M. A. Aizerman, È. M. Braverman, L. I. Rozonoèr, Probability problem of pattern recognition learning and potential functions method, *Avtomat. i Telemekh.*, 1964, Volume 25, Issue 9, 1307–1323

Том XXV «АВТОМАТИКА И ТЕЛЕМЕХАНИКА» **№ 9**

1964

УДК 62-507

ВЕРОЯТНОСТНАЯ ЗАДАЧА ОБ ОБУЧЕНИИ АВТОМАТОВ РАСПОЗНАВАНИЮ КЛАССОВ И МЕТОД ПОТЕНЦИАЛЬНЫХ ФУНКЦИЙ

М. А. АЙЗЕРМАН, Э. М. БРАВЕРМАН, Л. И. РОЗОНОЭР

(*Москва*)

Дается вероятностная постановка задачи об обучении автоматов разделению входных ситуаций на классы. Излагается алгоритм, основанный на методе потенциальных функций и позволяющий построить функцию, аппроксимирующую степень достоверности того, что возникающая на входе автомата ситуация принадлежит тому или иному классу. Доказывается теорема о сходимости (в вероятностном смысле) выстраиваемой алгоритмом функции к этой степени достоверности.

1. Постановка задачи

В работе рассматривается задача об обучении автоматов распознаванию классов, являющаяся обобщением задачи, рассмотренной в [1]. Пусть на входе в автомат появляются ситуации, каждая из которых может относиться к одному из двух классов * — A или B. В отличие от [1] в процессе обучения одна и та же ситуация может быть при различных показах отнесена к разным классам. При этом естественно предположить, что для каждой ситуации существуют вероятности принадлежности этой ситуации к классам A и B, а в процессе обучения каждая ситуация относится к A или B в соответствии с этими вероятностями.

Множество всех ситуаций, которые могут появиться на входе в автомат, образует пространство X. В соответствии со сказанным выше предполагается, что объективно существуют заданные на всем пространстве X функции $D_A(x)$ и $D_B(x) = 1 - D_A(x)$ — вероятности того, что точка x принадлежит соответственно A или B. Эти функции в дальнейшем будем называть «степенями достоверности» принадлежности точки x классам A или B. Задача состоит в том, чтобы по появляющимся в процессе обучения точкам и по информации, которая сообщается «учителем», о том, к какому множеству (A или B) он относит эти точки **, восстановить $D_A(x)$ и $D_B(x)$ как функции, заданные на всем пространстве *** X.

* Решение задачи о разделении ситуаций на два класса позволяет решать также задачи о распознавании любого конечного числа классов (например, при помощи последовательной дихотомии).

** Как уже указывалось, «учитель» относит появляющиеся точки к A или B не однозначно, а в соответствии с объективно существующими (хотя, быть может, и неизвестными ему!) степенями достоверности.

*** Казалось бы, такая постановка задачи означает, что, в отличие от [1], каждая точка пространства X с вероятностью единица относится «учителем» к $A \cup B$ (поскольку $D_A + D_B = 1$). Однако, если в процессе обучения и в экзамене могут появляться не все точки из X, а лишь точки, принадлежащие некоторому подмножеству пространства X, то функции $D_A(x)$ и $D_B(x)$ фактически заданы только на этом подмножестве, а в остальной части пространства значения их безразличны и они доопределяются удобным для нас образом.

Детерминистская постановка задачи о разделении классов, сформулированная в [1], и рассматриваемая здесь вероятностная постановка подобной задачи отличаются, во-первых, предположениями о классах A и B и, соответственно, характером информации, сообщаемой машине в процессе обучения, и, во-вторых, тем, какие функции подлежат восстановлению в процессе обучения. Именно, в [1] предполагалось, что в пространстве X объективно существуют непересекающиеся множества точек (например, A и B) и что поэтому всегда существуют разделяющие их функции; при показе точек из A или B «учитель» достоверно сообщает, к какому множеству они принадлежат; цель обучения состояла в построении какой-либо из этих разделяющих функций, т. е. функции, принимающей положительные значения на всех (а не только показанных в процессе обучения) точках из A и отрицательные значения на всех точках из B.

В настоящей работе предполагается, что множества A и B могут «пересекаться». Поэтому не существует разделяющая их функция, но существуют указанные выше функции степеней достоверности, в связи с чем указание «учителя» о принадлежности точки к A или B не является достоверным. Цель же процесса обучения состоит в восстановлении степеней достоверности.

Вероятностная постановка задачи охватывает детерминистскую постановку как частный случай, характеризующийся тем, что $D_A(x)$ и $D_B(x)$ принимают лишь значения, равные 0 или 1 на точках из A или B.

В качестве примера рассмотрим задачу об обучении машины прогнозированию исхода заболеваний по клиническим данным. Встречаются случаи, когда исход болезни может быть однозначно предсказан. В таких случаях возникает детерминистская задача, описанная в [1]. Часто, однако, клинические данные не дают оснований для однозначного предсказания исхода болезни, однако накопленная опытом медицины статистика дает вероятности исхода болезни. Если сообщать машине клинические данные о конкретных больных и исходы их болезней, а от машины требовать, чтобы она в результате обучения правильно определяла вероятность исхода в новых случаях, то как раз и возникает задача, рассматриваемая в этой работе.

В качестве технического примера рассмотрим типовую задачу об обнаружении какого-либо объекта локатором на фоне сильных помех. Одна и та же «картинка», появляющаяся на экране локатора, может из-за сильных помех соответствовать как наличию, так и отсутствию искомого объекта. Поэтому с каждой «картинкой» связывается лишь вероятность (степень достоверности) наличия объекта. Задача состоит в том, чтобы по отдельным наблюдаемым в процессе обучения случаям, когда факт наличия или отсутствия объекта точно установлен, научить машину правильно определять степень достоверности наличия того же объекта для новых «картинок».

Настоящая работа посвящена решению задачи о восстановлении степеней достоверности в предположении, что функции $D_A(x)$ и $D_B(x)$ могут быть представлены конечным рядом в разложении по некоторой системе ортонормированных функций. В принципе эта задача могла бы быть решена методом максимального правдоподобия *. Однако при принятом предположении о виде функции степени достоверности уравнения, вытекающие из метода максимума правдоподобия, практически неразрешимы (в частности, из-за большого числа оцениваемых параметров — коэффициентов разложения). В настоящей работе для решения задачи используется некоторая модификация метода потенциальных функций, идея которого была изложена в [1].

---

* Авторы благодарят С. Я. Виленкина за обсуждение вопроса о применимости метода максимума правдоподобия к решению поставленной задачи.

Однако возможен и в настоящее время используется другой («бейесовский») способ восстановления степеней достоверности. Именно, по показанным в процессе обучения точкам первоначально восстанавливаются не функции $D_A(x)$ и $D_B(x)$, а плотности вероятности появления в X точек из A и из B — $P(x|A)$ и $P(x|B)$ соответственно. Одновременно оцениваются вероятности P_A и P_B появления точки из A и B.

По окончании процесса обучения, при появлении новой точки x^* степени достоверности, представляющие собой условные вероятности принадлежности классам A и B при условии появления точки x^*, подсчитываются по формуле Бейеса

$$D_A(x^*) \equiv P(A\,|\,x^*) = \frac{P(x^*\,|\,A)\,P_A}{P(x^*)}, \qquad D_B(x^*) \equiv P(B\,|\,x^*) = \frac{P(x^*\,|\,B)\,P_B}{P(x^*)}, \tag{1}$$

где $P(x^*) = P(x^*|A)P_A + P(x^*|B)P_B$ — плотность вероятности появления x^*.

Выходом автомата могут быть найденные значения $D_A(x)$ и $D_B(x)$ или решения автомата о том, к какому множеству относится точка x^* (определяемое тем, какое из этих чисел больше).

