**© 2019 г. В.Н. ВАПНИК, д-р техн. наук (vladimir.vapnik@gmail.com)**
**(Колумбийский университет, Нью-Йорк, США;**
**Отдел исследования ИИ Фэйсбук, Нью-Йорк, США)**

# ПОЛНАЯ СТАТИСТИЧЕСКАЯ ТЕОРИЯ ОБУЧЕНИЯ

*Памяти выдающегося ученого и замечательного человека Я.З. Цыпкина*

Описывается одновременное решение двух задач выбора для функций из гильбертова пространства с воспроизводящим ядром: найти в *данном подмножестве* допустимых функций функцию, которая минимизирует средние потери; используя обучающую выборку, из обширного множества функций в гильбертовом пространстве выбирается допустимое подмножество, включающее в себя искомую функцию, затем в этом допустимом подмножестве выбирается хорошая аппроксимация данной функции. Получено аналитическое решение.

*Ключевые слова*: статистическая теория обучения, первая задача выбора, вторая задача выбора, гильбертово пространство с воспроизводящим ядром, обучающая выборка.

**DOI:** 10.1134/S0005231019110023

## 1. Введение

Современная статистическая теория обучения ставит перед собой проблему поиска на основании эмпирических данных функции, доставляющей возможно меньшую величину среднему значению заданной функции потерь. В зависимости от вида функции потерь решаются либо задача классификации, либо задача построения функции регрессии, либо задача оценивания условной функции вероятности. Несмотря на качественное различие этих задач, в [1–3] развит единый подход к их решению, позволяющий путем минимизации величины функции потерь, вычисленной на основании эмпирических данных, находить функцию, для которой среднее значение функции потерь близко к минимальному. Мера близости, естественно, зависит от количества эмпирических данных и от свойств подмножества допустимых функций, в котором ищется решение поставленной задачи.

Часто выбор подмножества допустимых функций рассматривается как проблема, лежащая вне теории обучения, относя ее к предметной части решаемой задачи. Методы теории обучения используются лишь для снижения размерности задачи — уменьшения числа оцениваемых параметров или построения малого числа простых зависимостей, в комбинации дающих искомую функцию. При этом повышается статистическая надежность ответа, но и, возможно, возрастает смещение найденного решения относительно истинного.

Проблема задания подмножества допустимых функций в зависимости от точности эмпирических данных известна в функциональном анализе в области решения некорректных задач [4]. Метод регуляризации задает процедуру поиска решения в подмножестве функций с ограниченной нормой. В зависимости от вида нормы получаются устойчивые решения в различных функциональных пространствах. Аналогичный подход применим и в статистической теории обучения.

В статье рассматривается общая математическая модель обучения, в рамках которой одновременно решаются как задача выбора подмножества допустимых функций, учитывающего специфику эмпирических данных, так и задача поиска в этом подмножестве решения поставленной задачи. При решении задач машинного обучения как с теоретической, так и с прикладной точек зрения оказывается удобно использовать гильбертовы пространства с воспроизводящим ядром, описываемые далее. При этом задача минимизации в бесконечномерном функциональном пространстве величины функции потерь, вычисленной на основании эмпирических данных, сводится к минимизации регуляризованного функционала в конечномерном пространстве, а искомое решение представляется в виде разложения по функциям воспроизводящего ядра.

## 2. Ограничения существующей статистической теории обучения

Пятьдесят лет назад была представлена так называемая VC теория обучения (теория Вапника–Червоненкиса)[1] [1, 2]. Эта теория рассматривает следующую математическую задачу:

Допустим, что в данном множестве индикаторных функций[2] $y = \theta(f(x,\alpha))$, $\alpha \in \Lambda$, функция $y = \theta(f(x,\alpha_\ell))$ минимизирует число ошибочных классификаций $y$ в $\ell$ наблюдаемых парах

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \quad x \in \mathrm{R}^n,\ y \in \{0, 1\},$$

где пары $(x_i, y_i)$ генерируются случайно и независимо в соответствии с неизвестной функцией распределения $P(x, y) = P(y|x)P(x)$.

Определим *эмпирическую функцию потерь* правила $\theta(f(x,\alpha_*))$ как

$$P_\ell(\alpha_*) = \sum_{i=1}^{\ell} |y_i - \theta(f(x_i,\alpha_*))|$$

и *функцию средних потерь* относительно распределения $P(y, x)$ как

$$P(\alpha_*) = \int |y - \theta(f(x,\alpha_*)| dP(x, y).$$

[1] Теория Вапника–Червоненкиса также называется Статистической теорией обучения.

[2] Здесь используется обозначение индикаторной функции как $y = \theta(f(x,\alpha))$, $\alpha \in \Lambda$, где $\theta(u)$ — ступенчатая функция: $\theta(u) = 1$, если $u \geqslant 0$, и $\theta(u) = 0$, если $u < 0$. Функции $f(x,\alpha)$, $\alpha \in \Lambda$, являются параметрическим подмножеством действительных функций, где $\alpha \in \Lambda$ — абстрактный параметр. Таким образом, нет никаких ограничений на выбор подмножества функций.

Вопрос состоит в том, когда с вероятностью $1-\eta$ можно утверждать, что наименьшее значение эмпирической функции потерь сходится к наименьшему значению функции средних потерь с ростом числа наблюдений $\ell$. Другими словами, когда точность выбранной аппроксимирующей функции $\varepsilon$-близка к точности наилучшей функции для данного множества, т.е. когда

$$\lim_{\ell\to\infty} P\left\{\left|\min_{\alpha\in\Lambda} P_\ell(\alpha) - \min_{\alpha\in\Lambda} P(\alpha)\right| > \varepsilon\right\} = 0 \quad \forall\varepsilon > 0. \tag{2.1}$$

Требуется найти необходимые и достаточные условия, которые обеспечивают выполнение (2.1), и оценить скорость сходимости.

VC теория показала, что для обеспечения условия (2.1), т.е. состоятельности методов, которые минимизируют количество ошибок обучения (так называемые методы минимизации эмпирического риска), необходимо и достаточно, чтобы специально определенная мера мощности данного (бесконечного) множества функций $y = \theta(f(x,\alpha)), \alpha\in\Lambda$ (так называемая комбинаторная размерность $h$ этого множества), была конечна (определение комбинаторной размерности и ее значение для теории обучения обсуждается далее).

Было показано, что с вероятностью $1-\eta$ одновременно для всех функций из множества $\theta(f(x,\alpha), \alpha\in\Lambda$, комбинаторная размерность которого $h^*$, справедлива оценка

$$P(\alpha) \leqslant P_\ell(\alpha) + \frac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4P_\ell(\alpha)}{\varepsilon}}\right), \tag{2.2}$$

где

$$\varepsilon = O\left(\frac{h^* - \ln\eta}{\ell}\right).$$

Так как (2.2) справедлива для всех $\alpha\in\Lambda$, то она также справедлива для функций $\theta(f(x,\alpha_\ell))$, которые минимизируют частоту $P_\ell(\alpha_\ell)$ ошибки обучения. В соответствии с (2.2) эта функция определяет для данного множества функций наименьшую гарантированную (с вероятностью $1-\eta$) вероятность ошибочной классификации.

Оценка (2.2) показывает, что для гарантии успеха обучения требуется, чтобы комбинаторная размерность допустимого множества функций $\theta(f(x,\alpha))$, $\alpha\in\Lambda$, была мала (чтобы $\varepsilon$ в (2.2) было мало), и чтобы допустимое множество функций содержало функцию, "хорошо" аппроксимирующую $y = \theta(f(x,\alpha_\ell))$ (чтобы значение $P_\ell(\alpha_\ell)$ было мало).

Поэтому полное решение задачи машинного обучения требует решения не одной, как в существующих теориях обучения, а двух задач выбора:

1. Выбрать из широкого множества функций (с бесконечной комбинаторной размерностью) малое допустимое подмножество функций, у которого комбинаторная размерность мала и содержит хорошую аппроксимирующую функцию $\theta(f(x,\alpha_\ell))$ (с малым значением $P_\ell(\alpha_\ell)$);

2. В допустимом подмножестве функций $\theta(f(x,\alpha)), \alpha \in \Lambda$, выбрать наилучшую аппроксимирующую функцию (для которой $P_\ell(\alpha_\ell)$ мало).

Существующие методы обучения главным образом посвящены решению второй задачи выбора — выбору наилучшей аппроксимирующей функции в подмножестве[3] и не рассматривают первую, возможно более важную, задачу выбора — выбор допустимого подмножества, которое содержит хорошую аппроксимирующую функцию (с малым значением функции потерь) с малой комбинаторной размерностью.

Метод решения первой задачи выбора — основная тема статьи. Однако перед тем как начать обсуждение первой задачи выбора, опишем некоторые результаты VC теории, которые используются в этой статье.

## 3. Некоторые результаты VC теории

### *3.1. Определение комбинаторной размерности*

В англоязычной литературе для обозначения комбинаторной размерности принят термин VC dimension (размерность Вапника–Червоненкиса). Пусть функции $f(x,\alpha)$, $\alpha \in \Lambda$, определены на $x \in X \in \mathrm{R}^n$. Рассмотрим множество индикаторных функций $\theta(f(x,\alpha))$, $\alpha \in \Lambda$. Известно, что $\ell$ различных векторов $x_1, \ldots, x_\ell$ допускают $2^\ell$ различных разделений на два подмножества.

Будем говорить, что множество индикаторных функций $\theta(f(x,\alpha))$, $\alpha \in \Lambda$, имеет комбинаторную размерность $h$, если для этого множества:

1) *существует* $h$ векторов, разделяемых всеми $2^h$ способами;
2) *не существует* $h+1$ векторов, разделяемых всеми $2^{(h+1)}$ способами;
3) если для любого $\ell$ существуют векторы, которые могут быть разделены всеми $2^\ell$ способами, то комбинаторная размерность такого множества считается бесконечной.

Это определение комбинаторной размерности описывает разнообразие данного множества функций: было показано в [1], что если множество функций имеет конечную комбинаторную размерность $h$, то любые $\ell$ векторов могут быть разделены (используя функции множества) не более чем $O(\ell^h)$ различными способами. Если комбинаторная размерность $h$ бесконечна, то для любого $\ell$ существуют $\ell$ векторов $x_i$, которые могут быть разделены всеми $2^\ell$ возможными способами. Комбинаторная размерность — мера разнообразия множества функций. Такое определение может быть применено к *произвольному множеству функций без ограничений* и поэтому описывает внутреннее свойство разнообразия (иногда называемого емкостью или сложностью) данного множества функций.

Справедлива следующая теорема [1–3]:

*Теорема 1. Если комбинаторная размерность множества допустимых функций равна $h$, тогда любые $\ell$ векторов $x_i$ могут быть разделены*

[3] Отметим, что наилучшая функция в допустимом множестве может иметь большое значение функции потерь.

*на два подмножества не более чем*

$$N \leqslant \left(\frac{e\ell}{h!}\right)^h$$

*различными способами.*

### *3.2. Комбинаторная размерность множества линейных функций*

В теории обучения важную роль играет теорема, которая оценивает комбинаторную размерность множества линейных функций [1].

