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»

Шишковец Светлана Сергеевна

Аддитивная регуляризация наивного линейного байесовского классификатора

03.03.01 — Прикладные математика и физика

Бакалаврская диссертация

Научный руководитель:
ст.н.с ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва, 2016 г.

Содержание

Аннотация

Рассматривается задача построения наивного байесовского классификатора(NB) и его аддитивной регуляризации. Основная цель регуляризации — ослабить ограничение независимости признаков, сохранив при этом высокую скорость обучения NB. Идея исследования опирается на теорему о линейности NB классификатора в случае принадлежности параметров к экспоненциальному семейств плотностей. Так же рассматривается несколько вариантов регуляризаций для введения отбора признаков и расширения задачи классификации на многоклассовый случай.

Ключевые слова: *машинное обучение, линейный классификатор, наивный байесовский классификатор, регуляризация, отбор признаков.*

1 Введение

Наивный байесовский классификатор (NB) является одним из простейших методов машинного обучения. В его основе лежит предположение о статистической независимости признаков, что ограничивает его применимость. Тем не менее, NB неплохо зарекомендовал себя во многих прикладных областях, в частности, в задачах медицинской диагностики, биоинформатики, классификации текстов. Для обучения (оптимизации параметров) модели достаточно вычислить средние значения каждого признака в каждом классе. Таким образом, вычислительная сложность алгоритма оптимизации параметров NB линейна по объёму обучающей выборки и по числу признаков. Простые эвристики отбора признаков позволяют улучшать качество классификации, не увеличивая сложность алгоритма обучения. Всё это делает NB перспективным методом для анализа больших данных, при условии расширения модели и преодоления ограничения независимости признаков.

В данной работе предлагается обобщение NB по двум направлениям.

Во-первых, вводится предположение, что признаки имеют вероятностные распределения из экспоненциального семейства. Это позволяет рассматривать задачи с разнотипными признаками, гарантирует линейность классификатора и возможность записи решения в аналитическом виде.

Во-вторых, для получения смещённых оценок максимального правдоподобия вводится линейная комбинация регуляризаторов, что смягчает ограничение независимости признаков. При этом вычислительная сложность метода обучения остаётся линейной по объёму выборки и числу признаков. Для оценивания параметров модели по-прежнему достаточно вычисления лишь средних значений каждого признака в каждом классе, что допускает возможность онлайновой обработки данных.

Эксперименты на задачах классификации символьных последовательностей подтверждают эффективность метода.

2 Байесовский классификатор

Рассмотрим задачу обучения классификации в вероятностной постановке. Имеется множество *объектов* $\mathbb{X}$ и конечное множество *классов* $\mathbb{Y}$. Их декартово произведение $\mathbb{X} \times \mathbb{Y}$, т. е. множество пар «объект, ответ», является вероятностным пространством с неизвестной плотностью $p(x, y)$. *Классификатором* называется отображение вида $a\colon \mathbb{X} \to \mathbb{Y}$. Согласно байесовской теории классификации [1], минимальной ошибкой классификации обладает *оптимальный байесовский классификатор*, который относит объект x к тому классу y, для которого *апостериорная вероятность* $P(y|x)$ максимальна:

$$a(x) = \arg\max_{y\in\mathbb{Y}} P(y|x) = \arg\max_{y\in\mathbb{Y}} P(y)p(x|y),$$

где $P(y)$ — *априорная вероятность* класса y, $p(x|y)$ — *условная плотность распределения* класса y. В задачах обучения классификации плотность $p(x, y)$ не известна; известна лишь конечная обучающая выборка наблюдений $(x_i, y_i) \in \mathbb{X} \times \mathbb{Y},\ i = 1, \ldots, \ell$.

Предполагается, что эти наблюдения получены в результате случайных и независимых испытаний из распределения с плотностью $p(x, y)$. Кроме того, предполагается, что условные плотности распределения каждого класса y описываются параметрической моделью $p(x|y) = p(x|\theta_y)$ с параметром θ_y.

В этих предположениях справедливы несмещённые частотные оценки априорных вероятностей классов

$$P(y) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[y_i = y\big],$$

а для несмещённого оценивания параметров условных плотностей применяется принцип максимума правдоподобия:

$$\ln\prod_{i=1}^{\ell} p(x_i, y_i) = \sum_{i=1}^{\ell}\ln p(x_i|y_i)P(y_i) \ \to\ \max_{\{\theta_y\}}.$$

Таким образом, задача обучения классификатора a по обучающей выборке распадается на $|\mathbb{Y}|$ независимых подзадач оптимизациии:

$$\sum_{y\in\mathbb{Y}}\sum_{x_i\in X_y}\ln p(x_i|\theta_y) + \underbrace{|X_y|\ln P(y_i)}_{\text{const}} \ \to\ \max_{\{\theta_y\}},$$

где X_y — выборка объектов x_i класса y, а второе слагаемое не зависит от параметров модели θ_y и может быть отброшено.

В случае двух классов, $\mathbb{Y} = \{-1, +1\}$ оптимальный байесовский классификатор записывается эквивалентным образом в виде

$$a(x) = \operatorname{sign}\Big(P(+1)p(x|+1) - P(-1)p(x|-1)\Big) = \operatorname{sign}\Big(\ln \tfrac{p(x|+1)}{p(x|-1)} + \ln \tfrac{P(+1)}{P(-1)}\Big). \qquad (2.1)$$

3 Наивный байесовский классификатор

Пусть объекты описываются n-мерными векторами признаков, $x = (x^1, \ldots, x^n)$, где каждый признак имеет свою область допустимых значений, $x^j \in D_j$. Когда множества D_j различны, говорят, что признаки являются разнотипными. В прикладных задачах классификации и распознавания образов наиболее распространены случаи вещественных, целочисленных и бинарных признаков.

Сделаем сильное дополнительное предположение, что признаки являются статистически независимыми случайными величинами. Тогда n-мерная условная плотность представляется произведением одномерных условных плотностей значений признаков:

$$p(x|y) = p(x^1, \ldots, x^n|y) = p(x^1|y) \cdots p(x^n|y),$$

где каждая одномерная плотность значений признака j в классе y описывается параметрической моделью с параметром θ_y^j:

$$p(x^j|y) = p(x^j|\theta_y^j).$$

Благодаря предположению независимости, задача максимизации логарифма правдоподобия распадается на независимые подзадачи как по классам y, так и по признакам j:

$$\mathcal{L}(\Theta) = \sum_{j=1}^{n} \sum_{y \in \mathbb{Y}} \sum_{x_i \in X_y} \ln p(x_i^j|\theta_y^j) \;\to\; \max_{\Theta}. \qquad (3.1)$$

где $\Theta = \big(\theta_y^j \colon y \in \mathbb{Y}, j = 1, \ldots, n\big)$ — вектор параметров всех одномерных плотностей.