Поскольку вероятности P_A и P_B легко оцениваются, то задача определения $D_A(x)$ и $D_B(x)$ фактически сводится к восстановлению плотностей вероятности $P(x|A)$ и $P(x|B)$.

При таком способе восстановления степеней достоверности необходимо сделать некоторые предположения о классе функций, к которому принадлежат восстанавливаемые плотности вероятности $P(x|A)$ и $P(x|B)$. Так, часто предполагают, что эти плотности вероятности имеют известный вид (например, являются гауссовскими), а задача их восстановления сводится к статистической оценке неизвестных заранее параметров распределения.

Однако класс функций, к которому принадлежат плотности вероятности, может быть таким, что их восстановление требует недопустимо большого числа показов, в то время, как непосредственное восстановление $D_A(x)$ и $D_B(x)$ может быть произведено по небольшому числу показов *. Это имеет место, например, когда $P(x|A)$ и $P(x|B)$ разрывны, а $D_A(x)$ и $D_B(x)$ непрерывны и достаточно гладки. Поэтому способы непосредственного восстановления степени достоверности, не связанные с промежуточным восстановлением $P(x|A)$ и $P(x|B)$ и с использованием формулы Бейеса, вообще говоря, являются предпочтительными. В тех же случаях, когда восстанавливаемые плотности вероятности достаточно «гладкие», «бейесовский» подход вполне оправдан. В таких случаях для задачи восстановления плотности вероятности также может быть использован метод потенциальных функций. Поскольку при этом применение метода потенциальных функций наиболее удобно, в следующем разделе будет рассмотрена задача о восстановлении неизвестной плотности вероятности. Другие разделы посвящены непосредственному восстановлению степеней достоверности.

2. Восстановление плотности вероятности

Метод решения задачи о восстановлении неизвестной плотности вероятности, основанный на определенных предположениях о свойствах коэффициентов разложения этой функции по некоторой ортонормированной

* Более того, может оказаться, что функции $P(x\,|\,A)$ и $P(x\,|\,B)$, строго говоря, не существуют вообще, в то время как $D_A(x)$ и $D_B(x)$ существуют и могут быть восстановлены с достаточной точностью по небольшому числу показов.

системе функций, подробно рассмотрен в [2, 3] (см. также [4]). В настоящем разделе, имеющем методический характер, аналогичная задача, но при более простых предположениях о классе функций, к которому принадлежит плотность вероятности, решается при помощи метода потенциальных функций. Применительно к этой задаче он совпадает, по существу, с методом, рассмотренным в [2, 3].

Пусть в пространстве X существует плотность вероятности $P(x)$. В соответствии с этой плотностью вероятности независимо появляются точки $x^1, x^2, \ldots, x^k$. Задача состоит в восстановлении $P(x)$ по этой последовательности точек. Основное предположение, с которым будем иметь дело в настоящем разделе, заключается в следующем. В пространстве X задана система функций $\varphi_i(x)$ $(i = 1, 2, \ldots)$, удовлетворяющая условию ортонормированности:

$$\int_X \varphi_i(x)\varphi_k(x)\,dx = \delta_{ik}, \tag{2}$$

где δ_{ik} — символ Кронекера, и такая, что восстанавливаемая функция $P(x)$ представима конечным числом N членов в разложении по этой системе

$$P(x) = \sum_{i=1}^{N} c_i\varphi_i(x). \tag{3}$$

В качестве потенциальной функции возьмем функцию двух переменных *

$$K(x, y) = \sum_{i=1}^{N} \varphi_i(x)\varphi_i(y). \tag{4}$$

В дальнейшем для простоты предполагается, что функции $\varphi_i(x)$ ограничены, так что ограниченными будут и функции $P(x)$ и $K(x, y)$. После появления k точек $x^1, x^2, \ldots, x^k$ составляется функция

$$\Psi_k(x) = \frac{1}{k}\sum_{s=1}^{k} K(x, x^s). \tag{5}$$

Приводимые ниже теоремы I и II показывают, что $\Psi_k(x)$ с ростом k в определенном смысле стремится к функции $P(x)$.

Теорема I. В каждой фиксированной точке $x \in X$ при $k \to \infty$ $\Psi_k(x)$ стремится по вероятности ** к $P(x)$.

Доказательство. Последовательность точек $x^1, x^2, \ldots, x^k$ есть последовательность независимых случайных одинаково распределенных величин. Для каждой фиксированной точки x числа $K(x, x^1)$, $K(x, x^2), \ldots$ — также одинаково распределенные независимые случайные величины. Их среднее

* Если функции $\varphi_i(x)$ комплексные, то условие ортонормированности (2) принимает вид

$$\int_X \varphi_i(x)\,\overline{\varphi_k(x)}\,dx = \delta_{ik},$$

а потенциал функции (4)

$$K(x, y) = \sum_{i=1}^{N} \varphi_i(x)\,\overline{\varphi_i(y)}.$$

** Говорят, что случайная величина α_k при $k \to \infty$ стремится по вероятности к α и пишут $\alpha_k \xrightarrow{p} \alpha$, если для любой пары чисел $\varepsilon > 0$, $\delta > 0$ существует такое K, что при $k < K$ вероятность неравенства $|\alpha_k - \alpha| < \varepsilon$ больше $1 - \delta$.

арифметическое, как известно, стремится по вероятности при $k \to \infty$ к математическому ожиданию *. Поэтому

$$\Psi_k(x) \xrightarrow{p} \int K(x, \xi) P(\xi) d\xi, \quad k \to \infty \tag{6}$$

Учитывая (3) и (4), вычислим интеграл в правой части выражения (6)

$$\int K(x, \xi) P(\xi) d\xi = \int K(x, \xi) \sum_{i=1}^{N} c_i \varphi_i(\xi) d\xi =$$

$$= \int \sum_{i,j=1}^{N} \varphi_j(x) \varphi_j(\xi) c_i \varphi_i(\xi) d\xi = \sum_{j=1}^{N} \varphi_j(x) \sum_{i=1}^{N} c_i \int_X \varphi_j(\xi) \varphi_i(\xi) d\xi = P(x).$$

Поэтому в силу (2)

$$\int K(x, \xi | p(\xi) d\xi = P(x), \tag{7}$$

что в силу (6) и доказывает теорему.

Предел в выражении (6) определяет лишь, что $\Psi_k(x)$ сходится по вероятности к $P(x)$ в каждой фиксированной точке. Поэтому не исключено, что при любом фиксированном k найдутся точки x, в которых $\Psi_k(x)$ с большой вероятностью сколько угодно сильно отличается от $P(x)$. Однако имеет место следующая теорема.

Теорема II. Функция $\Psi_k(x)$ с ростом числа показанных точек сходится по вероятности в среднем к $P(x)$, т. е.

$$\int (\Psi_k(x) - P(x))^2 dx \xrightarrow{p} 0, \quad k \to \infty. \tag{8}$$

Доказательство теоремы II приведено в Приложении. В заключение раздела сделаем два замечания.

Замечание 1. В тех случаях, когда функция распределения представима интегралом

$$P(x) = \int_\Omega c_\omega \varphi_\omega(x) d\omega,$$

изложенный метод восстановления $P(x)$ применим, если потенциальную функцию задать в виде

$$K(x, y) = \int_\Omega \varphi_\omega(x) \varphi_\omega(y) d\omega.$$

Доказательства теорем I и II в этом случае проводятся совершенно аналогично, если предположить ограниченность $P(x)$ и $K(x, y)$.