*Теорема 2. Комбинаторная размерность множества линейных функций*

$$f(x, w) = x^{\mathrm{T}} w, \quad w \in \mathrm{R}^n,$$

*где векторы $x$ принадлежат шару радиуса $R$ (т.е. $||x|| \leqslant R$) и векторы $w$ принадлежат шару радиуса $\Delta$ (т.е. $||w|| \leqslant \Delta$), ограничена выражением*

$$h \leqslant \min([\mathrm{R}^2\Delta^2], n) + 1. \tag{3.1}$$

В соответствии с этой теоремой в пространствах большой размерности $n$ комбинаторная размерность может быть меньше размерности пространства[4]. Этот факт играет важную роль в построении метода обучения, рассматриваемого в этой статье (см. оценку (2.2)).

### *3.3. Принцип структурной минимизации риска*

Для выбора наилучшего правила, используя оценку (2.2), в [1] был предложен так называемый принцип структурной минимизации риска. Пусть на допустимом множестве функций $S = \{\theta(f(x, \alpha)), \alpha \in \Lambda\}$ задана структура вложенных подмножеств функций $S_m = \{\theta(f(x, \alpha)), \alpha \in \Lambda^m\}$

$$S_1 \subset S_2 \subset \ldots \subset S_m \subset \ldots,$$

для которых соответствующие значения комбинаторной размерности $h^m$ упорядочены по возрастанию

$$h^1 < h^2 < \ldots < h^m < \ldots$$

Теперь можно минимизировать правую часть в (2.2) выбирая как подходящее подмножество $\Lambda^m$, комбинаторная размерность которого равна $h^m$, так и функцию $\theta(f(x, \alpha_\ell^m))$, которая минимизирует частоту ошибок обучения в $\Lambda^m$.

Верна следующая теорема [4–6].

[4] В [3] приведены оценки комбинаторной размерности множества функций, используемых для выбора правил в задаче распознавания цифр по данным из почтовой службы. Эти оценки показывают, что в пространствах очень большой размерности $d \sim 10^{10}$ оценки комбинаторной размерости составляли всего несколько сотен.

*Теорема 3. Если покрытие множества*

$$\overline{\bigcup_{k=1}^{\infty} S_k} = S$$

*компактно, тогда принцип структурной минимизации риска является сильно равномерно состоятельным.*

Этот факт и оценка (2.2) обосновывают многие алгоритмы обучения, включая алгоритмы обучения с использованием ядерных функций, рассматриваемых в этой статье.

### *3.4. Выбор допустимых множеств в методах обучения*

Отсутствие в теории обучения механизмов выбора допустимого подмножества функций привело к замене математически обоснованных методов выбора допустимых подмножеств функций различными эвристическими идеями, отражающими различные типы предположений. Далее опишем два из предположений.

1. Допущение о роли признаков в классификационных методах. В 1970-х гг. XX в. была представлена идея о том, что допустимое подмножество функций определяется разложением (например, линейным) по специальным функциям (имеющим смысл синтетических производных признаков и различным для различных задач) и что хорошее решение задачи обучения должно быть выражено в виде функции (скажем, линейной) от этих синтетических признаков.

Знание свойств таких признаков рассматривалось как априорное знание о решении задачи, которое косвенно должно быть включено в постановку задачи.

Согласно формальным рассуждениям чем больше признаков использовано, тем выше комбинаторная размерность соответствующего подмножества функций, тем больше данных необходимо для нахождения хорошей аппроксимации наилучшего правила для данного подмножества решений. С другой стороны, чем больше признаков использовано, тем выше шанс найти хорошую аппроксимирующую функцию в допустимом подмножестве. Поэтому когда используют конечное число наблюдений, сталкиваются с противоречивой ситуацией.

Чтобы разрешить это противоречие, рассматриваются методы сокращения числа признаков. Однако во всех случаях суть таких методов заключается в предположении, что существует “хорошее малопризнаковое пространство” и можно неформально найти его. Формулы (2.2) и (3.1) часто противоречат такому предположению, так как хорошее обобщение зависит не от размерности пространства, а от комбинаторной размерности допустимого подмножества (см. теорему 3) и существования в подмножестве хорошо аппроксимирующей функции.

Построение признаков не гарантирует решения какой-либо из задач выбора.

2. Допущение в обучении с использованием глубоких нейронных сетей. В 2005-х гг. возникла новая идея структурирования допустимого подмножества функций: так называемая идея архитектуры глубокого обучения, где используются нейронные сети со многими слоями. Построение глубоких нейронных сетей использует различные эвристики.

С формальной точки зрения, фиксируя архитектуру нейронной сети, выбирается подмножество функций, которые могут быть реализованы этой нейронной сетью. Считается, что (из-за "умной" конструкции глубокой сети) это подмножество содержит хорошо аппроксимирующую функцию и пытается найти ее, используя стохастический градиентный спуск (решение второй задачи выбора).

С теоретической точки зрения успех решения обеих задач путем выбора архитектуры глубокой сети (выбор допустимого подмножества функций и выбор желаемой функции в этом подмножестве) не обоснован, поскольку:

1. Глубокая нейронная сеть не обязательно содержит хорошую аппроксимацию желаемой функции. Более того, в соответствии с *теоремой о представителе* (описанной далее) для широких множеств функций гильбертова пространства решения, которые минимизируют функционал эмпирического риска на множестве функций с ограниченной нормой, определяются "мелкой" сетью (всего с одним скрытым слоем).
2. Стохастический градиентный спуск, используемый для нахождения решения, может не найти наилучшую функцию в подмножестве (этот метод находит локальный, но не глобальный минимум).

Следовательно, с теоретической точки зрения глубокие сети не могут гарантировать решения какой-либо из задач выбора, которые составляют полную задачу обучения.

## 4. Теория информации и машинное обучение

Перед тем как начать обсуждение методов полного решения задачи обучения, рассмотрим задачу обучения с точки зрения теории информации.

### *4.1. Основная модель Шеннона*

Развитие теории информации К. Шеннон начал со следующей основной идеи. Допустим, что цель — найти желаемый объект среди $N$ различных объектов, делая некоторое число запросов, на которые возвращаются ответы "да" ($y = 1$) или "нет" ($y = 0$), предоставляя запрашивающему один *бит* информации.

Теоретически, можно найти желаемый объект, делая $n$ запросов, где $n = \log_2 N$ (для простоты предполагаем, что $N = 2^n$). В самом деле, можно разделить множество из $N$ объектов на два подмножества и сделать запрос: к какому из двух подмножеств принадлежит объект: к первому (ответ 1) или ко второму (ответ 0). После получения ответа можно удалить подмножество, которое не содержит желаемого объекта, разделить оставшуюся часть на два подмножества и продолжать процедуру таким же образом, убирая половину

от оставшихся объектов после каждого ответа. Итак, после $n = \log_2 N$ запросов найдем желаемый объект. Поэтому чтобы найти один объект в множестве из $N$ объектов, необходимо получить не более чем $n = \log_2 N = \frac{\ln N}{\ln 2}$ бит информации.

### *4.2. Основная модель Шеннона в терминах теории обучения*

Повторим эту модель в терминах задачи распознавания образов. Допустим, что рассматриваемые объекты являются конечным множеством $y = \theta(f(x, a_t)),\ t = 1, \ldots, N$, индикаторных функций в $x \in \mathrm{R}^n$ и поставлена цель — найти в этом множестве неизвестную функцию классификации $y = \theta(f(x, \alpha_*))$, запрашивая значение функции для некоторого вектора $x$.

Допустим, что можно найти такой вектор $x_1 \in \mathrm{R}^n$, для которого половина функций из имеющегося множества принимает значение $\theta(f(x_1, a_{t_j})) = 1$, $j = 1, \ldots, N/2$, а другая половина принимает значение $\theta(f(x_1, a_{t_j})) = 0$, $j = N/2 + 1, \ldots, N$. Спрашивается, какая классификация вектора $x_1$ задает искомую функцию.

Метка $y_1$ вектора $x_1$ предоставляет первый элемент обучающей выборки $(x_1, y_1)$. Используя первый элемент обучающей выборки $(x_1, y_1)$, удаляем из данного множества половину функций, для которых $\theta(f(x_1, \alpha)) \neq y_1$ и продолжаем данную процедуру на оставшихся $N/2$ функциях.

После

$$n = \log_2 N = \frac{\ln N}{\ln 2} \tag{4.1}$$

повторений этого процесса найдем желаемую функцию. Итак, чтобы найти одну функцию в множестве из $N$ функций, используя запрос о специально выбранных векторах, необходимо $\ell = \log_2 N$ пар $(x_i, y_i)$.

### *4.3. Первая модификация модели обучения*

Чтобы найти требуемую функцию в рамках базовой модели, необходимо решить сложную задачу: на каждом шаге процедуры найти вектор $x_i$, который разбивает оставшийся набор данных на две равные части, по-разному классифицирующие данный вектор (считаем, что такое разделение существует). Чтобы упростить модель, рассмотрим ситуацию, когда векторы $x_i$ получены в результате независимых испытаний при фиксированной (но неизвестной) мере $P(x)$ и для любого $x_i$ можем узнать соответствующую метку $y_i$, формируя обучающую выборку $(x_i, y_i)$. Как и раньше, после каждого запроса функции, значения которых на векторе $x_i$ отличаются от сообщенного значения $y_i$, удаляются из рассмотрения (они могут составлять меньше половины от текущего набора функций).

Для описанной модели задача состоит в определении числа запросов, необходимых для нахождения функции, которая с вероятностью $1 - \eta$ $\varepsilon$-близка к искомой функции (напомним, что как $\varepsilon$-близость, так и вероятность $1 - \eta$ $\varepsilon$-близости определяются относительно меры $P(x)$). Решение этой проблемы

составляет специальный случай VC теории [3, 7]: число необходимых запросов (наименьший требуемый размер обучающей выборки) равно

$$\ell = \frac{\ln N - \ln \eta}{\varepsilon}. \tag{4.2}$$

Полученное выражение отличается от оценки (4.1) на постоянную величину: $\varepsilon^{-1}$ вместо $(\ln 2)^{-1}$. После этого числа запросов с вероятностью $1 - \eta$ любая функция в оставшемся наборе является $\varepsilon$-близкой к искомой функции. Данная оценка неулучшаема.

### *4.4. Вторая модификация модели обучения*

До сих пор рассматривалась ситуация, в которой набор из $N$ функций включал функцию, не делающую ошибок. Ослабим это допущение: любая из $N$ функций может ошибаться. Проблема заключается в поиске функции, которая обеспечивает наименьшую вероятность ошибки относительно меры $P(x)$. В этой ситуации нельзя воспользоваться методом выбора искомой функции, определенным в первой модели: удалять из рассмотрения функции, значения которых не совпадают с полученным в запросе значением. Воспользуемся обобщением данного метода, в котором среди $N$ заданных функций выбирается функция, имеющая наименьшее число несовпадений с полученными в запросах значениями, т.е. минимизирующая эмпирические потери на обучающей выборке

$$(x_1, y_1), \ldots, (x_\ell, y_\ell).$$

Можно показать, что для того чтобы с вероятностью $1 - \eta$ гарантировать выбор $\varepsilon$-близкой к лучшей из $N$ функций, необходимо использовать не более

$$\ell = \frac{\ln N - \ln \eta}{\varepsilon^2} \tag{4.3}$$

примеров [3, 6]. В этой модификации числитель $\ln N - \ln \eta$ остается прежним, но в отличие от (4.2) в знаменателе стоит $\varepsilon^2$.