Предположение о независимости считается чрезмерно сильным. Статистические эксперименты на реальных задачах классификации, как правило, отвергают гипотезу независимости. Поэтому байесовский классификатор, основанный на данном предположении, называют *наивным*.

Тем не менее, *наивный байесовский классификатор* (Naïve Bayes, NB) иногда используется на практике, и в некоторых задачах показывает неплохие результаты, причём даже в тех случаях, когда признаки заведомо не являются независимыми.

Этот парадокс заслуживает внимания хотя бы потому, что NB вычислительно очень эффективен. Сведение задачи оценивания n-мерной плотности распределения по конечной выборке к n задачам оценивания одномерных плотностей приводит к тому, что время обучения классификатора составляет $O(n\ell)$, то есть является линейным по объёму выборки и числу признаков.

4 Аналитический обзор

Попытки улучшить качество классификации NB неоднократно упоминались в литературе. Можно выделить три направления исследований в данной области:

1. Применение стандартных методов отбора признаков к NB
2. Компенсация статистической зависимости с помощью введения поправок в виде дополнительных функций в постановку задачи NB
3. Ослабление гипотезы о независимости с помощью оценки плотностей совместных распределений признаков

Одним из стандартных методов отбора является фильтрация признаков, при которой признаки сначала ранжируются по некоторому критерию, а затем отбираются первые k штук. В качестве критерия может быть статистика χ^2(чем выше значение, тем вероятнее, что признак и целевая переменная зависимы), коэффициент корреляции между признаком и целевой переменной, или их взаимная информация(сколько информации привносит признак для того, чтобы правильно определить класс):

$$I(i) = \int_{x_i} \int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy$$

В работе [2] к NB применялся модифицированный метод фильтрации, в котором признаки взвешивались пропорционально мере Кульбака–Лейблера, при этом авторы обнаружили прирост качества классификации.

В [3] рассматривался NB с пуассоновским распределением. Чтобы отобрать информативные признаки и повысить AUC, для каждого признака j в каждом классе из множества $\{+1, -1\}$ оценивался параметр распределения λ, а в качестве критерия ранжирования рассчитывался логарифм их отношения:

$$R(i) = \log \frac{\lambda_j^{+1}}{\lambda_j^{-1}}$$

В [4] в качестве ранжирующего критерия использовалась мера симметричной неопределенности:

$$SU = 2\frac{E(X) + E(Y) - E(X, Y)}{E(X) + E(Y)},$$

где величины $E(X), E(Y), E(X, Y)$ выражаются следующим образом:

$$E(X) = -\sum_{i=1}^{l} Pr(x_i) \log_2 Pr(x_i)$$

$$E(Y) = -\sum_{j=1}^{J} Pr(y_j) \log_2 Pr(y_j)$$

$$E(X,Y) = -\sum_{i=1}^{l} \sum_{j=1}^{J} Pr(x_i, y_j) \log_2 Pr(x_i, y_j)$$

В работе [5] для NB протестированно 6 ранжирующих методов:

- Information Gain (IG)
- Gain Ratio (GR)
- Symmetrical Uncertainty(SU)
- Relief-F (RF)
- One-R (OR)
- Chi-Squared (CS)

Из 11 наборов данных из репозитория UCI только метод One-R показал прирост качества классификации на всех наборах. Остальные методы были эффективными на 5-6 наборах, тогда как на оставшейся части данных качество классификации осталось на прежнем уровне или ухудшилось. Это говорит о том, что ранжирование признаков не является универсальным методом в контексте NB для произвольной структуры в данных.

Методы фильтрации просты в вычислительном плане, поскольку требуют расчета n критериев. К недостаткам стоит отнести ненулевую вероятность выбора избыточного количества признаков в ходе фильтрации, поскольку известно, что NB чувствителен к корреллирующим и нерелевантным признакам (однако, в некоторых ситуациях это может стать достоинством, например, см. [6])

Другой группой стандартных методов являются обертки(wrappers) - из множества признаков отбирается то подмножество, на котором алгоритм классификации показывает наилучшее качество. Сам алгоритм при этом рассматривается как черный ящик. В простейших случаях, когда число признаков невелико, можно использовать перебор или метод ветвей и границ, что требует экспоненциальное по числу

признаков время. С повышением размерности задачи может применяться жадный поиск с квадратичным по числу признаков временем работы. Обертки показывают лучшее качество отбора признаков, чем фильтры, но при этом требуют больше вычислительных ресурсов. В работе [7] в качестве классификатора использовался NB, а для отбора признаков алгоритм Sequential Forward Selection(SFS) и его улучшенная версия Incremental Wrapper Subset Selection (IWSS). Показанно, что SFS отбирает меньшее количество признаков, чем алгоритм FCBF из группы методов-фильтров, но при этом они соизмеримы по качеству классификации. Алгоритм IWSS отбирает меньше признаков, чем FCBF и качество классификации при этом выше.

Последняя группа методов из класса стандартных - встроенные методы. Основная их особенность состоит в том, что отбор признаков является частью процесса обучения, т.е. ищется такое подмножество переменных, которое приводит к наибольшей обобщающей способности алгоритма(минимизирует риск). Каждое подмножество признаков может быть представлено в виде характеристического вектора $\sigma \in \{0,1\}^n$. Если координата этого вектора равна единице, то соответствующий признак присутствует в подмножестве. Задача состоит в том, чтобы найти такой вектор $\sigma^* \in \{0,1\}^n$ и вектор параметров $\alpha^* \in \Lambda$, что для параметрического семейства функций $f : \Lambda \times \mathbb{R}^n \to \mathbb{R}$ минимизируется риск [8]:

$$R(\alpha, \sigma) = \int L[f(\alpha, \sigma \circ x), y] dP(x, y)$$

Свойством отбора признаков обладают такие методы как деревья решений, SVM с l_1-регуляризатором, а также линейные методы, которые можно рассматривать как результат минимизации выражения:

$$\min_{w,b} \frac{1}{m} \sum_{k=1}^{m} L(w \cdot x_k + b, y_k) + C\Omega(w),$$

где $L(f(x_k), y_k)$ является функцией потерь, $\Omega(w) : \mathbb{R}^n \to \mathbb{R}_+$ - регуляризатор. Поскольку NB не обладает свойством отбора признаков, то возможно использовать для этих задач другой метод. Так, в [9] заявлено увеличение качества классификации NB,если применить к небольшой части обучающей выборки алгоритм C4.5 и выбрать только те признаки, которые участвуют в разбиении на первых уровнях дерева, как наиболее значимые, а затем передать их на вход классификатору.