Замечание 2. Легко показать, что если функция $P(x)$ представима рядом

$$P(x) = \sum_{i=1}^{\infty} c_i \psi_i(x),$$

где не все C_i обращаются в нуль при $i > N$, а потенциальная функция по-прежнему выбирается в виде (4), то $\psi_k(x)$ при $k \to \infty$ стремится (в смысле теорем I и II) к

$$P^N(x) = \sum_{i=1}^{N} c_i \varphi_i(x).$$

3. Алгоритм восстановления степени достоверности

Как указывалось в разделе 1, всюду в дальнейшем будем предполагать, что степени достоверности $D_A(x)$ и $D_B(x)$ представимы конечным разложением в ряд по ортонормированной системе функций $\varphi_i(x)$. Так

* В силу ограниченности $K(x, y)$ дисперсия случайной величины существует и, следовательно, можно воспользоваться законом больших чисел. Это обстоятельство без специальных оговорок используется и в дальнейшем.

как $D_B(x) = 1 - D_A(x)$, то существен лишь вид функции $D_A(x)$, которую для краткости будем обозначать просто через $D(x)$, так что

$$D_A(x) \equiv D(x) = \sum_{i=1}^{N} c_i \varphi_i(x). \tag{9}$$

Переходя к описанию алгоритма восстановления функции $D(x)$ методом потенциальных функций, введем потенциальную функцию вида

$$K(x, y) = \sum_{i=1}^{N} \varphi_i(x) \varphi_i(y). \tag{10}$$

В предлагаемом алгоритме фигурирует оператор «черта сверху», определенный следующим образом:

$$\overline{\Phi}(x) = \begin{cases} \Phi(x), & \text{если} \quad 0 \leqslant \Phi(x) \leqslant 1, \\ 0, & \text{если} \quad \Phi(x) < 0, \\ 1, & \text{если} \quad \Phi(x) > 1, \end{cases} \tag{11}$$

причем при использовании алгоритма выстраивается последовательность функций $\Phi_i(x)$ таких, что при $i \to \infty$ функция $\overline{\Phi_i}(x)$ в некотором смысле сходится к $D(x)$.

В процессе обучения последовательно показываются точки x^i ($i = 1, 2, \ldots$) и при каждом, например $(i+1)$-м, показе алгоритм служит для построения функций $\Phi_{i+1}(x)$ и $\overline{\Phi_{i+1}}(x)$ по следующим данным: по функциям $\Phi_i(x)$ и $\overline{\Phi_i}(x)$, построенным на предыдущем i-м шаге, по показанной точке x^{i+1} и по имеющейся информации о том, к какому классу, A и B, эта точка принадлежит.

На первом шаге в качестве $\Phi_0(x)$ можно принять любую функцию, удовлетворяющую основному предположению (9), в частности можно положить $\Phi_0(x) = \overline{\Phi_0}(x) = 0$.

Особенность алгоритма состоит в использовании случайного акта («бросание монеты»). Именно, пусть к произвольному $(i+1)$-му шагу построены функции $\Phi_i(x)$ и $\overline{\Phi_i}(x)$ и показана $(i+1)$-я точка x^{i+1}. Тогда с вероятностью $\overline{\Phi_i}(x^{i+1})$ алгоритм относит точку x^{i+1} к классу A и с вероятностью $1 - \overline{\Phi_i}(x^{i+1})$ — к B. Так как при появлении в процессе обучения точки x^{i+1} сообщается, к какому классу ее отнес «учитель», то возникает один из следующих четырех случаев, условно обозначаемых через AA, AB, BA, BB. Здесь первая буква указывает, к какому классу отнес точку x^{i+1} «учитель», а вторая буква — к какому классу отнес эту точку алгоритм.

Далее, в зависимости от того, какой из этих случаев имеет место, строится функция $\Phi_{i+1}(x)$ (и соответственно $\overline{\Phi_{i+1}}(x)$) следующим образом:

$$\Phi_{i+1}(x) = \begin{cases} \Phi_i(x) & \text{в случае } AA, \\ \Phi_i(x) + \gamma_{i+1} K(x, x^{i+1}) & \text{в случае } AB, \\ \Phi_i(x) - \gamma_{i+1} K(x, x^{i+1}) & \text{в случае } BA, \\ \Phi_i(x) & \text{в случае } BB, \end{cases} \tag{12}$$

где γ_i — «весовые» множители, зависящие лишь от номера показа.

Иначе говоря, в случаях, когда «предположение» алгоритма о классе точки x^{i+1} оказалось верным, функция $\Phi_i(x)$, а значит и $\overline{\Phi_i}(x)$, не изменяются. В случае же, когда предположение неверно, $\Phi_i(x)$ изменяется путем добавления (случай AB) или отнятия (случай BA) величины потенциальной функции $K(x, y)$ при $y = x^{i+1}$ с весом γ_{i+1}.

В качестве множителей γ_{i+1} могут быть использованы члены любой последовательности положительных чисел $\gamma_1, \gamma_2, \ldots$, лишь бы она удовлетворяла условию: ряд $\sum_{i=1}^{\infty} \gamma_i$ расходится, а ряд $\sum_{i=1}^{\infty} \gamma_i^2$ сходится. Например, можно положить $\gamma_i = 1 / i$ либо $\gamma_i = 1 / i^{1/2+\varepsilon}$, где $\varepsilon > 0$, и т. д.

При следующем показе (точка x^{i+2}) процедура повторяется, но уже с новой функцией $\Phi_{i+1}(x)$, полученной в силу (12) на $(i+1)$-м шаге, и т. д.

Функция $\Phi_i(x)$ есть случайная функция, так как для ее построения используются два случайных процесса: случайно появляются точки x^i и, кроме того, случайно (в силу самого алгоритма) эти точки относятся к классу A или B. Далее доказывается, однако, что эта случайная функция с ростом i сходится к функции $D(x)$. Доказательству этого факта посвящен следующий раздел. Раньше, чем перейти к изложению этого доказательства, сделаем несколько замечаний о машинной реализации алгоритма.

Как следует из описания алгоритма, на каждом этапе его использования требуется «помнить» функцию $\Phi_i(x)$, заданную на всем пространстве X. Однако хранить в памяти ее таблицу, т. е. значения на достаточно частой сетке ее аргументов, практически нецелесообразно и, более того, часто просто невозможно. Поэтому вопрос о способе запоминания функции $\Phi_i(x)$ требует специального рассмотрения. Ниже рассмотрено два таких способа.

При реализации алгоритма, например на универсальной цифровой машине, удобно принять $\Phi_0(x) = \overline{\Phi}_0(x) = 0$. Далее, к концу любого, например i-го, шага в памяти машины достаточно хранить лишь три ряда символов:

коды точек	$x_1, x_2, \ldots, x_k$,
числа	$s_1, s_2, \ldots, s_k$,
числа	$\mu_1, \mu_2, \ldots, \mu_k$.

Здесь $x_1, x_2, \ldots, x_k$ — те из точек, появлявшихся на входе машины до i-го шага включительно, при появлении которых реализовались случаи AB или BA. Те же точки, для которых реализовались случаи AA или BB, не запоминаются. Нижний индекс в обозначении точки указывает на порядок поступления этой точки в память машины. Число s_j отмечает номер показа, в котором появилась x_j. Число $\mu_j = +1$, если при появлении x_j реализовался случай AB, и $\mu_j = -1$, если при появлении точки x_j реализовался случай BA.

При появлении на $(i+1)$-м шаге процесса обучения x^{i+1} машина подсчитывает число

$$\Phi_i(x^{i+1}) = \sum_{j=1}^{k} \mu_j \gamma_{s_j} K(x_j^{i+1} x_j) \tag{13}$$

и с вероятностью $\overline{\Phi}_i(x^{i+1})$ делает предположение, что точка x^{i+1} принадлежит классу A. Легко видеть, что число, подсчитанное в соответствии с (13), совпадает со значением функции $\Phi_i(x)$ в точке $x = x^{i+1}$, вычисляемой в силу рекуррентных соотношений (12).