Оценкам (4.2) и (4.3) можно дать следующую интерпретацию: в статистической постановке один элемент обучающих данных вносит в проблему выбора искомой функции меньше одного бита информации, в то время как в модели обучения Шеннона один обучающий пример вносит один бит информации. Оценки (4.2) и (4.3) неулучшаемы.

### *4.5. Третья модификация (VC модель)*

Теперь рассмотрим набор функций $\theta(f(x, \alpha))$, $\alpha \in \Lambda$, с бесконечным числом элементов. Вообще говоря, в этой ситуации нельзя гарантировать, что найдется хорошая аппроксимация лучшей функции даже при большом числе обучающих примеров.

Тем не менее если этот набор функций имеет конечную комбинаторную размерность $h$, то в соответствии с VC теорией с вероятностью $1 - \eta$ $\varepsilon$-близкое

решение может быть найдено, используя не более

$$\ell \sim O\left(\frac{h - \ln \eta}{\varepsilon}\right)$$

примеров, если аппроксимирующая функция не делает ошибок. Если ошибки возникают, то минимальное число примеров ограничивается величиной

$$\ell \sim O\left(\frac{h - \ln \eta}{\varepsilon^2}\right).$$

Отметим, что эти оценки имеют вид границ (4.2), (4.3), где величина комбинаторной размерности замещает логарифм числа функций в наборе. Эти оценки неулучшаемы [3, 7].

### *4.6. Три категории больших чисел*

Следуя А.Н. Колмогорову, выделим три категории целых чисел.

1. Обычные числа. Назовем *обычными* целые числа $n$, если они соответствуют ряду понятий из обычной жизни. Например, можно определить как обычные (маленькие) целые числа от 1 до миллиона (или миллиарда).
2. Большие числа. Назовем целые числа $\mathcal{N}$ *большими* числами, если $10^9 < \mathcal{N} << 2^n$ ($n$ принадлежит категории обычных чисел).
3. Гигантские числа. Назовем целые числа $\mathcal{N}$ *гигантскими* числами, если больше больших чисел $\mathcal{N} >> 2^n$ (скажем, имеют порядок $\mathcal{N} = O(2^{2^n})$).

Проблему обучения можно рассмотреть как проблему выбора требуемой функции из множества $N$ функций, когда один пример вносит до одного бита информации (число кандидатов уменьшается максимум вдвое). Следовательно, число функций, из которых производится выбор, должно быть по крайней мере больши́м (но не гигантским).

В рассмотренных моделях, используя обычное число $\ell$ обучающих примеров, нельзя выбрать требуемую функцию из набора с гигантским числом элементов (либо из набора функций с большой или бесконечной величиной комбинаторной размерности). Чтобы это сделать, необходимо использовать механизм, в котором один «пример» может внести гораздо больше, чем один бит информации.

Такой механизм задается специальным типом сходимости в гильбертовом пространстве, так называемой *слабой сходимостью.*

## 5. Два вида сходимости в гильбертовом пространстве

В настоящей статье в качестве множества аппроксимирующих функций используются действительные функции $f(x, \alpha)$, $\alpha \in \Lambda$, которые принадлежат гильбертовому пространству. Гильбертово пространство $H$ определяется операцией, называемой *скалярным произведением*, которое сопоставляет для любых двух элементов $\mathbf{u}, \mathbf{v}$ пространства значение $(\mathbf{u}, \mathbf{v})$ *скалярного произведения.* Скалярное произведение определяется как величина, удовлетворяющая следующим трем свойствам:

1. Симметричность. Скалярное произведение двух элементов — симметричная функция

$$(\mathbf{u}, \mathbf{v}) = (\mathbf{v}, \mathbf{u}).$$

2. Линейность. Скалярное произведение линейно по первому аргументу

$$((a_1\mathbf{u_1} + a_2\mathbf{u_2}), \mathbf{v}) = a_1(\mathbf{u_1}, \mathbf{v}) + a_2(\mathbf{u_2}, \mathbf{v}).$$

3. Положительная определенность. Скалярное произведение положительно определено

$$(\mathbf{u}, \mathbf{u}) \geqslant 0 \quad \forall \mathbf{u},$$

$(\mathbf{u}, \mathbf{u}) = 0$, если и только если $\mathbf{u} = 0$.

При использовании скалярного произведения $(\mathbf{u}, \mathbf{v})$ расстояние определяется как

$$\rho(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}|| = \sqrt{((\mathbf{u} - \mathbf{v}), (\mathbf{u} - \mathbf{v}))}.$$

Гильбертово пространство определяется как полное метрическое пространство относительно этой метрики.

### *5.1. Сильная и слабая сходимости в гильбертовом пространстве*

Две числовые характеристики связи двух функций в гильбертовом пространстве — расстояние и скалярное произведение — позволяют ввести два разных вида сходимости: сильную и слабую.

1. Сильная сходимость функций $f_\ell(x)$ к $f_0(x)$ требует, чтобы

$$\lim_{\ell \to \infty} ||f_\ell(x) - f_0(x)|| = 0$$

(сходимость в пространстве функций).

2. Слабая сходимость функций$f_\ell(x)$ к $f_0(x)$ требует, чтобы

$$\lim_{\ell \to \infty} ((f_\ell(x), \phi(x)) - (f_0(x), \phi(x)) = 0 \quad \forall \phi \in L_2$$

(сходимость в пространстве функционалов). Заметим, что равенство должно выполняться для всех функций $\phi(x) \in L_2$.

Используя неравенство Коши—Шварца, легко показать, что сильная сходимость влечет за собой слабую[5]. В самом деле,

$$((f_\ell(x) - f_0(x)), \phi(x))^2 \leqslant ||f_\ell(x) - f_0(x)||^2 ||\phi(x)||^2,$$

т.е. сильная сходимость в правой части неравенства влечет слабую в левой части.

[5] Вообще говоря, слабая сходимость не влечет за собой сильную. Однако для функций, рассматриваемых в этой статье (принадлежащие к гильбертовому пространству с воспроизводящим ядром ограниченной нормы), слабая сходимость влечет за собой сильную.

Существующая статистическая теория обучения посвящена решению *Второй задачи выбора*. Соответствующие механизмы обучения используют сильную сходимость аппроксимирующих функций $f_\ell(x)$, построенных на основании данных

$$(x_i, y_1), \ldots, (x_\ell, y_\ell)$$

и искомой функции $f_0(x)$.

Решение *Первой задачи обучения* основывается на механизмах слабой сходимости. Допустим, дано множество из $m$ функций $\phi_1(x), \ldots, \phi_m(x)$ ($\phi_i \in L_2$) (называем их *предикатами*). Идея решения первой задачи выбора состоит в выборе подмножества функций $f(x, \alpha), \alpha \in \Lambda$, для которых выполнено равенство

$$(f(x, \alpha), \phi_s(x)) = u_s, \quad s = 1, \ldots, m, \tag{5.1}$$

где

$$u_s = (f_0(x), \phi_s(x)), \quad s = 1, \ldots, m. \tag{5.2}$$

Назовем подмножество функций $\{f(x)\}$ из $L_2$, которое удовлетворяет $m$ интегральным соотношениям (5.1), допустимым подмножеством[6]. Заметим, что любое допустимое множество функций включает искомую функцию $f_0(x)$.

Допустим, что дано наряду с предикатами $\phi_s(x)$, $s = 1, \ldots, m$, значение $u_s$, $s = 1, \ldots, m$, определенное в (5.2). Тогда задача выбора допустимого подмножества функций может быть сформулирована следующим образом:

для заданных пар

$$(\phi_1(x), u_1), \ldots, (\phi_m(x), u_m)$$

найти в множестве функций $f(x, \alpha)$, $\alpha \in \Lambda$, подмножество, удовлетворяющее (5.1).

Далее покажем, что для задачи распознавания образов можно оценить значение $u_s$ и найти подмножество функций $f(x, \alpha)$, $\alpha \in \Lambda$, удовлетворяющих (5.1).

Таким образом, идея выбора искомой функции в подмножестве допустимых функций основана на механизмах сильной сходимости, в то время как идея выбора допустимого подмножества функций из широкого множества функций основывается на механизмах слабой сходимости.

Важно отметить, что так как в гильбертовом пространстве существуют только два вида сходимости, то существуют только два основных механизма обучения.

[6] Так как слабая сходимость определена всеми функциями $\phi(x) \in L_2$, то любое подмножество $\phi_i(x), i = 1, \ldots, m$, определяет допустимое подмножество.

## 6. Инструменталистские и реалистические модели обучения

### *6.1. Феноменологическая модель обучения*

Рассмотрим следующую схему обучения: допустим, что неизвестный генератор $G$ случайно и независимо генерирует векторы $x \in X$. Предположим, что некоторый объект $O$ возвращает на каждом входе $x_i$ бинарный выход $y_i \subset \{1, 0\}$, используя неизвестную условную функцию распределения $P(y|x)$. Это означает, что объект $O$ использует функцию $0 \leqslant P(y = 1|x) \leqslant 1$ и механизм, который определяет классификацию $y_i$ вектора $x_i$ подбрасыванием монетки: с вероятностью $P(y = 1|x_i)$ возвращается значение $y_i = 1$ и с вероятностью $1 - P(y = 1|x_i)$ возвращается $y_i = 0$.

Пары $(x_*, y_*)$ (классификация $y_*$ входного вектора $x_*$) наблюдаются обучающейся машиной $L$, которая способна выбрать функцию из заданного множества функций $f(x, \alpha)$, $\alpha \in \Lambda$. Требуется по наблюдениям $\ell$ пар

$$(x_1, y_1), \ldots, (x_\ell, y_\ell)$$

найти индикаторную функцию $y = \theta(f(x, \alpha_\ell))$, которая минимизирует математическое ожидание числа несовпадающих классификаций, задаваемых функцией $y = \theta(f(x, \alpha_\ell))$ и правилом $P(y|x)$.

Существуют две различные философские позиции для постановки задачи обучения, основанные на описанной феноменологической модели:

1. Позиция *философского инструментализма*, которая утверждает, что целью исследования является *имитация* работы объекта $O$ (чтобы предсказать выход $y$ объекта $O$ на любом векторе $x$, сгенерированном $G$), т.е. найти наилучшую функцию, прогнозирующую $y = \theta(f(x, \alpha_\ell))$ *в допустимом подмножестве* $f(x, \alpha)$, $\alpha \in \Lambda$;
2. Позиция *философского реализма*, которая утверждает, что целью исследования является оценка условной функции распределения $P(y = 1|x)$, которая использует объект $O$. Используя аппроксимацию $P(y = 1|x)$, может быть определено классификационное правило.

Короче, философский инструментализм утверждает, что целью науки является предсказание результатов во вселенной, в то время как философский реализм утверждает, что целью науки является понимание внутренних механизмов вселенной.

1. Модель, основанная на подходе инструментализма. С позиции философского инструментализма математическая модель задачи двухклассовой классификации (обучения) требует нахождения функции $\theta(f(x, \alpha_0))$, минимизирующей функционал

$$R_{inst}(\alpha) = \int (y - \theta(f(x, \alpha)))^2 dP(x, y)$$

на множестве допустимых индикаторных функций $\theta(f(x, \alpha)), \alpha \in \Lambda$.