Стандартные методы отбора признаков являются универсальными и применимы ко многим классификаторам. Существуют и специализированные решения для

более тонкой настройки и модификации NB. В [10] предлагается скомпенсировать невыполненное предположение о независимости признаков с помощью введения дополнительной функции в классификатор. Тогда отношение логарифма апостериорных вероятностей классов можно записать в виде:

$$\log \frac{P(Y=+1|X_1,\ldots,X_n)}{P(Y=-1|X_1,\ldots,X_n)} = \alpha + \sum_{j=1}^{n}(g_j(x_j) + b_j(x_j)),$$

где $\alpha = \log \frac{P(Y=+1)}{P(Y=-1)}$, $g_j(x_j) = \log \frac{f(x_j|Y=+1)}{f(x_j|Y=-1)}$, а $b_j(x_j)$ - гладкие функции для компенсации зависимости, которые обращаются в 0 если наивное предположение выполнено и признаки условно независимы.

Отдельно стоит выделить группу методов,ослабляющих предположение о независимости признаков. Исследование [11] описывает модификацию под названием Tree Augmented Naive Bayes(TAN). В данной модификации признак может зависеть от ограниченного числа других признаков, что позволяет учитывать корреляцию между ними. Степень зависимости определяется по количеству взаимной информации между признаками.

5 Наивный линейный байесовский классификатор

Рассмотрим важный частный случай, когда одномерные плотности $p(x^j|\theta_y^j)$ принадлежат *экспоненциальному семейству плотностей*, т. е. представляются в виде

$$p(x|\theta) = \exp\left(\frac{x\theta - c(\theta)}{\varphi} + h(x,\varphi)\right),$$

где $c(\theta)$, $h(x,\varphi)$ — функциональные параметры распределения, θ и φ — числовые параметры, θ называется параметром *сдвига*, φ — параметром *разброса*.

Экспоненциальное семейство интересно по ряду причин. Во-первых, это очень широкое семейство, ему принадлежат многие полезные распределения. Во-вторых, задача максимизации правдоподобия для экспоненциальных плотностей имеет аналитическое решение, которое вычисляется за время $O(\ell)$. В-третьих, именно в этом случае наивный байесовский классификатор является линейным. Рассмотрим эти свойства подробнее.

Пример 5.1. Нормальное (гауссовское) распределение подходит для описания действительных признаков ($x \in \mathbb{R}$) и соответствует случаю $\theta = \mu$, $c(\theta) = \frac{1}{2}\theta^2$, $\varphi = \sigma^2$:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \exp\left(\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right). \quad (5.1)$$

Пример 5.2. Пуассоновское распределение подходит для описания целочисленных признаков с неограниченным числом значений ($x \in \{0,1,2,\dots\}$) и соответствует случаю $\theta = \ln(\mu)$, $c(\theta) = e^\theta$, $\varphi = 1$:

$$p(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} = \exp\left(\frac{x\ln(\mu) - \mu}{1} - \ln x!\right).$$

Пример 5.3. Биномиальное распределение подходит для описания целочисленных признаков с ограниченным числом значений ($x \in \{0,1,\dots,k\}$) и соответствует случаю $\theta = \ln\frac{\mu}{1-\mu}$, $c(\theta) = k\ln(1+e^\theta)$, $\varphi = 1$:

$$p(x|\mu,k) = C_k^x \mu^x (1-\mu)^{k-x} = \exp\Big(x\ln\tfrac{\mu}{1-\mu} + k\ln(1-\mu) + \ln C_k^x\Big).$$

Пример 5.4. Распределение Бернулли является частным случаем биномиального при $k = 1$, оно подходит для описания бинарных признаков ($x \in \{0,1\}$) и соответствует случаю $\theta = \ln\frac{\mu}{1-\mu}$, $c(\theta) = \ln(1+e^\theta)$, $\varphi = 1$:

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp\Big(x\ln\tfrac{\mu}{1-\mu} + \ln(1-\mu)\Big).$$

В экспоненциальное семейство попадают и другие известные распределения: мультиномиальное, геометрическое, χ^2-распределение, бета-распределение, гамма-распределение, распределение Дирихле, распределение Лапласа с фиксированным математическим ожиданием, и другие. Не являются экспоненциальными распределения Коши, Вейбулла, Стьюдента, гипергеометрическое, и другие.

Обозначим через $\langle f(x_i)\rangle_y$ среднее значение функции от объекта $f(x_i)$ по всем обучающим объектам класса y:

$$\langle f(x_i)\rangle_y = \frac{1}{|X_y|}\sum_{x_i\in X_y} f(x_i),$$

например, $\langle x_i^j\rangle_y$ — среднее значение j-го признака по всем объектам класса y.

Покажем, что задача максимизации правдоподобия (3.1) решается аналитически, если плотности распределения признаков принадлежат экспоненциальному семейству.

Теорема 5.1. *Если одномерные плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству и $\Theta = (\theta_y^j)$ является точкой максимума правдоподобия* (3.1)*, то*

$$\theta_y^j = [c']^{-1}\big(\langle x_i^j\rangle_y\big).$$

Доказательство.

Подставим логарифмы экспоненциальных плотностей

$$\ln p(x^j|\theta_y^j) = \frac{x^j\theta_y^j - c(\theta_y^j)}{\varphi_y^j} + h(x^j,\varphi_y^j)$$

в необходимое условие максимума правдоподобия:

$$\frac{\partial\mathcal{L}}{\partial\theta_y^j} = 0; \qquad \frac{\partial}{\partial\theta_y^j}\sum_{j=1}^{n}\sum_{y\in\mathbb{Y}}\sum_{x_i\in X_y}\ln p(x_i^j|\theta_y^j) = 0.$$

Продифференцировав по θ_y^j, получим аналитическое решение для параметра θ_y^j:

$$\frac{1}{\varphi_y^j}\sum_{x_i\in X_y}\big(x_i^j - c'(\theta_y^j)\big) = 0, \qquad c'(\theta_y^j) = \frac{1}{|X_y|}\sum_{x_i\in X_y} x_i^j = \langle x_i^j\rangle_y,$$

откуда следует утверждение теоремы. ■

Таким образом, для получения оценок максимального правдоподобия параметров θ_y^j по обучающей выборке необходимо вычислить среднее значение каждого признака j по обучающим объектам каждого из классов y, что занимает время $O(n\ell)$.