Далее, в зависимости от реализовавшегося случая (AA, AB, BA, BB) машина либо не заносит в память код точки x^{i+1} (случаи AA и BB), либо запоминает код точки $x_{k+1} = x^{i+1}$, число $s_{k+1} = i + 1$ и число μ_{k+1} (равное $+1$, если реализовался случай AB, и равное -1, если реализовался случай BA); величины же $\Phi_i(x^{i+1})$ и $\overline{\Phi}_i(x^{i+1})$ «забываются».

В процессе экзамена при появлении нового символа x^* машина подсчитывает величину

$$\overline{\sum_{j=1}^{k} \mu_j \gamma_{s_j} K(x^*, x_j)},$$

где k — число точек, занесенных в память машины в процессе обучения, и выдает это число в качестве оценки $D_A(x^*)$.

Возможен и другой («персептронный» — см. [1]) путь реализации алгоритма, не требующий запоминания кодов точек x^i.

Так как для каждого i функция $\Phi_i(x)$ есть сумма взятых с некоторыми весами функций $K(x, x^s)$, представимых конечным рядом в разложении по системе $\varphi_k(x)$, то и сама функция $\Phi_i(x)$ может быть представлена в виде

$$\Phi_i(x) = \sum_{k=1}^{N} b_k^i \varphi_k(x). \tag{14}$$

Но так как функции $\varphi_k(x)$ заданы, то процесс построения $\Phi_i(x)$ сводится к процессу построения коэффициентов разложения b_k^i. Именно, в силу формулы (12) легко получить, что при появлении точки x^{i+1}

$$b_k^{i+1} = \begin{cases} b_k^i & \text{в случае } AA, \\ b_k^i + \gamma_{i+1}\varphi_k(x^{i+1}) & \text{в случае } AB, \\ b_k^i - \gamma_{i+1}\varphi_k(x^{i+1}) & \text{в случае } BA, \\ b_k^i & \text{в случае } BB. \end{cases} \tag{15}$$

Поэтому при такой реализации алгоритма в памяти машины к $(i+1)$-му шагу хранятся лишь числа

$$b_1^i, \; b_2^i, \; \ldots, \; b_N^i,$$

а запоминание точек x^i не требуется.

На $(i+1)$-м шаге при появлении точки x^{i+1} подсчитывается число

$$\overline{\sum_{k=1}^{N} b_k^i \varphi_k(x^{i+1})},$$

которое и определяет вероятность отнесения x^{i+1} к классу A.

В зависимости от возникающей ситуации (AA, AB, BA, BB) при подсчете b_k^{i+1} используется соответствующая строка формулы (15). Подсчитанные числа b_k^{i+1} и запоминаются машиной вместо b_k^i. При экзамене после r шагов обучения по появляющейся точке x^* подсчитывается число

$$\overline{\sum_{k=1}^{N} b_k^r \varphi_k(x^*)},$$

которое отождествляется с $D_A(x^*)$ и выдается на выход машины.

4. Доказательство сходимости алгоритма

Для доказательства сходимости алгоритма предположим, что в пространстве X существует функция $P(x)$ плотности вероятности * появления точки x, причем в процессе обучения точки x^1, x^2, … появляются не-

* Понятие плотности вероятности употребляется здесь лишь для простоты изложения. Доказательство теоремы о сходимости алгоритма остается в силе, если появление точек в X определяется вероятностной мерой $P(d\mu)$, где μ есть некоторая мера в X. Встречающиеся в тексте интегралы по x следует понимать тогда как интегралы по мере μ.

зависимо в соответствии с $P(x)$. Предположим также, что функции $\varphi_i(x)$ ограниченны. В этих предположениях справедлива следующая теорема.

Теорема III. Пусть степень достоверности $D(x)$ существует и удовлетворяет основному предположению

$$D(x) = \sum_{i=1}^{N} c_i \varphi_i(x).$$

Тогда в силу алгоритма при $i \to \infty$

$$\int [\overline{\Phi}_i(x) - D(x)]^2 P(x)\,dx \xrightarrow{p} 0. \qquad (16)$$

Теорема III устанавливает, таким образом, стремление в среднем по вероятности функции $\overline{\Phi}_i(x)$ к $D(x)$ в тех точках пространства X, где $P(x) > 0$. Если предположить, что вероятности появления точек в X совпадают в процессе обучения и в экзамене, то только такие точки нас и интересуют.

Доказательство теоремы III основано на следующей лемме.

Лемма 1. Пусть задан случайный процесс

$$y_1,\ y_2,\ \ldots,\ y_i,\ \ldots,$$

определяемый условиями вероятностями $P_{i+1}(y_{i+1} | y_1, y_2, \ldots, y_i)$ того, что на $(i+1)$-м шаге появляется величина y_{i+1} при условии, что на предыдущих i шагах появились величины $y_1, \ldots, y_i$.

Пусть заданы также две последовательности неотрицательных скалярных функций $\alpha_i(y_1, y_2, \ldots, y_i)$ и $\beta_i(y_1, y_2, \ldots, y_i)$ $(i = 1, 2, \ldots)$ таких, что для всех i имеют место неравенства

$$M\{\alpha_{i+1}(y_1, y_2, \ldots, y_{i+1}) | y_1, y_2, \ldots, y_i\} \leqslant \alpha_i(y_1, y_2, \ldots, y_i) - \\ - \gamma_i \beta_i(y_1, y_2, \ldots, y_i) + \delta_i, \qquad (17)$$

$$M\{\beta_{i+1}(y_1, y_2, \ldots, y_{i+1}) | y_1, y_2, \ldots, y_i\} \leqslant \beta_i(y_1, y_2, \ldots, y_i) + k\gamma_i, \qquad (18)$$

где M — символ условного математического ожидания, k — некоторая константа, γ_i и δ_i — неотрицательные числа такие, что ряд $\sum\limits_{i=1}^{\infty} \gamma_i$ расходится, а ряды $\sum\limits_{i=1}^{\infty} \gamma_i^2$ и $\sum\limits_{i=1}^{\infty} \delta_i$ сходятся.

Тогда при $i \to \infty$ случайная величина β_i стремится по вероятности к нулю.

Доказательство леммы 1 приведено в Приложении II. Оно опирается на лемму 2, изложенную в том же приложении.

Для того чтобы свести доказательство теоремы III к лемме 1, установим ряд свойств случайного процесса, определяемого алгоритмом.

Рассмотрим случайный процесс появления пар (x^1, ρ^1), (x^2, ρ^2), ..., (x^i, ρ^i), ..., где x^i — появившаяся при i-м показе точка пространства X, а ρ^i равно 0, если на i-м шаге реализовались случаи AA или BB, равно 1, если реализовался случай AB, и равно -1, если реализовался случай BA. Плотность вероятности события: «на $(i+1)$-м шаге появилась точка x^* и возникла ситуация AA (соответственно ситуации AB или BA, или BB)» равна

$$\begin{aligned}
&\text{для } AA\colon\ V_{AA}(x^*) = P(x^*) D(x^*) \overline{\Phi}_i(x^*),\\
&\text{для } AB\colon\ V_{AB}(x^*) = P(x^*) D(x^*) [1 - \overline{\Phi}_i(x^*)],\\
&\text{для } BA\colon\ V_{BA}(x^*) = P(x^*) [1 - D(x^*)] \Phi_i(x^*),\\
&\text{для } BB\colon\ V_{BB}(x^*) = P(x^*) [1 - D(x^*)]\ [1 - \overline{\Phi}_i(x^*)].
\end{aligned} \qquad (19)$$

В правых частях каждого из этих четырех равенств выписано произведение трех членов. Первый сомножитель $P(x^*)$ — плотность вероятности появления точки x^*, второй сомножитель $D(x^*)$ или $-D(x^*)$ — вероятность того, что точка x^* принадлежит классу A или соответственно B, третий сомножитель $\overline{\Phi}_i(x^*)$ или $-\overline{\Phi}_i(x^*)$ — вероятность того, что случайный шаг алгоритма отнес точку x^* к классу A или B. Все эти три события, соответствующие указанным трем сомножителям, независимы. Очевидно, $V_{AA}(x^*) + V_{BB}(x^*)$ есть плотность вероятности того, что $\rho^{i+1} = 0$, $x^{i+1} = x^*$. Аналогично $V_{AB}(x^*)$ — плотность вероятности события $\rho^{i+1} = 1$, $x^{i+1} = x^*$, а $V_{BA}(x^*)$ — плотность вероятности события $\rho^{i+1} = -1$, $x^{i+1} = x^*$.