2. Модель, основанная на подходе реализма. С позиции философского реализма математическая модель задачи обучения требует использования обучающих данных для того, чтобы оценить на множестве допустимых действительных функций $0 \leqslant f(x, \alpha) \leqslant 1$, $\alpha \in \Lambda$, условную функцию распределения

$f(x, \alpha_0) = P(y = 1|x)$. Используя соответствующую оценку $P_\ell(y = 1|x)$, можно получить классификационное правило

$$r(\alpha_\ell) = \theta(P_\ell(y = 1|x) - 0{,}5). \tag{6.1}$$

В [3], обосновывая постановку задачи обучения, которая отражает позицию инструментализма, автором сформулирован императив: "Решая интересующую вас задачу, не решайте более общую задачу как промежуточную. Решайте интересующую вас задачу, а не более сложную". В настоящей статье автор отказывается от этого императива и будет рассматривать постановку задачи обучения с реалистической точки зрения оценивания условной функции вероятности и получения при использовании этой функции требуемого классификационного правила (6.1).

Такой подход позволит включить в обучение оба механизма сходимости, существующие в гильбертовом пространстве: механизм сильной и слабой сходимости (см. соответствующий раздел далее).

В статье в качестве основной задачи обучения рассматривается задача двухклассовой классификации путем оценивания условной функции вероятности $P(y = 1|x)$ и получения правила классификации (6.1). Обобщение на $p$-классовую классификацию $y \in \{0, \ldots, (p-1)\}$ основано на $p$-оценках условных функций вероятности $P(y = t|x)$, $t = 0, \ldots, (p-1)$. Используя эти функции, можно получить правило классификации

$$y = \text{argmax}(P(y = 0|x), \ldots, P(y = (p-1)|x))$$

(специальные детали для мультиклассовой классификации даны далее).

## 7. Выбор допустимого подмножества функций

### *7.1. Преимущество реалистического подхода в задачах обучения*

Далее для решения двухклассовой задачи классификации оцениваем условную функцию вероятности $P(y|x)$, где $y \subset \{0, 1\}$. Для этого необходимо найти функцию $f(x, \alpha_0) = P(y = 1|x)$.

Слабая сходимость последовательности условных функций распределения

$$P_v(y = 1|x) \longrightarrow_{v\to\infty} P(y = 1|x)$$

означает равенство

$$\lim_{v\to\infty} \int \phi(x) P_v(y = 1|x) dP(x) = \int \phi(x) P(y = 1|x) dP(x) \quad \forall \phi(x) \in L_2$$

(для всех функций $\phi(x)$ из $L_2$). Для условных функций вероятности это равенство эквивалентно равенству

$$\lim_{\ell\to\infty} \int \phi(x) P_\ell(y = 1|x) dP(x) = \int \phi(x) dP(y = 1, x) \quad \forall \phi(x) \in L_2. \tag{7.1}$$

Ослабим требования к слабой сходимости. Пусть решение (7.1) принадлежит множеству функций $f(x,\alpha)$, $\alpha \in \Lambda$, которые удовлетворяют конечному числу равенств $m$ (как в (7.1))

$$\int \phi_s(x)P_\ell(y=1|x)dP(x) = \int \phi_s(x)dP(y=1,x), \quad s=1,\ldots,m, \tag{7.2}$$

заданных с помощью предикат $\phi_s(x)$, $s=1,\ldots,m$. Заметим, что так как бесконечное число предикатов заменено конечным, то решение (7.2) не единственное, но принадлежит подмножеству допустимых функций, которое включает в себя искомую функцию.

Допустим, что имеем достаточно большую обучающую выборку

$$(x_1,y_1),\ldots,(x_\ell,y_\ell), \quad x \in X,\ y \subset \{1,0\}.$$

Тогда согласно закону больших чисел справедливо следующее приближение равенства (7.2) для оценки условных функций вероятности $P_\ell(y=1|x)$:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\phi_s(x_i)P_\ell(y=1|x_i) \approx \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\phi_s(x_i), \quad s=1,\ldots,m. \tag{7.3}$$

Определения:

1. Говорим, что функция $P_*(y=1|x)$ принадлежит подмножеству *допустимых* функций, если для данных предикатов $\phi_s(x)$, $s=1,\ldots,m$, то она удовлетворяет (7.3).
2. Называем равенства (7.3) *статистическими инвариантами.* Правую часть равенства (7.3) называем *эмпирическими данными* о предикате $\phi_s$ и левую часть (7.3) — *теоретическими данными* об этом предикате.
3. Называем метод оценки правила классификации (6.1), использующий условную функцию вероятности, которая удовлетворяет (7.3), *обучением, использующим статистические инварианты* (или LUSI).

Равенство (7.3) требует, чтобы *данные, полученные для теоретической модели условной вероятности, подтверждались эмпирическими данными* (полученными при использовании обучающей выборки, сгенерированной, на основывании истинной условной функции вероятности).

Для полного решения задачи обучения (которое включает в себя две задачи выбора: выбор допустимого подмножества функций и выбор из этого подмножества желаемой функции) требуется три шага:

1. Выбрать подмножество допустимых функций, используя наблюдения и предикаты.
2. Найти требуемую условную функцию вероятности в допустимом подмножестве.
3. Построить правило (6.1), используя оцененную условную функцию вероятности.

Замечания о механизме слабой сходимости.

*Замечание 1.* Так как слабая сходимости требует существования равенства (7.1) для всех функций $\phi(x) \in L_2$, можно выбрать *произвольные* $m$ функций $\phi_s(x)$ из $L_2$ в качестве предикатов.

*Замечание 2.* Для любого заданного подмножества функций число функций, удовлетворяющих инвариантам, убывает с ростом числа $m$ инвариантов.

*Замечание 3.* Хорошая аппроксимация функции принадлежит множеству функций, удовлетворяющих инвариантам (так как эмпирические данные получены, основываясь на данных, сгенерированных с помощью истинной условной функции вероятности).

*Замечание 4.* Как упоминалось ранее, в классической модели любая новая пара в обучающей выборке несет в себе не более одного бита информации. В модели LUSI (которая использует наряду с обучающей выборкой данное множество предикатов) с добавлением одного предиката можно обеспечить более одного бита информации (так как возможно, что менее половины функций из подмножества удовлетворяют инварианту с этим предикатом).

*Замечание 5.* Выбор "хороших" новых предикатов (которые отсекают большое количество функций, нарушающих инвариант соответствующего предиката) — это *творческая* часть задачи, которая частично обсуждается в разделе 8.

### *7.2. Пример логики вывода, основанной на слабой сходимости*

Существует так называемый "тест на утку" (Duck test), который иллюстрирует логику вывода, основанную на слабой сходимости. "Тест на утку", описанный в английской пословице, гласит: *"Если что-то выглядит как утка, плавает как утка, крякает как утка, значит, это, возможно, утка".*

В рамках машинного обучения "тест на утку" может быть выражен так:

Во-первых, выберите подмножество допустимых правил, каждое из которых удовлетворяет свойству "утка", для которого применимы два типа предикатов:

1. *Предикаты общего типа* — "выглядит как утка".
2. *Предикаты специального типа* — "плавает как утка", "крякает как утка".

Во-вторых, выберите из допустимого подмножества искомое правило, используя стандартную процедуру обучения на основе данных[7].

Рассматриваем предикат "выглядит как класс $y = 1$" (класс "утки" ) как *предикат общего типа*, потому что он выражает общие (математические) свойства элементов обучающей выборки, которые принадлежат классу $y = 1$, и не использует специальные знания о данной задаче классификации.

Называем предикаты "плавает как утка" и "крякает как утка" *предикатами специального типа* потому, что, представляя их, знаем, как утка плавает

[7] Смысл "теста на утку" в том, что существуют такие хорошие предикаты, что любое правило из допустимого множества будет достаточно хорошим. Пословица гласит, что достаточно трех предикатов. В предложенном методе будем выбирать требуемую аппроксимацию из допустимого множества не с помощью выбора, а используя второй метод выбора, предоставленный классической теорией статистического обучения.

и крякает. Так как любая функция может быть использована в качестве предиката, то можно рассматривать предикаты вида: “прыгает как утка” или “играет в шахматы как утка”. Однако такие предикаты не слишком полезны в задаче классификации птиц.

Выбор предиката общего типа является в большей степени математической частью задачи. Выбор предиката специального типа в большой степени — *творческая* часть задачи. Выбор требует знаний об определенных деталях решаемой задачи. Выбор специальных предикатов отражает влияние учителя на процесс обучения. Это предмет взаимодействия *Учитель — Студент*.

### *7.3. Примеры предикатов общего типа*

Можно ввести различные концепции предикатов общего типа, которые отражают идею “выглядит как утка”. В этом разделе рассмотрим три математические концепции, отражающие существующий статистический подход, называемый *методом моментов*. Пусть

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \quad x \in X,\ y \subset \{0, 1\},$$

— независимые одинаково распределенные элементы обучающей выборки, сгенерированной в соответствии с $P(x, y) = P(y|x)P(x)$.

1. Предикат для статистического инварианта момента нулевого порядка. Рассмотрим предикат $\phi(x) = 1$. Используя эту функцию в (7.3), получаем инвариант

$$\sum_{i=1}^{\ell} P_\ell(y = 1|x_i) = \sum_{i=1}^{\ell} y_i, \tag{7.4}$$

   который определяет баланс классов в обучающей выборке: число представителей класса $y = 1$, посчитанное на основе оценки условной функции вероятности, должно быть равно наблюдаемому числу представителей класса $y = 1$ в обучающей выборке.
2. Предикаты для статистического инварианта момента первого порядка. Рассмотрим вектор-функцию предикатов $\phi(\mathbf{x}) = \mathbf{x}$, где $\mathbf{x_i} = (x_i^1, \ldots, x_i^n)^{\mathrm{T}}$ — вектор размерности $n$, $\mathbf{x} \in X$. Используя эту функцию в (7.3), получаем (покоординатно) $n$ инвариантов

$$\sum_{i=1}^{\ell} \mathbf{x_i} P(y = 1|\mathbf{x_i}) = \sum_{i=1}^{\ell} y_i \mathbf{x_i}, \quad \mathbf{x_i} = (x_i^1, \ldots, x_i^n)^{\mathrm{T}}. \tag{7.5}$$

   Равенство (7.5) требует, чтобы центр масс векторов $x$ из обучающей выборки, принадлежащих к первому классу, вычисленный на основе оценки условной функции вероятности, и векторов $x_1, \ldots, x_\ell$ совпадал с центром масс, полученным непосредственно из элементов обучающей выборки первого класса.