Теорема 5.2. *Пусть* $\mathbb{Y} = \{-1, +1\}$, *плотности* $p(x^j|\theta_y^j)$ *принадлежат экспоненциальному семейству распределений и параметры разброса не зависят от класса,* $\varphi_y^j = \varphi^j$. *Тогда наивный байесовский классификатор представляется в линейном виде*

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} x^j w_j - w_0\right),$$

причём весовые коэффициенты w_j *выражаются через параметры распределений:*

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y\theta_y^j, \qquad j = 1, \ldots, n. \tag{5.2}$$

Доказательство.

Запишем формулу двухклассового классификатора (2.1) в виде

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} \ln p(x^j|\theta_+^j) - \ln p(x^j|\theta_-^j) + \ln \tfrac{P(+1)}{P(-1)}\right),$$

подставим в неё экспоненциальные плотности и перегруппируем слагаемые:

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} x^j \underbrace{\left(\tfrac{\theta_+^j}{\varphi_+^j} - \tfrac{\theta_-^j}{\varphi_-^j}\right)}_{w_j} + \underbrace{h(x^j, \varphi_+^j) - h(x^j, \varphi_-^j)}_{0} + \underbrace{\tfrac{c(\theta_-^j)}{\varphi_-^j} - \tfrac{c(\theta_+^j)}{\varphi_+^j} + \ln \tfrac{P(+1)}{P(-1)}}_{-w_0}\right).$$

Получили линейный классификатор с весами признаков w_j и свободным членом w_0.

Теорема доказана. ∎

Свободный член w_0 в линейных классификаторах обычно выбирается из соображений баланса чувствительности и специфичности, и его выбор зависит от соотношения цены ошибок первого и второго рода.

Подстановка оценок максимума правдоподобия в формулу линейного классификатора позволяет исключить этап явного вычисления параметров распределений θ_y^j и вычислять непосредственно весовые коэффициенты w_j по обучающей выборке:

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y[c']^{-1}\big(\langle x_i^j \rangle_y\big), \tag{5.3}$$

где φ^j — оценка параметра разброса, общая для всех классов. Если параметр разброса не является константой, то его оценка максимального правдоподобия находится из уравнения, которое для некоторых распределений удаётся решить аналитически:

$$(\varphi^j)^2 \sum_{i=1}^{\ell} h'_\varphi(x_i^j, \varphi^j) = \sum_{i=1}^{\ell} \big(x_i^j \theta_{y_i}^j - c(\theta_{y_i}^j)\big).$$

Например, для гауссовского распределения (5.1) оценка максимального правдоподобия параметра φ^j равна средневзвешенной по всем классам дисперсии значений j-го признака:

$$\varphi^j = \sum_{y\in\mathbb{Y}} \frac{|X_y|}{\ell} \big\langle x_i^j - \langle x_i^j\rangle_y \big\rangle_y^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} \big(x_i^j - \langle x_i^j\rangle_{y_i}\big)^2. \tag{5.4}$$

6 Регуляризация наивного байесовского классификатора

Наивный байесовский классификатор при условии экспоненциальности распределений признаков остаётся линейным независимо от того, как оцениваются параметры распределений. Вместо несмещённых оценок максимального правдоподобия могут быть использованы другие оценки. Общим подходом к получению смещённых оценок является регуляризация. Обычно регуляризация применяется для обеспечения устойчивости решения некорректно поставленных задач. В нашем случае регуляризация используется для того, чтобы сместить решение в сторону оптимума другого критерия качества и тем самым уйти от избыточно жёсткого ограничения независимости признаков. При этом сохраняется линейная вычислительная сложность наивного байесовского классификатора.

Пусть дополнительный критерий качества задан в виде требования максимизации функционала $\mathcal{R}(w) \to \max$. Если критериев несколько, будем максимизировать их взвешенную сумму с коэффициентами регуляризации τ_k:

$$\mathcal{R}(w) = \sum_{k=1}^{K} \tau_k \mathcal{R}_k(w) \to \max_w.$$

Согласно (5.2), вектор весов линейного классификатора является функцией от параметров распределений, $w = w(\Theta)$. Поэтому дополнительные ограничения можно накладывать на вектор весов, затем решать задачу максимизации регуляризованного правдоподобия по совокупности параметров Θ,

$$\mathcal{L}(\Theta) + \mathcal{R}(w(\Theta)) = \sum_{j=1}^{n}\sum_{y\in\mathbb{Y}}\sum_{x_i\in X_y} \left(\frac{x_i^j\theta_y^j - c(\theta_y^j)}{\varphi_y^j}\right) + \mathcal{R}(w(\Theta)) \;\to\; \max_{\Theta}, \tag{6.1}$$

и из полученного решения Θ получать вектор весов $w = w(\Theta)$ согласно (5.2).

Теорема 6.1. *Пусть $\mathbb{Y} = \{-1, +1\}$, плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству распределений и параметры разброса не зависят от класса, $\varphi_y^j = \varphi^j$.*

Тогда точка максимума регуляризованного правдоподобия (6.1) *удовлетворяет системе уравнений*

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y[c']^{-1} \left(\langle x_i^j \rangle_y + \frac{y}{|X_y|} \frac{\partial \mathcal{R}}{\partial w_j} \right). \tag{6.2}$$

Доказательство.

Запишем необходимое условие максимума по параметру θ_y^j:

$$\frac{\partial}{\partial \theta_y^j}(\mathcal{L} + \mathcal{R}) = 0; \qquad \sum_{x_i \in X_y} \left(\frac{x_i^j - c'(\theta_y^j)}{\varphi^j} \right) + \frac{y}{\varphi^j} \frac{\partial \mathcal{R}}{\partial w_j} = 0$$

и выразим отсюда параметр θ_y^j:

$$c'(\theta_y^j) = \frac{1}{|X_y|} \left(\sum_{x_i \in X_y} x_i^j + y \frac{\partial \mathcal{R}}{\partial w_j} \right) = \langle x_i^j \rangle_y + \frac{y}{|X_y|} \frac{\partial \mathcal{R}}{\partial w_j}.$$

Подставляя параметр θ_y^j в (5.2), получим требуемое уравнение относительно w_j.

Теорема доказана. ■

Система уравнений (6.2) в общем случае решается численно. Для некоторых регуляризаторов решение может быть найдено очень быстро, без обращения к исходным данным на каждой итерации, так как средние значения $\langle x_i^j \rangle_y$ достаточно вычислить один раз. Если регуляризатор аддитивен по признакам, $\mathcal{R}(w) = \sum_{j=1}^n \mathcal{R}_j(w_j)$, то система распадается на n независимых уравнений по параметрам w_j. Для некоторых представителей экспоненциального семейства систему удаётся решить аналитически.