Очевидно также, что

$$V_{AA}(x^*) + V_{AB}(x^*) + V_{BA}(x^*) + V_{BB}(x^*) = P(x^*). \tag{20}$$

Функция $\overline{\Phi}_i(x)$ однозначно определяется парами (x^1, ρ^1), (x^2, ρ^2), ..., (x^i, ρ^i), так что формула (19) однозначно определяет условную плотность вероятности пары (x^{i+1}, ρ^{i+1}) при условии, что на i предыдущих шагах реализовались пары (x^1, ρ^1), ..., (x^i, ρ^i).

Таким образом, случайный процесс, определяемый алгоритмом, имеет тот же характер, что и случайный процесс, рассмотренный в лемме, если только понимать под случайной переменной y_i пару (x^i, ρ^i).

Рассмотрим теперь последовательность случайных величин

$$I_i \equiv \int\limits_X [\Phi_i(x) - D(x)]^2 dx, \tag{21}$$

$$J_i \equiv \int\limits_X [\Phi_i(x) - D(x)]\,[\overline{\Phi}_i(x) - D(x)]\,P(x)\,dx, \tag{22}$$

являющихся неотрицательными функциями пар (x^1, ρ^1), ..., (x^i, ρ^i).

Неотрицательность интеграла (21) очевидна. Для того чтобы установить неотрицательность интеграла (22), рассмотрим два случая

$$\Phi_i(x) \geqslant D(x) \geqslant 0 \quad \text{и} \quad \Phi_i(x) < D(x) \leqslant 1.$$

В первом случае по определению оператора «черта сверху» имеет место соотношение $\Phi_i(x) \geqslant \overline{\Phi}_i(x)$, откуда следует неравенство

$$[\Phi_i(x) - D(x)]\,[\overline{\Phi}_i(x) - D(x)] \geqslant [\overline{\Phi}_i(x) - D(x)]^2 \geqslant 0.$$

Во втором $\Phi_i(x) \leqslant \overline{\Phi}_i(x)$, и выполнено это же неравенство. Поэтому в любом случае

$$J_i \geqslant \int\limits_X [\overline{\Phi}_i(x) - D(x)]^2 P(x)\,dx \geqslant 0. \tag{23}$$

В Приложении III показывается, что последовательности случайных величин I_i и J_i в силу алгоритма удовлетворяют неравенствам

$$M\{I_{i+1}|\Phi_i(x)\} \leqslant I_i - 2\gamma_{i+1}J_i + L\gamma_{i+1}^2. \tag{24}$$

$$M\{J_{i+1}|\Phi_i(x)\} \leqslant J_i + k\gamma_{i+1}, \tag{25}$$

где L и k — некоторые константы. Отождествляя теперь последовательности неотрицательных величин I_i и J_i с фигурирующими в лемме 1 последовательностями α_i и β_i соответственно, получаем, что в силу леммы 1 последовательность случайных величин J_i при $i \to \infty$ стремится по вероятности к нулю. Это в силу неравенства (23) и доказывает утверждение (16) теоремы III.

Заметим в заключение, что теорема III остается справедливой, если вместо предположения (9) относительно $D(x)$ выполнено предположение

$$D(x) = \int\limits_{\Omega} c_\omega \varphi_\omega(x)\, d\omega, \tag{26}$$

а потенциальная функция выбрана в виде

$$K(x, y) = \int\limits_{\Omega} \varphi_\omega(x)\, \varphi_\omega(y)\, d\omega \tag{27}$$

и удовлетворяет условию $|K(x, y)| < c$.

ПРИЛОЖЕНИЕ I

Доказательство теоремы II. Рассмотрим математическое ожидание

$$\overline{\int\limits_X [\Psi_k(x)\xi - P(x)]^2\, dx} = \int\limits_X \overline{[\Psi_k(x) - P(x)]^2}\ dx. \tag{I.1}$$

Подынтегральное выражение правой части (I.1) есть дисперсия случайной величины $\Psi_k(x)$, являющейся средним арифметическим k независимых одинаково распределенных величин $K(x, \xi)$). Поэтому для дисперсии $\Psi_k(x)$ получим

$$D\Psi_k(x) = \frac{1}{k} DK(x, \xi). \tag{I.2}$$

Но учитывая, что $\overline{K(x, \xi)} = P(x)$ (см. (7)), имеем

$$DK(x, \xi) = \int [K(x, \xi) - P(x)]^2\, P(\xi)\, d\xi.$$

Отсюда после интегрирования получим при помощи (I.2)

$$\overline{[\Psi_k(x) - P(x)]^2} = \frac{1}{k}\left[\int K^2(x, \xi)\, P(\xi)\, d\xi - P^2(x)\right]. \tag{I.3}$$

Из (I.1) тогда следует

$$\int \overline{[\Psi_k(x) - P(x)]^2\, dx} = \frac{1}{k}\left[\int K(\xi, \xi)\, P(\xi)\, d\xi - \int P^2(x)\, dx\right]. \tag{I.4}$$

Здесь учтено, что $\int K^2(x, \xi)\, dx = K(\xi, \xi)$. Поскольку интегралы в правой части (I.4) существуют $\left(K(\xi, \xi)\right.$ ограниченно, а $\left.\int P^2(x)\, dx = \sum\limits_1^N c_i^2\right)$, то из (I.4), следует, что при $k \to \infty$

$$\overline{\int [\Psi_k(x) - P(x)]^2\ dx} \to 0. \tag{I.5}$$

Но, как легко показать, из стремления к нулю последовательности математических ожиданий неотрицательных случайных величин следует стремление к нулю по вероятности этой последовательности. Поэтому из (I.5) вытекает, что $\int [\Psi_k(x) - P(-x)]^2 dx \xrightarrow{p} 0$ при $k \to \infty$. Теорема II доказана.

ПРИЛОЖЕНИЕ II

Лемма 2. Пусть ряд $J = \sum\limits_1^\infty \gamma_s M_s$ сходится, а числа M_s удовлетворяют неравенствам

$$M_s \geqslant 0, \tag{II.1}$$

$$M_{s+1} \leqslant (1 + A\gamma_s)M_s + B\gamma_s, \qquad A > 0,\ B > 0. \tag{II.2}$$

Тогда, если $\gamma_s > 0$ и ряд $\sum\limits_1^\infty \gamma_s$ расходится, а ряд $\sum\limits_1^\infty \gamma_s^2$ сходится, то при $s \to \infty$ предел M_s существует и равен нулю.

Доказательство леммы 2. Доказательство проводится «от противного». Предположим, что $\lim\limits_{s\to\infty} M_s$ не существует или существует, но не равен нулю. Тогда найдется такое ε и может быть выделена такая подпоследовательность

$$M_{s_1}, M_{s_2}, \ldots, M_{s_i}, \ldots \tag{II.3}$$

с возрастающими индексами $s_1, s_2, \ldots, s_i, \ldots$, что имеют место неравенства

$$M_{s_i} \geqslant \varepsilon, \tag{II.4}$$

$$1 - A\gamma_r > 0, \quad r \geqslant s_1, \tag{II.5}$$

$$\sum_{s_i}^{s_{i+1}-1} \gamma_r > \frac{\varepsilon}{k}, \qquad k = B + A\varepsilon. \tag{II.6}$$

Действительно, возможность выбрать подпоследовательность M_{s_i} так, чтобы удовлетворялось неравенство (II.3), вытекает из определения предела; возможность удовлетворить (II.5) следует из того, что $\gamma_s \to 0$ в силу сходимости ряда $\sum\limits_1^\infty \gamma_s^2$; возможность же удовлетворить неравенство (II.6) вытекает из расходимости ряда $\sum\limits_1^\infty \gamma_s$. При этом легко видеть, что все требования (II.4) — (II.6) могут быть выполнены одновременно.