3. Предикаты для статистического инварианта момента второго порядка. Чтобы ввести $0{,}5n(n+1)$ инвариантов относительного статистического момента второго порядка, рассмотрим предикаты, определяемые ковариационной матрицей $n \times n$

$$\phi(\mathbf{x_i}) = \mathbf{x_i}\mathbf{x_i^T}.$$

Соответствующие матричные инварианты имеют вид

$$\sum_{i=1}^{\ell} \mathbf{x_i}\mathbf{x_i}^{\mathrm{T}} P_\ell(y=1|\mathbf{x_i}) = \sum_{i=1}^{\ell} y_i \mathbf{x_i}\mathbf{x_i}^{\mathrm{T}}. \tag{7.6}$$

Матричное равенство (7.6) рассматривается поэлементно. Каждый элемент $a_{k,m}$ матрицы $n \times n$ в левой части, полученный при использовании оценки $P_\ell(y=1|x)$, должен быть равен соответствующему элементу $a^*_{km}$ матрицы в правой части равенства (7.3), посчитанному при использовании векторов первого класса из обучающей выборки. Из (7.6) может быть получено $0{,}5n(n+1)$ различных инвариантов[8], которые определяют инварианты для всех попарных координат входного вектора $x$.

Инварианты для нулевого (7.4), первого (7.5) и второго (7.6) порядков моментов являются характеристиками множества допустимых условных функций вероятности[9], которые могут быть использованы для решения определенной задачи распознавания образов. Они играют важную роль в приложениях.

Предикат подмножества пространства. Пусть $I_S(x)$ является индикатором, который определяет принадлежность вектора $x$ подмножеству $S$ в каком-то подпространстве $X^*$ пространства $X$ (например, $S(x) = \{||x-a|| \leqslant b\}$, где векторы $x, a \in X$ и $x^* = (x^1, \ldots, x^k, 0, \ldots, 0)$ определены первыми $k$ координатами вектора $x$). Рассмотрим предикат, определенный индикаторной функцией

$$I_S(x) = \begin{cases} 1, & \text{если } x \in S(x), \\ 0 & \text{иначе,} \end{cases}$$

и соответствующий инвариант

$$\sum_{i=1}^{\ell} I_S(x_i) P_\ell(y=1|\mathbf{x_i}) = \sum_{i=1}^{\ell} y_i I_S(x_i). \tag{7.7}$$

Инвариант (7.7) требует, чтобы число векторов обучающей выборки первого класса, которые принадлежат $S(x)$, т.е. число, посчитанное при использовании условной функции вероятности, совпадало с числом векторов первого класса в $S(x)$, непосредственно посчитанным по обучающей выборке.

Выбирая различные функции $S(x)$, определяем различные инварианты.

[8] Так как матрица симметрична, она содержит $0{,}5n(n+1)$ различных элементов.

[9] Напомним, что метод моментов является важным методом в задаче оценки плотности.

Предикаты специального типа. Предикаты общего типа являются чисто математическими конструкциями, они не полагаются на знания об интересующей задаче.

Предикаты специального типа (такие как "плавает как утка" и "крякает как утка" в пословице, но не "прыгает как утка" или "играет в шахматы как утка", которые допустимы) основываются на знаниях о задаче.

Введение предикатов специального типа составляет *творческую* часть решения задачи обучения. Они основаны на добавочных (специальных) знаниях о задаче. Эти предикаты формируют основу взаимодействия учителя и ученика в процессе обучения.

Продемонстрируем предикаты специального типа для задачи распознавания цифр. При построении функции условной вероятности для распознавания цифры 3 можно использовать наблюдение о том, что цифра 3 имеет свойство "горизонтальной симметрии" (верхняя часть соответствующей цифры остается ее нижней частью) и не имеет свойства вертикальной симметрии. Можно ввести число, которое измеряет коэффициент горизонтальной симметрии $\phi(x)$, и использовать его в качестве предиката для построения инвариантов специального типа.

Опишем один из возможных способов построения такой меры. Пусть $x^*$ является размытой версией изображения $x$, определенного на $(k \times p)$-пространстве пикселей $X$. Каждое (размытое) изображение цифры 3 может быть описано вектором $N = ((c_{1,1}, \ldots, c_{1,p}), \ldots, (c_{k,1}, \ldots, c_{k,p}))$ размерности $kp$ (конкатенация первой строки пикселей со второй и т.д.). Для того чтобы ввести коэффициент симметрии, наряду с вектором $N$, рассмотрим вектор $N^* = ((c_{k_1}, \ldots, c_{k,p}), \ldots, (c_{1,1}, \ldots, c_{1,p}))$ размерности $k \times p$ (конкатенация последней строки пикселей с предыдущей и т.д.).

В качестве коэффициента симметрии цифры $x$ рассмотрим значение скалярного произведения векторов $N$ и $N^*$ или $\phi_{sim}(x) = (N, N^*)$.

Аналогичным образом может быть введена мера вертикальной симметрии.

В качестве предиката можно использовать меру затененности некоторой части пространства пикселей (используя среднее значение пикселей в данной части пространства).

## 8. Оценка функции условной вероятности

Для построения инвариантов необходимо найти методы оценивания требуемой функции условной вероятности. Этот раздел описывает идею таких методов.

Согласно аксиоматике теории случайности фундаментальным понятием как *теории вероятностей*, так и *статистики* является так называемая функция распределения $P(x)$ случайной величины $x$.

Задачи *теории вероятностей* могут быть описаны следующим образом: даны $P(x)$ описание эксперимента со случайной величиной $x$, нужно найти функцию распределения результатов эксперимента.

Основная задача статистики может быть описана следующим образом: дано наблюдение результатов $\ell$ случайных (независимых, одинаково распреде-

ленных) испытаний $x$ с $P(x)$, найти функцию распределения $P_\ell(x)$ (или некоторые ее характеристики, как, скажем, первый момент).

Этот раздел посвящен формализации задачи оценки условной функции распределения $P(y=1|x)$ на основании наблюдений

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \quad x \in \mathrm{R},\ y \in \{0, 1, \},$$

сгенерированных случайно и независимо в соответствии с $P(y,x) = P(y|x)P(x)$: рассматриваем задачу оценки условной функции распределения как задачу решения интегрального уравнения Фредгольма[10]

$$\int \theta(x - x')P(y=1|x')dP(x') = P(y=1, x), \tag{8.1}$$

определенного с помощью ядра $\theta(x-x')$. Необходимо решить интегральное уравнение (8.1) (относительно неизвестной функции $P(y=1|x)$), если функции распределения $P(x)$ и $P(y=1,x)$ неизвестны, но дана выборка $(x_i, y_i)$.

Заметим, что решение интегрального уравнения Фредгольма составляет *плохо обусловленную* задачу. Более того, в рассматриваемом случае — плохо обусловленная задача специального типа, в которой не только правая часть равенства $P(y=1,x)$ должна быть определена приближенно, но и оператор интегрального уравнения (функция $P(x')$ в левой части (8.1)) должен быть оценен по данным, поэтому он также определен приближенно.

## 9. Оценка функции распределения

Оценка функции распределения является основной задачей статистики. Для решения любой другой задачи статистики можно применить методы теории вероятности, используя оцененную функцию распределения вместо истинной.

### *9.1. Эмпирическая функция распределения и теорема Гливенко–Кантелли*

Для оценки функции распределения $P(x)$ по наблюдениям $x_1, \ldots, x_\ell$ необходимо задать форму аппроксимирующей функции. Статистика предлагает следующую форму для оценки функции распределения

$$P_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - x_i). \tag{9.1}$$

[10] Уравнение (8.1) эквивалентно

$$\int_{-\infty}^{x} P(y=1|t)dP(t) = P(y=1, x),$$

и беря производную по $x$, можно получить условную функцию распределения.

Эта кусочно-постоянная монотонная функция называется эмпирической функцией распределения. Введение этой функции для использования вместо реальной функции распределения является главным индуктивным шагом в статистике[11].

*Теорема 4. Эмпирическая функция распределения $P_\ell(z)$ сходится к реальной функции распределения равномерно с вероятностью единица, т.е.*

$$\sup_z |P_\ell(z) - P(z)| \longrightarrow^P_{\ell\to\infty} = 0.$$

В 1933 г. А.Н. Колмогоров нашел асимптотически точную скорость сходимости для случая $z \in \mathrm{R}^1$. Он показал, что справедливо соотношение

$$P\left[\lim_\ell \sqrt{\ell} \sup_x |P_\ell(z) - P(z)| > \varepsilon\right] = 2\sum_{k=1}^{\ell} (-1)^{k-1} \exp -2\varepsilon^2 k^2 \quad \forall\varepsilon.$$

Позже (в 1956–1990 гг.), Дворецкий–Кифер–Вольфовиц–Массарт нашли точную неасимптотическую оценку, в которой правая часть в точности совпадает с первым членом колмогоровского соотношения

$$P\left[\sup_x |P_\ell(x) - P(x)| > \varepsilon\right] \leqslant 2\exp -2\varepsilon^2\ell \quad \forall\varepsilon.$$

Обобщение оценки на $n$-мерный случай $x = (x^1, \ldots, x^n) \in \mathrm{R}^n$, где

$$P_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\prod_{k=1}^{n}\theta(x^k - x_i^k),$$

было получено в 1970 г. с использованием VC теории:

$$P\left[\sup_z |P_\ell(z) - P(z)| > \varepsilon\right] \leqslant 2\exp\left\{-\left(\varepsilon^2 - \frac{n\ln\ell}{\ell}\right)\ell\right\} \quad \forall\varepsilon.$$

Во всех проблемах статистического вывода будем использовать одинаковый индуктивный прием: будем замещать функцию распределения ее эмпирической оценкой (9.1).

Конструктивная постановка проблемы оценивания. Чтобы оценить условную функцию распределения, заменим в (8.1) неизвестные функции распределения $P(x)$ и $P(y = 1, x)$[12] выражениями:

$$P_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - x_i), \quad P_\ell(y = 1, x) = \frac{1}{\ell}\sum_{j=1}^{\ell} y_j\theta(x - x_j).$$

[11] Поскольку данные и функции являются объектами различной природы, то для получения функции из данных необходима индуктивная идея. Аппроксимация (9.1) является главным индуктивным шагом в статистике.

[12] Идея такой замены отражает стандартный индуктивный шаг в статистическом выводе.

Получим, что

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x-x_i)P(y=1|x_i)\approx\frac{1}{\ell}\sum_{j=1}^{\ell}y_j\theta(x-x_i). \tag{9.2}$$

Цель заключается в решении уравнения (9.2) на множестве функций $P(y=1|x)\in\{f(x,\alpha),\ \alpha\in\Lambda\}$, которые удовлетворяют инвариантам

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\phi_s(x_i)P(y=1|x_i)\approx\frac{1}{\ell}\sum_{i=1}^{\ell}y_i\phi_s(x_i),\quad s=1,\ldots,m,$$

где $\phi_s(x),\ s=1,\ldots,m,$ – выбранные предикаты.

Решение этой проблемы составляет главную часть полной статистической теории обучения.

## 10. Оценка функции условной вероятности является плохо обусловленной задачей

Проблема решения операторного уравнения

$$Af=F, \tag{10.1}$$

которое отражает функции $\{f\}$ в $\{F\}$, принадлежит к так называемым *плохо обусловленным задачам*, т.е. если решение уравнения существует, единственно, но неустойчиво, то имеются маленькие возмущения правой части уравнения (10.1) ($||F_\delta-F||\leqslant\delta$), которые могут привести к большим возмущениям решения ($||f_\delta-f_\delta||\geqslant C$). Другими словами, обратное отображение (которое отображает $\{F\}$ в $\{f\}$ ) не непрерывно.

