer D., Lemeshow S. *Applied Logistic Regression*, 2nd Edition, John Wiley and Sons, 2000.

[3] Bishop, C. M., *A New Framework for Machine Learning.* J.M. Zurada et al. (Eds.): WCCI 2008 Plenary/Invited Lectures, LNCS 5050, pp. 1-24, 2008.

[4] Efron, B., *The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis.* Journal of the American Statistical Association, 70(352), 892-898, 1975.

[5] Ng, A. Y. and Jordan, M. I., *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes.* In Advances in Neural Information Processing Systems. Volume 14, p. 841-848. 2002.

[6] P. Liang and M. I. Jordan, *An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators.* In Proceedings of the 25th International Conference on Machine Learning (ICML), 2008.

[7] J.-H Xue and D.M. Titterington, *Comment on "discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes".* Neural Processing Letters, 28(3), 169-187, 2008.

[8] Bouchard, G. and Triggs, B., *The tradeoff between generative and discriminative classifiers.* In J. Antoch, editor, Proc. of COMPSTAT'04, 16th Symposium of IASC, volume 16. Physica-Verlag, 2004.

[9] Raina, R.; Shen, Y.; Ng, A. Y. and McCallum, A., *Classication with hybrid generative/discriminative models.* Advances in Neural Information Processing Systems, volume 16. Cambridge, MA: The MIT Press, 545-552. 2003.

[10] Bishop, C. M., *Pattern Recognition and Machine Learning.* Berlin: Springer-Verlag. 2006.

[11] Bishop, C. M. and Lasserre, J., *Generative or Discriminative? getting the best of both worlds.* In Bayesian Statistics 8, Bernardo, J. M. et al. (Eds), Oxford University Press. 3-23, 2007.

[12] Azuaje F., Devaux Y., Wagner D. *Computational biology for cardiovascular biomarker discovery.* // Brief Bioinform. 2009. V. 10, № 4. P. 367–377.

[13] Медведникова М. М. 2014. Согласование агрегированных непараметрических прогнозов временных рядов. Машинное обучение и анализ данных. Т.1 № 8. (принято в печать)

[14] S. Kullback. 1959. Information Theory and Statistics. New York: Wiley.

[15] H. Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. Ann. Math. Statist. 4(23):493-655

[16] A. N. Kolmogorov. 1965. On the approximation of distributions of sums of independent summands by infinitely divisible distributions. Contributions to statistics. 158-174.

[17] S. M. Ali, S. D. Silvey. 1966. A general class of coefficients of divergence of a distribution from another. Journal of Royal Statistical Soiety. Series B (Methodoligical). 1(28):131-142.

[18] I. Csiszar and P. Shields. 2004. Information theory and statistics: A tutorial. Foundations and Trend in Communications and Information Theory, 4:417–528.

[19] A. L. Gibbs, F. E. Su. 2002. On Choosing and bounding probability metrics. International Statistical Review. 3(70):419–435.

[20] C. Mallows. 1972. A note on asymptotic joint normality. Annals of Mathematical Statistics, 42(2):508–515.

[21] A. Irpino, R. Verde, and Y. Lechevallier. 2006. Dynamic clustering of histograms using Wasserstein metric. COMPSTAT. 869-876.

[22] А. П. Мотренко. 2014. Статистический тест для проверки гипотезы о принадлежности двух выборок одному распределению на основе расстония Кульбака-Лейблера.
http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group874/Motrenko2014KL/code/KLtest.

[23] A. Finkler. Goodness of fit statistics for sparse contingency tables. arXiv:1006.2620 [math.ST]. 2010.

[24] U. Keich, N. Nagarajan. A Fast and Numerically Robust Method for Exact Multinomial Goodness-of-Fit Test. Journal of Computational and Graphical Statistics. 15(4): 779–802.