Для выбора коэффициента регуляризации τ будем использовать кроссвалидацию. На этапе обучения будем вычислять весовые коэффициенты $w_j(\tau_s)$ для заданной сетки значений $\tau = \tau_s$, $s = 1, \ldots, S$. При этом все вычисления, требующие обращения к данным, можно сделать один раз для всех значений τ_s. Для выбора оптимального значения τ_s качество классификации будем оценивать по проверочной выборке.

Рассмотрим частные случаи регуляризаторов.

1. При отсутствии регуляризации, $\mathcal{R}(w) = 0$, система уравнений (6.2)
2. Сжимающий регуляризатор по L_1-норме:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^n |w_j|.$$

Этот регуляризатор применяется также для устранения эффектов мультиколлинеарности и переобучения, но имеет важный побочный эффект отбора признаков.

По мере увеличения коэффициента τ увеличивается число коэффициентов w_j, обращающихся в нуль. Этот регуляризатор применяется в методе LASSO.

Для того, чтобы к L_1-регуляризатору можно было применить (6.2) необходимо провести замену переменных:

$$\begin{cases} u_j = \frac{1}{2}(|w_j| + w_j) \\ v_j = \frac{1}{2}(|w_j| - w_j) \end{cases}$$

После замены :

$$w_j = u_j - v_j,$$

$$|w_j| = u_j + v_j.$$

Таким образом

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} (v_j + u_j).$$

К такому регуляризатору применима **Теорема 6.1.** Подстановка $\mathcal{R}(w)$ в (6.2) приводит к уравнению относительно w_j

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y[c']^{-1} \left(\langle x_i^j \rangle_y - \frac{\tau y}{|X_y|} \operatorname{sign} w_j \right),$$

которое решается численно. Если распределения признаков нормальны, то решение, как и в предыдущем случае, выписывается аналитически:

$$w_j = \operatorname{sign}(w_{0j}) \max\big(|w_{0j}| - \tfrac{\tau'}{\varphi^j}, 0\big), \qquad w_{0j} = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y \langle x_i^j \rangle_y,$$

где w_{0j} — нерегуляризованный весовой коэффициент.

Таким образом, L_1-регуляризатор также приводит к сжатию вектора коэффициентов, однако теперь коэффициенты w_j обращаются в нуль при $|w_{0j}|\varphi^j < \tau'$.

Отдельно рассмотрим эвристику, а именно сжимающий регуляризаотр по L_0-норме:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} [|w_j| > 0],$$

$[w_j > 0] = \#\left\{ j = 1...|Y| \;\middle|\; w^j \neq 0 \right\}$ —количество ненулевых весов для каждого класса. Основной идеей данной регуляризации является удаление признаков с малым весом, неинформативных признаков. Параметр τ находится перебором, следовательно, его всегда можно таким образом, чтобы обнулить определенное количество наименее

информативных признаков. В таком случае, вместо параметра τ можно рассматривать просто k — количество наиболее информативных признаковов, обнуление всех остальных, кроме них, упрощает работу классификатора и незначительно влияют на качество классификации. В качестве неинформативных признаков рассматриваются те, у которые наименьший по модулю вес. Таким образом, работа регуляризатора заключается в сортировке по модулю убывания весов и обнулении $n - k$ наименее значимых из них.

6.1 Многоклассовая классификация

Существует два основных подхода для сведения многоклассовой классификации с непересекающимися классами к бинарной.

- **One-vs-All approach** (OVA) заключается в обучении N классификаторов по следующему принципу

$$a_j(x) = \begin{cases} \geqslant 0, & \text{если } f(x) = y, \\ < 0, & \text{если } f(x) \neq y, \end{cases}$$

которые отделяют каждый класс от остальных. Далее, для каждого $x \in X$ вычисляются все классификаторы и выбирается класс, соответствующий классификатору с большим значением:

$$a(x) = \arg\max_{j \in \mathbb{Y}} a_j(x).$$

- **One-vs-One approach**. Его так же называют All-vs-All (AVA) approach. В этом случае строятся $N(N-1)$ классификаторов, которые разделяют объекты пар различных классов:

$$f_{jz}(x) = \begin{cases} +1, & \text{если } f(x) = j, \\ -1, & \text{если } f(x) = z. \end{cases}$$

После обучения бинарных классификаторов решение принимается следующим образом:

$$a(x) = \arg\max_{j \in \mathbb{Y}} \sum_{\substack{z=1,\ldots,N \\ z \neq j}} f_{jz}(x).$$

Сравнение первых двух подходов проведено в [12].

Для сведения задачи многоклассовой классификации с пересекающимися классами к бинарной воспользуемся комбинацией двух вышеописанных подходов. В задачах дифференциальной медицинской диагностики существует выделенный класс, а именно класс «абсолютно здоровых», именно с ним сравниваются все болезни при подходе One-vs-All approach. Таким образом, получена $K-1$ подзадача бинарной классификации. Для каждого класса i обучающей подвыборкой является не вся выборка, а только та ее часть, которой соответствует "эталонный"и i-ый классы. Таким образом, каждый классификатор будет давать оценку принадлежности объекта к классу i. Для того, чтобы хорошо решать задачу сравнения болезней, воспользуемся подходом All-vs-All как регуляризатором.

6.2 Регуляризация наивного байсовского многоклассового классификатора

Основная задача многоклассового регуляризатора - как можно лучше различать классы, дистанцировать все болезни друг от друга. Это можно добиться, если сделать векторы весов у всех классов как можно уникальнее. Такой регуляризатор получатается путём минимизации всевозможных попарных скалярных произведений весов:

$$R = -\sum_{j=1}^{n}\sum_{y>z} w_{yj}w_{zj} \longrightarrow \max.$$

Зафиксируем класс u:

$$R = -\sum_{j=1}^{n}\sum_{y\in\mathbb{Y}/u} w_{uj}w_{yj} \longrightarrow \max.$$

При фиксированном классе u данный регуляризатор является так же регуляризатором для бинарной задачи. Учтем, что

$$w_j = \left(\frac{\theta_+^j}{\varphi_+^j} - \frac{\theta_-^j}{\varphi_-^j}\right),$$

где $\theta_-^j = [c']^{-1}\left(\langle x_i^j\rangle_y + \frac{\tau y}{|X_y|}\frac{\partial R}{\partial w_j}\right)$,

$\theta_+^j = [c']^{-1}\left(\langle x_i^j\rangle_y + \frac{\tau y}{|X_y|}\frac{\partial R}{\partial w_j}\right)$.