В рассматриваемом случае имеется плохо обусловленная задача (8.1), где не только правая часть уравнения $P(y=1,x)$ оценивается по данным и, значит, задана приближенно, но также и оператор интегрального уравнения задан приближенно (функция $P(x')$ в левой части (8.1) и также оценивается по данным.

При решении подобных задач используем так называемые методы регуляризации, разработанные для решения плохо обусловленных задач [4].

### *10.1. Методы решения плохо обусловленных задач*

Основная идея решения плохо обусловленных задач связана с леммой об обратном операторе. Допустим, что существует единственное решение $f_0\in$ $\in\{f\}$ операторного уравнения

$$Af=F,\quad f\in\{f\},$$

где $A$ – непрерывный линейный оператор, отображающий функции $f\in\{f\}$ в функции $F\in\{F\}$. Пусть $A^{-1}$ является обратным оператором к $A$, задающим отображение из $\{F\}$ в $\{f\}$

$$A^{-1}F=f.$$

Справедлива следующая лемма.

*Лемма* (об обратном операторе). *Оператор $A^{-1}$, обратный к непрерывному оператору $A$, определенному на множестве $\{f\}$, непрерывен, если это множество компактно.*

Эта лемма означает, что для того чтобы найти устойчивое решение интегрального уравнения (заданного непрерывным оператором $A$), необходимо выбрать некоторое компактное подмножество функций $\{f^*\}$, которое включает искомое решение, и затем решить уравнение в этом подмножестве.

При решении плохо обусловленных задач обычно используется следующая форма задания компактных множеств:

$$f \in \{W(f) \leqslant C\}, \tag{10.2}$$

где $W(f)$ — положительная выпуклая функция (выпуклость требуется из вычислительных соображений), $C$ константа.

Пусть решение принадлежит множеству (10.2) при некоторой фиксированной константе $C$. В этом случае для решения операторного уравнения (10.1) необходимо минимизировать функционал

$$R(f) = \rho^2(Af, F)$$

при ограничениях (10.2). Подобная форма решения плохо обусловленных задач называется *квазирешением*. Для линейного оператора $A$ и выпуклого функционала $W(f)$ квазирешение — единственно.

С помощью множителей Лагранжа проблему условной минимизации можно записать как проблему безусловной минимизации функционала

$$R(f) = \rho(Af, F) + \gamma W(f). \tag{10.3}$$

Такая форма решения называется регуляризацией. Параметр $\gamma > 0$ в регуляризованной проблеме (10.3) определяется константой $C$ в (10.2).

Главные теоремы из области плохо обусловленных задач утверждают, что если решение принадлежит компакту (10.2) с некоторой константой $C$ (или $\gamma$ в эквивалентной форме (10.3)), то существует такой закон выбора параметра $C_\delta$ (или $\gamma_\delta$) в зависимости от точности представления правой части операторного уравнения (10.1), что соответствующие решения уравнения (10.1) сходятся к искомому решению при $\delta \to 0$ [4]. Кроме того, показано, что при некоторых общих условиях описанные решения плохо обусловленных задач с приближенно заданным оператором $A_\ell$ и приближенно заданной правой частью также сходятся к искомому решению [8].

### *10.2. Оценка функции условного распределения*

Для решения плохо обусловленной задачи оценки условной функции распределения при использовании регуляризованного функционала (10.3) (или метода квазирешений) необходимо определить три элемента:

1. Квадрат расстояния $\rho^2(Af, F)$. В настоящей статье используем расстояние $L_2(\mu)$

$$\rho(Af, F) = \int (Af(x) - F(x))^2 d\mu(x),$$

где $\mu(x)$ — некоторая заданная вероятностная мера (например $\mu(x) = P(x)$).
2. Множество функций $\{f\}$, в котором ищется решение, автор использует функции, принадлежащие гильбертовому пространству с воспроизводящим ядром $K(x, x')$, которое будет определено далее.
3. Регуляризирующий функционал $W(f)$: автор использует квадрат нормы функции $W(f) = (f, f)$.

## 11. Оценки с использованием V-матриц

Зададим расстояние в рассматриваемой плохо обусловленной задаче оценки функции условной вероятности в множестве $f(x, \alpha)$, $\alpha \in \Lambda$, в виде

$$\rho(A_\ell f, F_\ell) = \int \left( \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i) f(x_i, \alpha) - \frac{1}{\ell} \sum_{j=1}^{\ell} y_j \theta(x - x_j) \right)^2 d\mu(x),$$

где $\mu(x)$ — заданная вероятностная мера.

В статье вместо меры $\mu(x)$ автор использует ее эмпирическую оценку

$$\mu_\ell(x) = \frac{1}{L} \sum_{t=1}^{L} \theta(x - x'_t),$$

вычисленную на основании $L$ элементов универсума (набора элементов, который содержит $\ell$ элементов $x_i$ обучающих данных и $(L - \ell)$ векторов, сгенерированных мерой $\mu(x) \approx P(x)$)[13]. Получим

$$\rho^2(A_\ell f, F_\ell) = \sum_{i,j=1}^{\ell} \left( f(x_i, \alpha) f(x_j, \alpha) - 2 y_j f(x_i, \alpha) + y_i y_j \right) V(x_i, x_j), \tag{11.1}$$

где обозначено

$$V(x_i, x_j) = \int \theta(x - x_i) \theta(x - x_j) d\mu_\ell(x) = \frac{1}{L} \sum_{t=1}^{L} \prod_{k=1}^{n} \theta(x_t^k - x_{i \vee j}^k). \tag{11.2}$$

В (11.2) используется обозначение $x_{i \vee j}^k = \max\{x_i^k, x_j^k\}$, где $x_i^k$ — $k$-я координата векторов $x_i = (x_i^1, \ldots, x_i^n)$ из обучающей выборки. Векторы $x_t$ обозначают элементы из универсума. Чем больше число $L$ векторов $x_t$, тем точнее оценка матрицы $V$. Уравнение (11.1) можем переписать в виде

$$\rho(A_\ell f, F_\ell) = \sum_{i,j=1}^{\ell} (f(x_i, \alpha) - y_i)(f(x_j, \alpha) - y_j) V(x_i, x_j). \tag{11.3}$$

[13] Например, при распознавании цифр элементы универсума наряду с примерами написания цифр могут включать и примеры написания букв.

Таким образом, расстояние (11.3) между двумя функциями, определяющими левую и правую части уравнения, учитывает наряду с невязкой $\delta_i = (y_i - f(x_i, \alpha))$ попарные положения наблюдаемых векторов относительно векторов из универсума. Функционал (11.3) отличается от стандартного функционала из метода наименьших квадратов

$$\rho(A_\ell f, F_\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2.$$

*Замечание 6.* Для удобства вычислений рекомендуется вместо матрицы $V$ использовать нормализованную матрицу $V/||V||$, где $||V||$ — любая норма матрицы $V$.

### 12. Функции из гильбертова пространства с воспроизводящим ядром

Пусть $K(x, x')$ — положительно определенная функция[14]. Говорят, что функция $f(x)$ принадлежит гильбертовому пространству с воспроизводящим ядром $K(x, x')$, если выполняется следующее *воспроизводящее* свойство:

$$(K(x, x'), f(x')) = f(x) \tag{12.1}$$

(в скалярном произведении (12.1) автор рассматривает $K(x, x')$ как функцию $x'$ при фиксированном $x$). Легко видеть, что существует простая конструкция скалярного произведения, которая определяет гильбертово пространство с воспроизводящим ядром $K(x, x')$. Действительно, согласно теореме Мерсера любая положительно определенная функция $K(x, x')$ (ядро) имеет представление

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_p \psi_p(x) \psi_p(x'),$$

где $\psi_1(x), \ldots, \psi_p(x), \ldots$ — система ортонормированных функций и $\lambda_1 \geqslant \geqslant \lambda_2 \ \ldots \geqslant \lambda_p \ \ldots \geqslant 0$, причем $\lambda_p$ сходится к нулю при $p \to \infty$. Рассмотрим гильбертово пространство функций вида

$$f(x) = \sum_{p=1}^{\infty} c^p \psi_p(x) \tag{12.2}$$

со скалярным произведением

$$(f_1(x), f_2(x)) = \sum_{p=1}^{\infty} \frac{c_1^p c_2^p}{\lambda_p}, \tag{12.3}$$

[14] Положительно определенная функция удовлетворяет условию $\sum\limits_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0$ для любого $m$, любых $x_i, \ldots, x_m$ и любых $\alpha_i,\ i = 1, \ldots, m$.

где $c_1 = (c_1^1, \ldots, c_1^p, \ldots)$ и $c_1 = (c_2^1, \ldots, c_2^p, \ldots)$ — коэффициенты разложения (37) функций $f_1(x)$ и $f_2(x)$. Норма функции в этом пространстве равна

$$||f(x)||^2 = \sum_{p=1}^{\infty} \frac{(c^p)^2}{\lambda_p}. \tag{12.4}$$

Заметим, что в гильбертовом пространстве, в котором функции представимы разложением по ортонормированным базисным функциям $\psi_p(x)$, скалярное произведение определяется соотношением

$$(f_1(x), f_2(x)) = \sum_p c_1^p c_2^p,$$

а норма —

$$||f(x)||^2 = \sum_p (c^p)^2$$

(т.е. имеют вид (12.3) и (12.4) при $\lambda_p = 1$).

### *12.1. Гильбертово пространство с воспроизводящим ядром*

В статье при оценке условной вероятности автор использует в качестве регуляризирующего функционала квадрат нормы $||f||^2$ с параметром регуляризации $\gamma > 0$ (см. 10.3)). Это эквивалентно поиску функций вида (12.2) с коэффициентами разложения $c^p$, принадлежащих области

$$||f(x)||^2 = \sum_{p=1}^{\infty} \frac{(c^p)^2}{\lambda_p} \leqslant C, \tag{12.5}$$

где константа $C$ определяется константой регуляризации $\gamma > 0$ в (10.3) (и наоборот).

Поскольку $\lambda_p$ в знаменателе в левой части (12.5) стремится к нулю при $p$, стремящемся к бесконечности, неравенство (12.5) задает множество гладких функций: коэффициенты в разложении (12.2) функции с ограниченной нормой должны стремиться к нулю при $p \to \infty$. Для любого фиксированного $C > 0$ множество функций из гильбертова пространства с воспроизводящим ядром формирует компакт и, следовательно, квадрат нормы может использоваться как регуляризирующий функционал $W(f) = ||f||^2$ в (10.3) при решении плохо обусловленных задач.

### *12.2. Теорема о представителе*

Важную роль в теории и методах Reproducing Kernel Hilbert Space (RKHS) играет так называемая теорема о представителе.

*Теорема 5. Функция $f(x)$ из RKHS, которая минимизирует функционал*

$$R(f) = \sum_{i=1}^{\ell} L(y_i - f(x_i)) + \gamma ||f(x)||^2,$$

*наряду с разложением* (12.2) *представима в виде*

$$f(x, \alpha_\ell) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x). \tag{12.6}$$

*Квадрат нормы этой функции наряду с* (12.4) *представим в виде*

$$||f(x, \alpha_\ell)||^2 = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i x_j). \tag{12.7}$$

Другими словами, в гильбертовом пространстве с воспроизводящим ядром $K(x, x')$ функция, минимизирующая $R(f)$, наряду с разложением по бесконечному числу ортонормированных базисных функций разложима по конечному числу функций, задаваемых ядром $K(x, x')$. Это делает функции из RKHS с ядром $K(x, x')$ важным инструментом аппроксимации функций (в особенности функций высокой размерности).