Из (II.2) получаем, что

$$M_{s-1} \geqslant \frac{M_s}{1 + A\gamma_{s-1}} - \frac{B\gamma_{s-1}}{1 + A\gamma_{s-1}}$$

или, усиливая это неравенство,

$$M_{s-1} \geqslant M_s(1 - A\gamma_{s-1}) - B\gamma_{s-1} \tag{II.7}$$

(здесь учтено, что $1 - A\gamma_{s-1} \leqslant 1 / (1 + A\gamma_{s-1}) \leqslant 1$). Выразим при помощи (II.7) M_s (для «текущего» s, $s_i \leqslant s < s_{i+1}$) через $M_{s_{i+1}}$, последовательно подставляя неравенства для M_r в неравенства для M_{r-1} ($r = s_{i+1} - 1, s_{i+1} - 2, \ldots, s$) *. Получим

$$M_s \geqslant M_{s_{i+1}} \prod_{r=s}^{s_{i+1}-1} (1 - A\gamma_r) - B\left\{\gamma_s + \sum_{r=s+1}^{s_{i+1}-1} \prod_{k=s}^{r-1} (1 - A\gamma_k)\gamma_r\right\}. \tag{II.8}$$

Учитывая (II.4), а также тот факт, что $1 - A\gamma_k < 1$, усилим (II.8)

$$M_s \geqslant \varepsilon \prod_{r=s}^{s_{i+1}-1} (1 - A\gamma_r) - B \sum_{s}^{s_{i+1}-1} \gamma_r \quad (s_i \leqslant s < s_{i+1}). \tag{II.9}$$

Легко доказывается справедливость неравенства

$$\prod_{r=s}^{s_{i+1}-1} (1 - A\gamma_r) \geqslant 1 - A \sum_{r=s}^{s_{i+1}-1} \gamma_r. \tag{II.10}$$

Действительно, проведем индукцию по числу членов l в произведении $\prod_l = \prod\limits_1^l (1 - \eta_k)$ и сумме $\sum\limits_l = \sum\limits_1^l \eta_k$, где η_k — произвольные неотрицательные числа, меньшие единицы. Если неравенство

$$\prod_l \geqslant 1 - \sum_l \tag{II.11}$$

верно для какого-либо l, то оно верно и при $l + 1$. В самом деле, умножая (II.11) на $1 - \eta_{l+1}$ (что не изменяет смысла неравенства, поскольку $1 - \eta_k > 0$), получим

$$\prod_{l+1} \equiv (1 - \eta_{l+1}) \prod_l \geqslant 1 - \eta_{l+1} - (1 - \eta_{l+1})\sum_l \geqslant 1 - \sum_{l+1}$$

Но при $l = 1$

$$\prod_1 = 1 - \sum_1,$$

что завершает доказательство (II.11). Из (II.11) следует (II.10), если учесть неотрицательность $A\gamma_r$ и условие (II.5).

* При такой подстановке смысл неравенств не изменяется в силу (II.5).

Подставляя (II.10) в (II.9), получим

$$M_s \geqslant \varepsilon - K \sum_{s}^{s_{i+1}-1} \gamma_r \quad (s_i \leqslant s < s_{i+1}), \tag{II.12}$$

где $K = B + \varepsilon A$.

Заметим, что, начиная с некоторого достаточно большого i (именно такого, что $\gamma_{s_{j+1}-1} < \varepsilon / K$ при $j \geqslant i$)* в силу (II.6) найдется число $s^* (c_i \leqslant s^* < s_{i+1})$ такое, что правая часть неравенства (II.12) положительна при $s \geqslant s^*$ и отрицательна при $s < s^*$.

Учитывая, кроме того, неравенство (II.1), можно написать

$$M_s \geqslant \begin{cases} 0 & \text{при } s_i \leqslant s \leqslant s^* - 1, \\ \varepsilon - K \sum\limits_{s}^{s_{i+1}-1} \gamma_r & \text{при } s^* \leqslant s \leqslant s_{i+1} - 1. \end{cases} \tag{II.13}$$

Покажем теперь, что из оценки (II.13) следует расходимость ряда $J = \sum\limits_1^{\infty} \gamma_s M_s$, что противоречит условию леммы, и тем самым завершим ее доказательство.

Представим ряд $J = \sum\limits_1^{\infty} \gamma_s M_s$ в виде

$$J = \sum_{i=1}^{\infty} J_i, \quad J_i = \sum_{s_i}^{s_{i+1}-1} \gamma_s M_s \tag{II.14}$$

и докажем расходимость этого ряда, установив, что суммы J_i при достаточно больших i превышают некоторую положительную константу, не зависящую от i.

Действительно, используя оценку (II.13), получаем неравенство

$$J_i \geqslant \sum_{s=s^*}^{s_{i+1}-1} \gamma_s \left(\varepsilon - K \sum_{r=s}^{s_{i+1}-1} \gamma_s \right) = \varepsilon \sum_{s=s^*}^{s_{i+1}-1} \gamma_s - K \sum_{s=s^*}^{s_{i+1}-1} \gamma_s \sum_{r=s}^{s_{i+1}-1} \gamma_r . \tag{II.15}$$

Преобразуем теперь второй член в правой части неравенства (II.15), используя легко доказываемое тождество

$$\sum_{s=a}^{b} \gamma_s \sum_{r=s}^{b} \gamma_r = \frac{\sum\limits_{s=a}^{b} \gamma_s^2 + \left(\sum\limits_{s=a}^{b} \gamma_s\right)^2}{2}. \tag{II.16}$$

Получаем

$$J_i \geqslant \varepsilon \sum_{s=s^*}^{s_{i+1}-1} \gamma_s - \frac{K}{2} \left(\sum_{s=s^*}^{s_{i+1}-1} \gamma_s \right)^2 - \frac{K}{2} \sum_{s=s^*}^{s_{i+1}-1} \gamma_s^2. \tag{II.17}$$

Но в силу определения числа s^* имеют место неравенства

$$\sum_{r=s^*}^{s_{i+1}-1} \gamma_r \leqslant \frac{\varepsilon}{K}, \quad \sum_{r=s^*-1}^{s_{i+1}-1} \gamma_r > \frac{\varepsilon}{K}$$

и поэтому

$$\frac{\varepsilon}{K} - \gamma_{s^*-1} < \sum_{r=s^*}^{s_{i+1}-1} \gamma_r \leqslant \frac{\varepsilon}{K} . \tag{II.18}$$

---

* Такое i существует, так как $\gamma_s \to 0$ в силу сходимости ряда $\sum\limits_1^{\infty} \gamma^2$.

Учитывая (II.18), преобразуем неравенство (II.17) к виду

$$J_i \geqslant \varepsilon\left(\frac{\varepsilon}{K} - \gamma_{s^*-1}\right) - \frac{K}{2}\left(\frac{\varepsilon}{K}\right)^2 - \frac{K}{2}\sum_{s^*}^{s_{i+1}-1} \gamma_s^2$$

или

$$J_i \geqslant \frac{\varepsilon^2}{K} - \varepsilon\gamma_{s^*-1} - \frac{K}{2}\sum_{s^*}^{s_{i+1}-1} \gamma_s^2. \qquad \text{(II.19)}$$

Выбирая достаточно большое i (а значит и s^*, так как $s^* \geqslant s_i$), можно сделать произвольно малыми два последних члена в правой части неравенства (II.19), так как по условию ряд $\sum\limits_1^\infty \gamma_s^2$ сходится и $\gamma_s \to 0$.