Таким образом, применив (6.2), получили:

$$w_{ju} = -\frac{1}{\varphi^j}\sum_{u\in\mathbb{Y}}[c']^{-1}\left(\langle x_i^j\rangle_y + \frac{\tau}{|X_y|}\sum_{y\in\mathbb{Y}/u} w_{yu}\right)$$

7 Вычислительные эксперименты

В экспериментах используются данные дифференциальной медицинской диагностики. Выборка состоит из 2016 обследуемых и 19 классов. Данные получены из обработанных при помощи технологии информационного анализа [3] кардиосигналов. Технология информационного анализа электрокардиосигналов основана на преобразовании каждой электрокардиограммы сначала в последовательность интервалов и амплитуд кардиоциклов, а затем — в символьную последовательность фиксированной длины, называемую кодограммой. Кодограмма разбивается на слова длинною в три символа - триграмму. В качестве признаков в нижеописанных экспериментах рассматриваются частоты встречаемости триграмм.

Аббревиатура	Мощность	Болезнь
АЗ	193	абсолютно здоров
ВДЭ	694	вегетососудистая дистония
ГБК	324	асептический некроз головки бедренной кости
ГБЭ	1894	гипертоническая болезнь
ГДЭ	324	хронический гастрит (гастродуоденит) гиперацидный
ДЖЭ	717	дискинезия желчевыводящих путей
ЖКЭ	278	желчнокаменная болезнь
ИБЭ	1265	ишемическая болезнь сердца
МКЭ	654	мочекаменная болезнь
ММЭ	781	миома матки
РОЭ	530	рак общий (онкопатология различной локализации)
СДЭ	871	сахарный диабет (СД1 и СД2)
УЩЭ	748	узловой (диффузный) зоб щитовидный железы
ХГЭ	700	хронический гастрит (гастродуоденит) гипоацидный
ХХЭ	340	холецистит хронический
ЭА	260	анемия железодефицитная
ЭАП	260	аденома простаты
ЭАХ	276	аднексит хронический
ЯБЭ	785	язвенная болезнь
ВСЕ	2016	все обследуемые

Таблица 1: Заболевания, их аббревиатуры и объёмы выборок

Цель экспериментов —повысить качество дифференциальной диагностики при помощи регуляризации линейного наивного байсовского классификатора. Провести эксперименты для различных распределий из экспоненциального семейства плотностей.

Для оценки качества работы классификатора предлагается строить ROC-кривую и считать AUC. Выборку разбивают при помощи $n \times k$-fold Cross Validation. Итоговый AUC получается усреднением по всем n, k и классам.

7.1 Отбор признаков

В данном эксперименте будем проводить отбор признаков при при помощи L_0-регуляризации. Алгоритм заключается в сортировке весов по убыванию модуля и построении классификатора, который использует только $1, \ldots, n$ наиболее информативных признака. Далее находится среднее по всем подвыборкам оптимальное число используемых признаков — минимальное количество признаков, при котором AUC отличается от наилучшего меньше, чем на один процент. После расчета числа признаков для всех классов строится NB на всех данных и веса сортируются по убыванию модуля. Таким образом определяется, какие признаки обнуляются для каждого класса.

На графиках ниже предоставлены графики зависимости AUC от число признаков для различных распределений из экспоненциального семейства для обучающей выборки. Для распределения Пуассона в среднем достаточно использовать 11 признаков, тогда как для распределения Бернулли и нормального их количество возрастает до 95 и 80 соответственно.

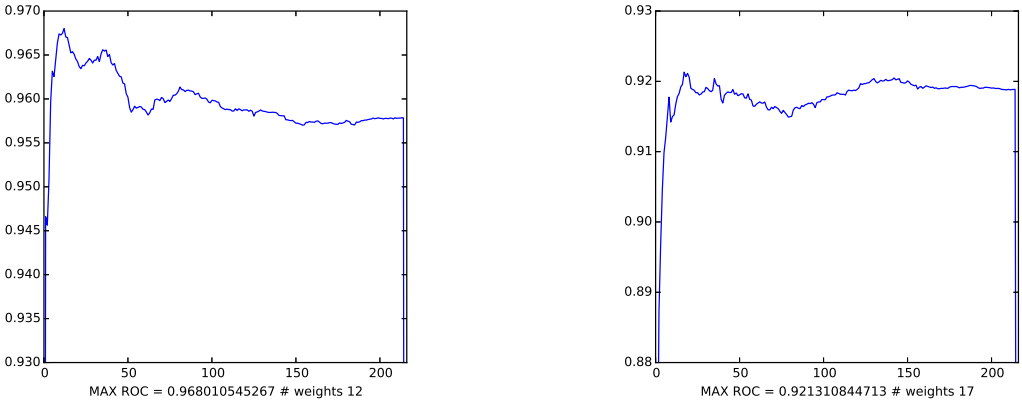

Рис. 1: Зависимость значения AUC от количества признаков для болезни ГБЭ(слева) и ЭА(справа) в случае распределения Пуассона

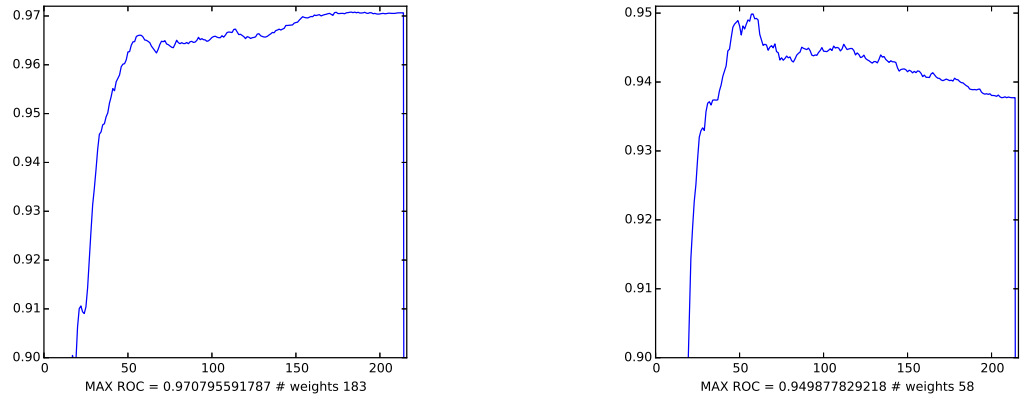

Рис. 2: Зависимость значения AUC от количества признаков для болезни ГБЭ(слева) и ЭА(справа) в случае нормального распределения