### *12.3. Примеры ядер: ядра Мелера и RBF*

В теории аппроксимации важную роль играют полиномы Эрмита и функции Эрмита. Полиномы Эрмита $H_1(x), \ldots, H_k(x), \ldots$, определяют ортогональные с весовой функцией $\exp\{-x^2\}$ полиномы, т.е.

$$\int_{-\infty}^{\infty} P_s(x) P_n(x) e^{-x^2} dx = 2^n n! \sqrt{\pi} \delta_{n,s},$$

где $\delta_{n,s}$ — символ Кронекера ($\delta_{s,n} = 0$, если $s \neq n$, и $\delta_{s,n} = 1$, если $s = n$).

Ортонормированные функции

$$e_n(x) = \left(2^n n! \sqrt{\pi}\right)^{-1/2} H_n(x) e^{-\frac{x^2}{2}}$$

называются функциями Эрмита. Показано, что

$$|e_n(x)| < 0{,}82.$$

Функции Эрмита образуют ортонормированный базис в $L_2$. Таким образом, любая функция из $L_2$ однозначно представима в виде разложения по функциям Эрмита.

В 1866 г. Мелер показал, что для любого $0 < \varepsilon < 1$ справедливо соотношение

$$K_M(x, x') = \sum_{k=0}^{\infty} \varepsilon^k e_k(x) e_k(x') = \frac{1}{\sqrt{1-\varepsilon^2}} \exp\left\{ -\frac{\varepsilon^2 (x-x')^2 + 2(\varepsilon - \varepsilon^2) x x'}{1-\varepsilon^2} \right\}.$$

Правая часть этого равенства называется ядром Мелера. Нормализованное ядро Мелера называется RBF (радиальные базисные функции) ядром

$$K(x, x') = \frac{K_M(x, x')}{\sqrt{K_M(x, x) K_M(x' x')}} = \exp\{-\delta (x - x')^2\},$$

где параметр $\delta > 0$ задается величиной $\varepsilon$:

$$\delta = \frac{\varepsilon^2}{1 - \varepsilon^2}.$$

## 13. Полное решение задачи классификации

### *13.1. Решение Второй задачи выбора*

Получим полное решение задачи распознавания образов для двухклассовой задачи классификации на множестве функций из гильбертова пространства с воспроизводящим ядром $K(x, x')$. Решим уравнение (9.2) на множестве функций (12.6), принадлежащих RKHS с ядром $K(x, x')$ как плохо обусловленную задачу, используя регуляризационный функционал (12.7).

Согласно теореме о представителе решение $P_\ell(y = 1|x) = f_\ell(x)$ имеет вид

$$f_\ell(x) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x)$$

и квадрат нормы $||f_\ell(x)||$ равен

$$||f_\ell(x)||^2 = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j).$$

Для простоты будем использовать следующие векторно-матричные обозначения. Пусть $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ — элементы обучающей выборки. Рассматриваем вектор $Y = (y_1, \ldots, y_\ell)^{\mathrm{T}}$, определенный элементами обучающей выборки. Для двухклассовой задачи $Y$ является бинарным вектором. Рассматриваем $(\ell \times \ell)$-матрицу $K = K(x_i, x_j)$, определяемую элементами обучающей выборки. Рассмотрим вектор-функцию $\mathcal{K}(x) = (K(x_1, x), \ldots, K(x_\ell, x))^{\mathrm{T}}$ размерности $\ell$ и вектор коэффициентов разложения $A = (\alpha_1, \ldots, \alpha_\ell)^{\mathrm{T}}$ размерности $\ell$.

В этих обозначениях по теореме о представителе функция $f_\ell(x)$ имеет вид

$$f_\ell(x) = A^{\mathrm{T}} \mathcal{K}(x).$$

Введем $(\ell \times \ell)$-матрицу $V = ||V(x_i, x_j)||$ с элементами $V(x_i, x_j)$. В данных обозначениях для решения второй задачи выбора необходимо минимизировать регуляризованный функционал

$$R(A) = (KA - Y)^{\mathrm{T}} V (KA - Y) + \gamma A^{\mathrm{T}} KA.$$

Однако автор будет искать решение задачи в более общем множестве функций, рассматривая множество функций из RKHS, которые имеют свободный член

$$f_\ell(x) = A^{\mathrm{T}} \mathcal{K}(x) + c.$$

Для нахождения оценки минимизируем функционал

$$R(A, c) = (KA + c1_\ell - Y)^{\mathrm{T}} V (KA + c1_\ell - Y) + \gamma A^{\mathrm{T}} KA,$$

где используем $\ell$-мерный вектор $1_\ell = (1, \ldots, 1)^{\mathrm{T}}$. Необходимыми условиями минимума $R(A, c)$ являются

$$\text{(13.1)} \qquad \begin{cases} \dfrac{\partial R(A,c)}{\partial A} \implies VKA + cV1_\ell - VY + \gamma_\ell A = 0, \\ \dfrac{\partial R(A,c)}{\partial c} \implies 1_\ell^{\mathrm{T}} VKA + c1_\ell^{\mathrm{T}} V1_\ell - 1_\ell^{\mathrm{T}} VY = 0. \end{cases}$$

Из первого равенства в (13.1) получим

$$(VK + \gamma I)A = VY - cV1_\ell$$

или

$$\text{(13.2)} \qquad A = (VK + \gamma I)^{-1} V(Y - c1_\ell).$$

Используя обозначения

$$\text{(13.3)} \qquad B = (VK + \gamma I)^{-1} V,$$

из (13.2) получим, что требуемый вектор $A$ имеет вид

$$\text{(13.4)} \qquad A = B(Y - c1_\ell).$$

Из второго равенства в (13.1) и (13.4) получим уравнение для определения $c$:

$$1_\ell^{\mathrm{T}} VKB(Y - c1_\ell) + c1_\ell^{\mathrm{T}} V1_\ell - 1_\ell^{\mathrm{T}} VY = 0,$$
$$\left[1_\ell^{\mathrm{T}} VKB1_\ell - 1_\ell^{\mathrm{T}} V1_\ell\right] \cdot c \;=\; \left[1^{\mathrm{T}} VKBY - 1_\ell^{\mathrm{T}} VY\right].$$

Значением свободного члена $c$ является

$$\text{(13.5)} \quad c = \frac{\left[1_\ell^{\mathrm{T}} VKB1_\ell - 1_\ell^{\mathrm{T}} VY\right]}{\left[1_\ell^{\mathrm{T}} VKB1_\ell - 1_\ell^{\mathrm{T}} V1_\ell\right]} = \frac{1_\ell^{\mathrm{T}} \left[VK(VK + \gamma I)^{-1} - I\right] VY}{1_\ell^{\mathrm{T}} \left[VK(VK + \gamma I)^{-1} - I\right] V1_\ell} = \frac{1_\ell^{\mathrm{T}} SVY}{1_\ell^{\mathrm{T}} SV1_\ell},$$

где введено обозначение

$$S = VK(VK + \gamma I)^{-1} - I.$$

Подставляя $c$ в (13.4), получим требуемые параметры $A$.

### *13.2. Калибровка n функций условных вероятностей*

Рассмотрим задачу $n$-классовой классификации, в которой величина $y$ принимает одно из $n$ значений $y \subset \{1, \ldots, n\}$. Определим за вектор $Y_p$ бинарный вектор, координата $s$ которого равна единице, если в паре $s$ обучающей выборки $y_s = p$, и равна нулю, если $y_s \neq p$. Пусть для любого фиксированного $p$ функция

$$f(x; A_p, c_p) = A_p^{\mathrm{T}} \mathcal{K}(x) + c_p, \quad p = 1, \tag{13.6}$$

определяет условную вероятность $P(y = p|x)$ класса $y = p$ при заданном $x$.

Допустим, что для всех $p$ оценок пар $(A_p, c_p)$ использовалось одно и то же ядро $K$, минимизируя функционалы

$$R(A_p, c_p) = (KA_p + c_p 1_\ell - Y_p)^{\mathrm{T}} V (KA_p + c_p 1_\ell - Y_p) + \gamma A_p^{\mathrm{T}} K A_p, \quad p = 1, \ldots, n, \tag{13.7}$$

для $n$ различных $Y_p$ с одним и тем же параметром регуляризации $\gamma > 0$. Легко убедиться (используя (13.5)), что поскольку $\sum_{s=1}^{n} Y_s = 1_\ell$, имеем, что

$$\sum_{p=1}^{n} c_k = \sum_{p=1}^{n} \frac{1_\ell^{\mathrm{T}} S V Y_p}{1_\ell^{\mathrm{T}} S V 1_\ell} = 1. \tag{13.8}$$

Используя (13.4), получаем

$$\sum_{p=1}^{n} A_p = \sum_{p=1}^{n} \left( (VK + \gamma I)^{-1} V Y_p \right) - \sum_{p=1}^{n} c_p \left( (VK + \gamma I)^{-1} V 1_\ell \right) V 1_\ell = 0_\ell, \tag{13.9}$$

где $0_\ell$ обозначает $\ell$-мерный вектор, состоящий из нулей. Из (13.8) и (13.9) получаем, что

$$\sum_{p=1}^{n} P_\ell(y = p|x) = \sum_{p=1}^{n} c_p = 1 \tag{13.10}$$

для всех $x \in X$.

Свойство (13.10) $n$ оценок называется *совместной калибровкой* $n$ условных плотностей.