Поэтому при достаточно больших i величина J_i больше некоторой константы, откуда в силу (II.14) сразу следует расходимость ряда J. Таким образом, предположение, что предел $\lim\limits_{s\to\infty} M_s$ не существует или существует, но не равен нулю, противоречит условию леммы. Лемма 2 доказана.

Доказательство леммы 1. Докажем сначала, что

$$\lim_{i\to\infty} M\,\{\beta_i\,(y_1,\ y_2,\ \ldots,y_i)\} = 0. \qquad \text{(II.20)}$$

С этой целью оценим математическое ожидание величины $\alpha_{j+1} \equiv \alpha_{j+1}(y_1,\ y_2,\ \ldots,\ y_{j+1})$. Очевидно, что

$$M\,\{\alpha_{j+1}\} = \int M\,\{\alpha_{j+1} \mid y_1,\ y_2,\ \ldots,y_j\}\, P_j\,(y_1,\ y_2,\ \ldots,y_j)\, dy_1, \ldots, dy_j, \qquad \text{(II.21)}$$

где $P_j\,(y_1,\ y_2,\ \ldots,y_j)$ — совместная плотность вероятности события $(y_1,\ \ldots,y_j)$.

Используя (II.21) и условие (17) леммы 1, имеем

$$M\,\{\alpha_{j+1}\} \leqslant \int \alpha_j P_j\,(y_1, \ldots, y_j)\, dy_1, \ldots, dy_j - \gamma_j \int \beta_j P_j\,(y_1, \ldots, y_j)\, dy_1, \ldots, dy_j +$$

$$+ \delta_j \int P_j\,(y_1, \ldots, y_j)\, dy_1, \ldots, dy_j,$$

т. е. получим

$$M\,\{\alpha_{j+1}\} \leqslant M\,\{\alpha_j\} - \gamma_j M\,\{\beta_j\} + \delta_j.$$

Суммируя это неравенство в пределах от $j = 1$ до $i - 1$, найдем

$$M\,\{\alpha_i\} \leqslant M\,\{\alpha_1\} - \sum_{j=1}^{i-1} \gamma_j M\,\{\beta_j\} + \sum_{j=1}^{i-1} \delta_j,$$

или учитывая, что $M\,\{\alpha_i\} \geqslant 0$ и усиливая неравенство, имеем

$$\sum_{j=1}^{i-1} \gamma_j M\,\{\beta_j\} \leqslant M\,\{\alpha_1\} + \sum_{j=1}^{i-1} \delta_j. \qquad \text{(II.22)}$$

Последовательность $S_i = \sum\limits_1^{i-1} \gamma_j M\,\{\beta_j\}$ неубывающая, так как $\gamma_j \geqslant 0$ и $M\{\beta_j\} \geqslant 0$, и, в силу неравенства (II.22), ограниченная $\left(\text{ряд } \sum\limits_1^\infty \delta_j \text{ сходится по условию леммы}\right)$. Поэтому последовательность S^i имеет предел и ряд

$$\sum_1^\infty \gamma_j M\,\{\beta_j\} < \infty. \qquad \text{(II.23)}$$

Из условия (18) леммы 1 следует, что

$$M\,\{\beta_{j+1}\} = \int M\,\{\beta_{j+1} \mid y_1, \ldots, y_j\}\, P_j\,(y_1, \ldots, y_j)\, dy_1, \ldots, dy_j \leqslant$$

$$\leqslant \int \beta_j P\,(y_1, \ldots, y_j) dy_1, \ldots, dy_j + K\gamma_j \int P_j\,(y_1, \ldots, y_j)\, dy_1, \ldots, dy_j,$$

т. е.

$$M\{\beta_{j+1}\} \leqslant M\{\beta_j\} + K\gamma_j. \qquad (II.24)$$

Если положить $M_s = M\{\beta_s\}$, то неравенства (II.23) и (II.24) составляют в совокупности условия леммы 2 и в силу этой леммы (II.20) доказаны.

Но из стремления к нулю последовательности математических ожиданий неотрицательных случайных величин следует стремление к нулю по вероятности этой последовательности. Поэтому из (II.20) и следует, что $\beta_i \xrightarrow{p} 0$ при $i \to \infty$. Лемма 1 доказана.

ПРИЛОЖЕНИЕ III

Доказательство неравенства (24). Пусть после i-го шага алгоритма построена функция $\Phi_i(x)$. Рассмотрим $(i+1)$-й шаг алгоритма и оценим условное математическое ожидание интеграла

$$I_{i+1} = \int [\Phi_{i+1}(x) - D(x)]^2 \, dx.$$

Используя правило (12) формирования $\Phi_{i+1}(x)$ в зависимости от $\Phi_i(x)$ в силу алгоритма и вспоминая определение плотностей вероятностей $V_{AA}(z)$, $V_{AB}(z)$, $V_{BA}(z)$, $V_{BB}(z)$ (см. 19)), получаем

$$M\{I_{i+1} \mid \Phi_i(x)\} = \int \left\{ \int [\Phi_i(x) - D(x)]^2 \, dx \right\} [V_{AA}(z) + V_{BB}(z)] \, dz + \\ + \int \left\{ \int [\Phi_i(x) + \gamma_{i+1} K(x, z) - D(x)]^2 \, dx \right\} V_{AB}(z) \, dz + \\ + \int \left\{ \int [\Phi_i(x) - \gamma_{i+1} K(x, z) - D(x)]^2 \, dx \right\} V_{BA}(z) \, dz.$$

Собирая отдельно члены, содержащие $K(x, z)$ в нулевой, первой и второй степени, получаем

$$M\{I_{i+1} \mid \Phi_i(x)\} = R_1 + R_2 + R_3, \qquad (III.1)$$

где

$$R_1 = \int \left\{ \int [\Phi_i(x) - D(x)]^2 \, dx \right\} [V_{AA}(z) + V_{AB}(z) + V_{BA}(z) + V_{BB}(z)] \, dz,$$

$$R_2 = 2\gamma_{i+1} \int \left\{ \int K(x, z) [\Phi_i(x) - D(x)] \, dx \right\} [V_{AB}(z) - V_{BA}(z)] \, dz,$$

$$R_3 = \gamma_{i+1}^2 \int \left\{ \int K^2(x, z) \, dx \right\} [V_{AB}(z) + V_{BA}(z)] \, dz.$$

Рассмотрим каждый из этих интегралов порознь. В соответствии с равенством (20) и учитывая, что $\int P(z)\, dz = 1$, имеем

$$R_1 = I_i. \qquad (III.2)$$

В интеграле R_2 выполним интегрирование по x. При этом примем во внимание следующие обстоятельства:

а) в силу условий (9) и (10) имеем

$$\int K(x, z) D(x) \, dx = D(z),$$

б) в силу алгоритма [условие (12) и вид потенциальной функции (10)] $\Phi_i(x)$ также разлагается в конечный ряд, содержащий N функций $\varphi_1(x), \ldots, \varphi_N(x)$, поэтому

$$\int K(x, z) \Phi_i(x) \, dx = \Phi_i(z),$$

в) в силу (19)

$$V_{AB}(z) - V_{BA}(z) = [D(z) - \overline{\Phi_i}(z)] P(z).$$

В результате получим

$$R_2 = 2\gamma_{i+1} \int [\Phi_i(z) - D(z)] [D(z) - \overline{\Phi}_i(z)] P(z) \, dz. \qquad (III.3)$$

Рассмотрим теперь интеграл R_3. Для этого выполним интегрирование по x, учитывая (10)

$$\int K^2(x, z) \, dx = K(z, z).$$

Используя это соотношение и очевидное неравенство

$$V_{AB}(z)+V_{BA}(z)\leqslant P(z),$$

получаем оценку

$$R_3\leqslant\gamma_{i+1}^2\int K(z,\ z)\,P(z)\,dz.$$

Полагая *

$$L=\int K(z,\ z)\,P(z)\,dz>0,$$

получим

$$R_3\leqslant\gamma^2{}_{i+1}L. \tag{III.4}$$

Возвращаясь к равенству (III.1) и используя полученные значения интегралов (III.2) и (III.3) и оценку (III.4), находим

$$M\{I_{i+1}\mid\Phi_i(x)\}\leqslant I_i-2\gamma_{i+1}\int[\Phi_i(z)-D(z)]\ [\overline{\Phi}_i(z)-D(z)]\,P(z)\,dz+\gamma_{i+1}^2L,$$

т. е. убеждаемся, что неравенство (24) выполняется.