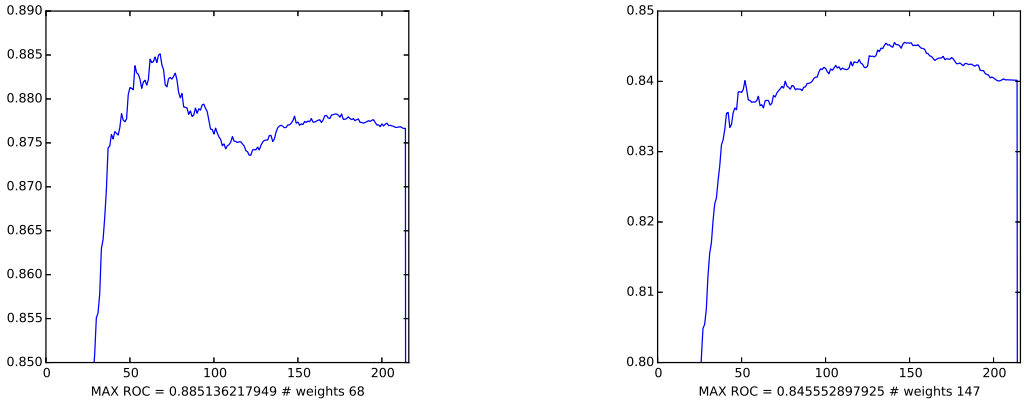

Рис. 3: Зависимость значения AUC от количества признаков для болезни ГБЭ(слева) и ЭА(справа) в случае распределения Бернулли

7.2 Подбор коэффициентов регуляризации

Для L_1 и многоклассовой регуляризации(MR) необходимо подобрать оптимальные коэффициенты τ. Это делается перебором, причем учитывается, что при использовании регуляризаций последовательно, коэффициенты зависят друг от друга. На графиках ниже предоставлена зависимость AUC от τ.

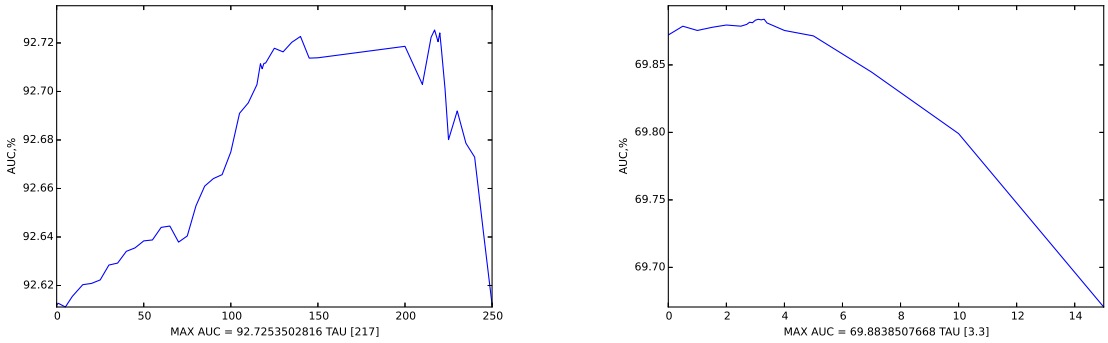

Рис. 4: Зависимость значения AUC от константы τ для L_1- регуляризатора(слева) и многоклассового регуляризатора(справа).

7.3 Результаты

В данном эксперименте строится NB-классификатор для трех распределений из экспоненциального семейства: нормального, пуассоновского и Бернулли. Для каждого из них был проведен отбор признаков при помощи L_0 и L_1-регуляризаций. Далее проводилась многоклассовая регулиризация(MR). Результаты приведены в таблицах ниже. Видно, что распределение Пуассона и нормальное дает схожие результаты, которым занчительно уступает распределение Бернулли. Для всех рассмотренных распределений лучший результат дает NB с комбинацией L_1 и MR, однако, L_0 вместе с MR незначительно им уступают, но при этом работают меньше, чем на половине признаков.

Болезнь	Pois	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8340	0.8202(13)	**0.8401**(203)	0.8245(13)	0.8385(202)
ГБК	0.9578	0.9516(4)	0.9579(157)	0.9526(4)	**0.9593**(157)
ГБЭ	0.9464	0.9454(14)	0.9479(215)	0.9466(14)	**0.9580**(214)
ГДЭ	0.9280	0.9196(14)	**0.9298**(162)	0.9178(14)	0.9295(162)
ДЖЭ	0.9188	0.9157(7)	0.9215(201)	0.9182(7)	**0.9259**(201)
ЖКЭ	0.9682	0.9604(16)	0.9704(135)	0.9670(16)	**0.9798**(135)
ИБЭ	0.9626	0.9357(17)	0.9636(209)	0.9415(17)	**0.9699**(207)
МКЭ	0.9290	0.9272(17)	0.9308(194)	0.9279(17)	**0.9386**(193)
ММЭ	0.9120	0.9099(7)	0.9130(206)	0.9129(7)	**0.9193**(205)
РОЭ	0.9456	0.9425(12)	**0.9112**(181)	0.9561(12)	0.9106(180)
СДЭ	0.9524	0.9511(7)	0.9572(205)	0.9573(7)	**0.9604**(205)
УЩЭ	0.9318	0.9275(8)	0.9328(206)	0.9301(8)	**0.9335**(206)
ХГЭ	0.9323	0.9291(17)	0.9317(197)	0.9314(17)	**0.9342**(197)
ХХЭ	0.9324	0.9292(8)	0.9420(157)	0.9313(8)	**0.9496**(157)
ЭА	0.8696	0.8631(8)	**0.8802**(156)	0.8761(8)	0.8764(155)
ЭАП	0.9515	0.8509(9)	0.9526(138)	0.8611(9)	**0.9578**(137)
ЭАХ	0.8806	0.8660(16)	**0.8873**(160)	0.8852(16)	0.8834(1158)
ЯБЭ	0.9160	0.9122(10)	0.9212(205)	0.9104(10)	**0.9265**(205)
ВСЕ	0.9261	0.9214 (11)	0.9280(194)	0.9286(11)	**0.9291**(193)