### *13.3. Одновременное решение обеих проблем выбора*

Чтобы найти функцию условной вероятности $P(y = p|x)$, которая является решением обеих проблем выбора, минимизируем (13.7) при $m$ ограничениях (инвариантах)

$$\Phi_s^{\mathrm{T}} K A_p + c_p \Phi_s^{\mathrm{T}} 1_\ell = \Phi_s^{\mathrm{T}} Y_p, \quad s = 1, \ldots, m,$$

где $\Phi_s$ — $\ell$-мерный вектор, сконструированный с использованием предикат $\phi_s(x)$ и элементов $x_i$ обучающей выборки:

$$\Phi_s = (\phi_s(x_1), \ldots, \phi_s(x_\ell))^{\mathrm{T}}, \quad s = 1, \ldots, m.$$

Метод условной минимизации. В [9] для случая $\ell > m$ найдено следующее решение этой проблемы: вектор параметров $A_p$ функции (13.6) имеет вид

$$A_p = B(Y_p - c_p 1_\ell) - \sum_{s=1}^{n} \mu_s A_s,$$

где наравне с обозначением (13.3) использовано обозначение

$$A_s = (VK + \gamma I)^{-1} \Phi_s, \quad s = 1, \ldots, n.$$

Коэффициенты $\mu_s$ и $c_p$, которые определяют вектор $A_p$ и свободный член $c_p$ в (13.6), являются решениями следующих линейных уравнений:

$$c_p \left[1_\ell^{\mathrm{T}} VKB 1_\ell - 1_\ell^{\mathrm{T}} V 1_\ell\right] + \sum_{s=1}^{m} \mu_s \left[1_\ell^{\mathrm{T}} VKA_s - 1_\ell^{\mathrm{T}} \Phi_s\right] = \left[1_\ell^{\mathrm{T}} VKBY_p - 1_\ell^{\mathrm{T}} VY_p\right],$$

$$c_p \left[\Phi_k KB 1_\ell - \Phi_k^{\mathrm{T}} 1_\ell\right] + \sum_{s=1}^{m} \mu_s A_s^{\mathrm{T}} K \Phi_k = \left[\Phi_s^{\mathrm{T}} KBY_p^{\mathrm{T}} - Y_p^{\mathrm{T}} \Phi_k\right], \quad k = 1, \ldots, m.$$

Метод безусловной минимизации. Далее найдем приближенное решение рассматриваемой проблемы (которое включает случай $\ell < m$) путем минимизации функционала

$$\begin{aligned} L(A, c) = {} & (KA_p + c_p 1_\ell - Y_p)^{\mathrm{T}} V (KA + c 1_\ell - Y_p) - \\ & - 2(KA_p + c_p 1_\ell - Y_p)^{\mathrm{T}} VY_p + \gamma A_p^{\mathrm{T}} KA_p + \\ & + \tau \sum_{s=1}^{m} \frac{\mu_s}{m} (\Phi_s^{\mathrm{T}} KA_p + c_p \Phi_s^{\mathrm{T}} 1_\ell - \Phi_s^{\mathrm{T}} Y_p)^2, \end{aligned} \tag{13.11}$$

где $\tau \geqslant 0$ — дополнительный параметр функционала и

$$\mu_s = \left(\Phi_s^{\mathrm{T}} \Phi_s\right)^{-1}.$$

Как и ранее, получим необходимые условия минимума

$$\begin{aligned} & \frac{\partial L(A_p, c_p, )}{\partial A_p} \Longrightarrow V(KA_p + c_p 1_\ell - Y_p) + \gamma A_p + \tau M(KA_p + c 1_\ell - Y_p) = 0, \\ & \frac{\partial L(A_p, c_p, )}{\partial c_p} \Longrightarrow 1_\ell^{\mathrm{T}} V(KA_p + c_p 1_\ell - Y_p) + \tau 1_\ell^{\mathrm{T}} M(KA_p + c_p 1_\ell - Y_p) = 0. \end{aligned} \tag{13.12}$$

В (13.12) использовано матричное обозначение

$$M = \frac{1}{m}\sum_{s=1}^{m}\mu_s\Phi_s\Phi_s^{\mathrm{T}} = \frac{1}{m}\sum_{s=1}^{m}\frac{\Phi_s\Phi_s^{\mathrm{T}}}{\Phi_s^{\mathrm{T}}\Phi_s}. \tag{13.13}$$

*Замечание 7.* В случае когда предикаты $\phi_s(x)$ таковы, что $d_s = \Phi_s^{\mathrm{T}}1_\ell = 0$ (это справедливо, если рассматривать в качестве предикат выражение $\phi'_s(x) = \phi_s(x) - d_s$), элементы $M_{i,j}$ матрицы $M$ задают "корреляцию" между векторами $x_i$ и $x_j$ относительно предикат $\phi'_1(x), \ldots, \phi'_m(x)$.

Чтобы получить решение в компактном виде, введем обозначение

$$W_\tau = V + \tau M.$$

Используя обозначение из первого уравнения (13.12), получим

$$(W_\tau K + \gamma I)A_p = W_\tau Y_p - c_p W_\tau 1_\ell$$

и выражение

$$A_p = (W_\tau K + \gamma I)^{-1}W_\tau(Y_p - c_p 1_\ell). \tag{13.14}$$

Введя векторы

$$A_b^p = (W_\tau K + \gamma I)^{-1}W_\tau Y_p$$

и

$$A'_c = (W_\tau K + \gamma I)^{-1}W_\tau 1_\ell,$$

из (13.14) найдем требуемый вектор $A_p$ в виде

$$A_p = A_b^p - c_p A'_c. \tag{13.15}$$

Подставляя выражение (13.15) во второе уравнение (13.12), найдем свободный член $c_p$

$$c_p = \frac{\left[1_\ell^{\mathrm{T}}W_\tau K A_p^b - 1_\ell^{\mathrm{T}}W_\tau Y_p\right]}{\left[1_\ell^{\mathrm{T}}W_\tau K A'_c - 1_\ell^{\mathrm{T}}W_\tau 1_\ell\right]} = \frac{1_\ell^{\mathrm{T}}\left[W_\tau K(W_\tau K + \gamma I)^{-1} - I\right]W_\tau Y_p}{1_\ell^{\mathrm{T}}\left[W_\tau K(W_\tau K + \gamma I)^{-1} - I\right]W_\tau 1_\ell}. \tag{13.16}$$

Подставляя $c_p$ в (13.15), найдем вектор $A_p$. Параметр $c_p$ и вектор $A_p$ определяют выражение (13.6) для оценки функции условной вероятности.

### *13.4. Решение n-классовой проблемы классификации*

При решении задачи классификации $n$ классов оцениваем $n$ функций условных вероятностей $P(y = p|x)$, $p \subset \{1, \ldots, n\}$. Используя эти оценки, сконструируем правило

$$r(x) = \operatorname{argmax}\{P_\ell(y = 1|x), \ldots, P_\ell(y = n|x)\}.$$

Допустим, что для оценки $n$ функций условных вероятностей используем одно и то же ядро $K(x,x')$, один и тот же набор предикатов $\phi_1(x),\ldots,\phi_m(x)$ и один и тот же параметр регуляризации $\gamma$. Тогда матрицы $K$ и $W_\tau$ не зависят от $p$ (класс $y=p$, для которого оцениваем условную вероятность). В этом случае поскольку

$$\sum_{p=1}^{n} Y_p = 1_\ell,$$

то, используя (13.15) и (13.16), получим, что

$$\sum_{p=1}^{n} c_p = 1, \quad \sum_{p=1}^{n} A_p = 0_\ell,$$

т.е.

$$\sum_{p=1}^{n} P_\ell(y=p|x) = 1$$

для всех $x \in X$. Другими словами, получаем $n$ совместно калиброванных решений, которые решают обе проблемы выбора.

## 14. Заключение

В настоящей статье рассмотрена общая модель проблемы обучения: найти в множестве функций $\{f(x)\}$ функцию $f_\ell(x)$, минимизирующую функционал среднего риска

$$R(f) = \int (y - f(x))^2 dP(x,y)$$

при неизвестной кумулятивной функции распределения $P(x,y)$, но при заданной независимой выборке одинаково распределенных пар

$$(x_1,y_1),\ldots,(x_\ell,y_\ell).$$

Показано, что для решения этой задачи необходимо в $\{f\}$ минимизировать целевой функционал

$$R_\ell(f) = (Y - F(f))^{\mathrm{T}} \mathcal{V} (Y - f(F)),$$

где $Y = (y_1,\ldots,y_\ell)^{\mathrm{T}}$, $F(f) = (f(x_1,\ldots,f(x_\ell))^{\mathrm{T}}$ и $\mathcal{V}$ — матрица, состоящая из элементов $V(x_i,x_j)$. Используя $m$ предикат $\phi_1(x),\ldots,\phi_m(x)$, можно извлечь из данных дополнительную информацию. Для этого необходимо в $\{f\}$ минимизировать функционал (13.11) при ограничениях

$$\Phi_k^{\mathrm{T}} F(f) = \Phi_k^{\mathrm{T}} Y, \quad k = 1,\ldots,m.$$

Задачу условной минимизации в $\{f\}$ можно переписать в виде задачи безусловной минимизации, минимизируя в $\{f\}$ функционал

$$R_{un}(f) = (Y - F(f))^{\mathrm{T}}\mathcal{V}(Y - f(F)) + \frac{\tau}{m}\sum_{k=1}^{m}\frac{1}{\Phi_k^{\mathrm{T}}\Phi_t}(\Phi_k^{\mathrm{T}}F(f) - \Phi_k^{\mathrm{T}}Y)^2,$$

который можно переписать в виде

$$R_{un}(f) = (Y - F(f))^{\mathrm{T}}(\mathcal{V} + \tau\mathcal{P})(Y - f(F)),$$

где матрица $\mathcal{P}$ равна:

$$\mathcal{P} = \frac{1}{m}\sum_{k=1}^{m}\frac{\Phi_k\Phi_k^{\mathrm{T}}}{\Phi_k^{\mathrm{T}}\Phi_k}.$$

В статье получено решение для функций с ограниченной нормой, принадлежащих RKHS. Функции из RKHS с ядром $K(x, x')$ допускают линейное представление с вектором параметров $A$

$$f_\ell(x) = A^{\mathrm{T}}\mathcal{K}(x),$$

при этом $F(f) = KA$, а норма функции $||f||^2 = A^{\mathrm{T}}KA$, где $K$ — матрица с элементами $K(x_i, x_j)$. В статье для этого случая получено решение.

Для множества функций $\{f\}$, задаваемых нейронными сетями, минимизация функционала (13.13) методом стохастического градиентного спуска приводит к методу *$\mathcal{VP}$-обратного распространения*, в котором, используя стандартную схему обратного распространения, вместо обратного распространения вектора ошибок $E = (e_1, \ldots, e_\ell)^{\mathrm{T}}$ разностей $e_i = y_i - z_i$ (значение целевой переменной $y_i$ из обучающей выборки $(x_i, y_i)$ минус выход $z_i$ последнего слоя нейронной сети в ответ на входной вектор $x_i$) осуществляется обратное распространение модифицированного вектора $(\mathcal{V} + \tau\mathcal{P})E$.

## СПИСОК ЛИТЕРАТУРЫ


1. *Вапник В.Н., Червоненкис А.Я.* О равномерной сходимости частот появления событий к их вероятностям // Теория вероятн. и ее примен. 1971. Т. 16. № 2. С. 264–279.

   *Vapnik V.N., Chervonenkis A.Ya.* On Uniform Convergence of the Frequencies of Events to Their Probabilities // Theory Probab. Apl. 1971. V. 16. No. 2. P. 264–280.
2. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974.
3. *Vapnik V.* The Nature of Statistical Learning Theory. Springer, 1995.
4. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М.: Наука, 1979.
5. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.
6. *Devroy L., Geofry L., Lugosi G.* A Probabilistic Theory of Pattern Recognition. Springer, 1996.

7. *Vapnik V.* Statistical Learning Theory. N.Y.: J. Wiley, 1998.
8. *Вапник В.Н., Стефанюк А.Р.* Непараметрические методы восстановления плотности вероятностей // АиТ. 1978. № 8. С. 38–52.
*Vapnik V.N., Stefanyuk A.R.* Nonparametric Methods for Restoring the Probability Densities // Autom. Remote Control. 1979. V. 39. No. 8. P. 1127–1140.
9. *Vapnik V., Izmailov R.* Rethinking Statistical Learning Theory: Learning Using Statistical Invariants // Machine Learning. 2018. V. 108. P. 381–423.

*Статья представлена к публикации членом редколлегии А.В. Назиным.*

Поступила в редакцию 13.07.2018

После доработки 05.09.2018

Принята к публикации 08.11.2018