Доказательство неравенства (25). Рассмотрим условное математическое ожидание величины

$$J_{i+1}=\int[\Phi_{i+1}(x)-D(x)]\ [\overline{\Phi}_{i+1}(x)-D(x)]\,P(x)\,dx$$

при условии, что на i-м шаге построена функция $\Phi_i(x)$

$$M\{J_{i+1}\mid\Phi_i(x)\}=\int\left\{\int[\Phi_i(x)-D(x)]\ [\overline{\Phi}_i(x)-D(x)]\,P(x)\,dx\right\}\times$$
$$\times[V_{AA}(z)+V_{BB}(z)]\,dz+\int\left\{\int[\Phi_i(x)+\gamma_{i+1}K(x,\ z)-D(x)]\times\right.$$
$$\left.\times\,[\overline{\Phi_i(x)+\gamma_{i+1}K(x,\ z)}-D(x)]\,P(x)\,dx\right\}V_{AB}(z)\,dz+\int\left\{\int[\Phi(x)-\right.$$
$$\left.-\gamma_{i+1}K(x,\ z)-D(x)]\,[\overline{\Phi_i(x)-\gamma_{i+1}K(x,z)}-D(x)]\,P(x)\,dx\right\}V_{BA}(z)\,dz.$$

После перегруппировки членов получаем

$$M\{J_{i+1}\mid\Phi_i(x)\}=S_1+S_2+S_3+S_4+S_5, \tag{III.5}$$

где

$$S_1=\int\left\{\int[\Phi_i(x)-D(x)]\ [\overline{\Phi}_i(x)-D(x)]\,P(x)\,dx\right\}[V_{AA}(z)+$$
$$+V_{AB}(z)+V_{BA}(z)+V_{BB}(z)]\,dt,$$

$$S_2=\gamma_{i+1}\int\left\{\int K(x,\ z)\,[\overline{\Phi_i(x)+\gamma_{i+1}K(x,z)}-D(x)]\,P(x)\,dx\right\}V_{AB}(z)\,dz.$$
$$S_3=-\gamma_{i+1}\int\left\{\int K(x,\ z)\,[\overline{\Phi_i(x)-\gamma_{i+1}K(x,z)}-D(x)]\,P(x)\,dx\right\}V_{BA}(z)\,dz,$$
$$S_4=\int\left\{\int[\Phi_i(x)-D(x)]\,[\overline{\Phi_i(x)+\gamma_{i+1}K(x,z)}-\overline{\Phi}_i(x)]\,P(x)\,dx\right\}V_{AB}(z)\,dz,$$
$$S_5=\int\left\{\int[\Phi_i(x)-D(x)]\,[\overline{\Phi_i(x)-\gamma_{i+1}K(x,z)}-\overline{\Phi}_i(x)]\,P(x)\,dx\right\}V_{BA}(z)\,dz.$$

Используя формулу (20) и выполнив в S_1 интегрирование по z, получаем

$$S_1=J_i. \tag{III.6}$$

Интегралы S_2 и S_3 оценим, принимая во внимание, что

$$|\,\overline{\Phi_i(x)\pm\gamma_{i+1}K(x,\ z)}-D(x)\,|\leqslant1,$$
$$V_{AB}(z)\leqslant P(z),\qquad V_{BA}(z)\leqslant P(z), \tag{III.7}$$

Поэтому, например,

$$S_2\leqslant\gamma_{i+1}\int\left\{\int|\,K(x,\ z)\,|\ P(x)\ dx\right\}P(z)\,dz.$$

---

* В соответствии с (10) $K(z,\ z)>0$. Существование интеграла гарантируется ограниченностью $K(z,z)$, вытекающей из предположения об ограниченности $\varphi_i(x)$.

В силу предполагаемой ограниченности $\varphi_i(x)$, функция $K(x, z)$ также ограничена [в соответствии с (10)]. Отсюда следует существование такой константы C, что

$$\int |K(x, z)| P(x)\, dx \leqslant C$$

и поэтому

$$S_2 \leqslant C\gamma_{i+1}. \tag{III.8}$$

Аналогично оцениваем

$$S_3 \leqslant C\gamma_{i+1}. \tag{III.9}$$

Для оценки интегралов S_4 и S_5 используем неравенство

$$|F(u)(\overline{u+a} - \bar{u})| \leqslant a \max |F(u)|, \qquad -|a| \leqslant u \leqslant 1 + |a|, \tag{III.10}$$

которое следует из определения оператора «черта сверху», а именно из того факта что $|\overline{u+a} - \bar{u}| \leqslant a$ при $-|a| \leqslant u \leqslant 1 + |a|$ и $|\overline{u+a} - \bar{u}| = 0$ при u вне этого диапазона. Применяя (III.10) к подынтегральному выражению в S_4 и S_5 и учитывая ограниченность $K(x, z)$, находим

$$[\Phi_i(x) - D(x)][\overline{\Phi_i(x) \pm \gamma_{i+1} K(x, z)} - \overline{\Phi}_i(x)] \leqslant \gamma_{i+1} |K(x, z)| \times$$
$$\times [1 + \gamma_{i+1} |K(x, z)| - D(x)] \leqslant B\gamma_{i+1},$$

где B — некоторая положительная константа. Используя последнее неравенство и (III.7), получаем для S_4

$$S_4 \leqslant B\gamma_{i+1}. \tag{III.11}$$

Аналогично оцениваем

$$S_5 \leqslant B\gamma_{i+1}. \tag{III.12}$$

Возвращаясь к (III.5) и используя (III.6), (III.8), (III.9), (III.11) и (III.12), имеем оценку

$$M\{J_{i+1} \mid \Phi_i(x)\} \leqslant J_i + K\gamma_{i+1},$$

где $K = 2(B + C)$, которая совпадает с неравенством (25).

Поступила в редакцию
13 февраля 1964 г.

Цитированная литература

1. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. Автоматика и телемеханика, т. XXV, № 6, 1964.
2. Фролов А. С., Ченцов Н. Н. Использование зависимых испытаний в методе Монте Карло для получения гладких кривых. Тр. VI Всесоюзн. совещ. по теории вероятностей и математической статистике. Госполитнаучиздат, Вильнюс 1962.
3. Ченцов Н. Н. Оценка неизвестной плотности распределения по наблюдениям. Докл. АН СССР, т. 147, № 1, 1962.
4. Кириллов И. Е. Метод измерения функций распределения вероятностей, использующий их разложение по ортогональным полиномам. Проблемы передачи информации, вып. 4, 1959.

PROBABILITY PROBLEM OF PATTERN RECOGNITION LEARNING AND POTENTIAL FUNCTIONS METHOD

M. A. AIZERMAN, É. M. BRAVERMAN, L. I. ROZONOÉR

A probability problem of pattern recognition learning is formulated. An algorithm is described which is based on the potential functions method and permits to obtain a function approximating the validity of the fact that the situation appearing at the automaton input belongs to this or that class. A theorem on the convergence (in the probabilistic sense) of the obtained function to this validity is proved.