Таблица 2: Результаты регуляризованного NB для пуассоновского распределения

Болезнь	Norm	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8316	0.8001 (123)	0.8348(216)	0.8079(123)	**0.8431**(215)
ГБК	0.8471	0.8354 (99)	0.8540(215)	0.8411(99)	**0.8588**(214)
ГБЭ	0.8653	0.8602 (117)	0.8674(216)	0.8621(117)	**0.8701**(216)
ГДЭ	0.8804	0.8782 (86)	0.8837(215)	0.8808(85)	**0.8842**(215)
ДЖЭ	0.8910	0.8895 (105)	**0.8844**(214)	0.8921(105)	0.8820(214)
ЖКЭ	0.9006	0.8990 (111)	**0.9178**(216)	0.9010(111)	0.9171(216)
ИБСЭ	0.8521	0.8399 (105)	0.8592(215)	0.8393(105)	**0.8607**(214)
МКЭ	0.9106	0.9010 (111)	**0.9278**(215)	0.9010(111)	0.9271(215)
ММЭ	0.9028	0.8959 (107)	**0.9108**(214)	0.8961(107)	0.9093(214)
РОЭ	0.9428	0.8859 (72)	0.9112(216)	0.8861(72)	**0.9106**(216)
СДЭ	0.8421	0.8390 (100)	0.8462(215)	0.8408(100)	**0.8479**(215)
УЩЭ	0.8807	0.8791 (78)	0.8819(214)	0.8825(78)	**0.8898**(214)
ХГЭ	0.8646	0.8498 (107)	0.8698(213)	0.8481(107)	**0.8719**(213)
ХХЭ	0.9021	0.8991 (99)	**0.9157**(215)	0.9053(99)	0.9134(215)
ЭА	0.9238	0.9159 (96)	0.9292(215)	0.9201(96)	**0.9309**(215)
ЭАП	0.9006	0.8973 (100)	0.9092(216)	0.9061(100)	**0.9109** (215)
ЭАХ	0.8427	0.8395 (102)	0.8502(216)	0.8420(102)	**0.8576**(216)
ЯБЭ	0.8602	0.8552 (99)	**0.8672**(215)	0.8661(99)	0.8668(215)
ВСЕ	0.8825	0.8749 (95)	0.8893(215)	0.8808(95)	**0.8903**(215)

Таблица 4: Результаты регуляризованного NB для распределения Бернулли

Болезнь	Norm	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8402	0.8290 (48)	0.8448(215)	0.80298(48)	**0.8478**
ГБК	0.9598	0.9454 (50)	0.9594(214)	0.9432(50)	**0.9631**(212)
ГБЭ	0.9389	0.9281 (112)	0.9397(216)	0.9264(112)	**0.9421**(215)
ГДЭ	0.9378	0.9323 (40)	0.9387(215)	0.9346(40)	**0.9412**(214)
ДЖЭ	0.9176	0.8993 (107)	**0.9214**(216)	0.8996(107)	0.9201(216)
ЖКЭ	0.9706	0.9696 (50)	0.9716(215)	0.9786(50)	**0.9791**(215)
ИБСЭ	0.9588	0.9464 (105)	0.9592 (216)	0.9493(105)	**0.9620**(216)
МКЭ	0.9267	0.9140 (108)	**0.9278** (216)	0.9210(108)	0.9271(215)
ММЭ	0.9103	0.9031 (115)	0.9112(216)	0.9161(115)	**0.9116**(215)
РОЭ	0.9418	0.9239 (106)	**0.9112**(216)	0.9261(105)	0.9209(216)
СДЭ	0.9503	0.9455 (89)	0.9112(214)	0.9471(88)	**0.9516**(214)
УЩЭ	0.9294	0.9079 (110)	0.9302(215)	0.9091(110)	**0.9304**(215)
ХГЭ	0.9296	0.9133 (115)	0.9315(216)	0.9161(115)	**0.9357**(214)
ХХЭ	0.9310	0.9180 (97)	**0.9323**(216)	0.9172(97)	0.9210(216)
ЭА	0.8766	0.8713 (44)	0.8812(216)	0.8861(44)	**0.8813** (215)
ЭАП	0.9488	0.9279 (110)	0.9492(215)	0.9283(110)	**0.9506** (214)
ЭАХ	0.8888	0.8825 (37)	**0.9001**(216)	0.8892(37)	0.8962(215)
ЯБЭ	0.9143	0.9043 (95)	**0.9145**(214)	0.9161(95)	0.9123(213)
ВСЕ	0.9262	0.9145 (80)	0.9273(215)	0.9258(80)	**0.9303**(214)

Таблица 3: Результаты регуляризованного NB для нормального распределения

8 Результаты, выносимые на защиту

- Предложен метод аддитивной регуляризации наивного байесовского классификатора
- Предложены регуляризаторы для отбора признаков и для повышения различности векторов весов признаков в многоклассовой классификации.
- Показано, что разработанные методы повышают качество дифференциальной диагностики заболеваний при использовании технологии информационного анализа электрокардиосигналов.
- Выявлено, что для распределения Пуассона необходимо минимальное количество признаков для задачи дифференциальной медицинской диагностики.

9 Литература

Список литературы

[1] Р. Дуда and П. Харт. *Распознавание образов и анализ сцен.* М.: Мир, 1976.

[2] Chang-Hwan Lee, Fernando Gutierrez, and Dejing Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In Diane J. Cook, Jian Pei, Wei Wang 0010, Osmar R. Zaiane, and Xindong Wu, editors, *ICDM*, pages 1146–1151. IEEE Computer Society, 2011.

[3] Uspensky K. Vorontsov V. Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. In *Математическая биология и биоинформатика*, 2014.

[4] Lutu P. E. N. Fast feature selection for naive bayes classi cation in data stream mining. *Proceedings of the World Congress on Engineering*, 3, 2013.

[5] Jasmina N. The impact of feature selection on the accuracy of naive bayes classifier. *2010 18TH TELECOMMUNICATIONS FORUM (TELFOR).*, page 1113–1116, 2010.

[6] Elissee A. Guyon I. An introduction to variable and feature selection. *Journal of Machine Learning Research.*, 3:1157–1182, 2003.

[7] Puerta J. M. Bermejo P., Gamez J. A. Speeding up incremental wrapper feature subset selection with naive bayes classi er. *Knowledge-Based Systems.*, 55:140–147, 2014.

[8] Nikravesh M. Guyon I., Gunn S. Feature extraction, foundations and applications. *Series Studies in Fuzziness and Soft Computing, Physica-Verlag*, 2006.

[9] Chotirat (Ann) Ratanamahatana and Dimitrios Gunopulos. Feature selection for the naive bayesian classifier using decision trees. *Applied Artificial Intelligence*, 17(5-6):475–487, 2003.

[10] Kim Larsen. Generalized naive bayes classifiers. *SIGKDD Explorations*, 7(1):76–81, 2005.

[11] Nir Friedman, Dan Geiger, Moises Goldszmidt, G. Provan, P. Langley, and P. Smyth. Bayesian network classifiers. In *Machine Learning*, pages 131–163, 1997.

[12] Duin R. P. W. Tax D. M. J. Using two-class classi ers for multiclass classification. *ICPR.*, (2):124–127, 2002.