Математические основы прогнозирования временных рядов

А. М. Миронов[1]

В статье излагаются основные понятия и методы прогнозирования временны́х рядов. Рассматриваются различные алгоритмы смешивающего прогнозирования, и приводятся оценки качества этих алгоритмов.

Ключевые слова: временные ряды, алгоритмы прогнозирования, смешивающие алгоритмы прогнозирования

Введение

Основным объектом исследования в данной статье являются алгоритмы прогнозирования временных рядов (называемые ниже просто алгоритмами прогнозирования), которые основаны на следующей идее: пусть заданы несколько алгоритмов прогнозирования $A_1, \ldots, A_N$, искомый алгоритм прогнозирования A (называемый смешивающим алгоритмом) должен использовать результаты работы алгоритмов $A_1, \ldots, A_N$, качество прогнозирования смешивающего алгоритма A (т.е. доля правильных прогнозов этого алгоритма) должно быть близким к качеству наилучшего из алгоритмов $A_1, \ldots, A_N$.

Цель настоящей работы заключается в систематизации изложения основных подходов к смешивающему прогнозированию. Содержание работы имеет следующий вид. Сначала рассматриваются простейшие алгоритмы смешивающего прогнозирования: алгоритм большинства, алгоритм взвешенного большинства, алгоритм оптимального распределения потерь. Далее рассматриваются более сложные алгоритмы смешивающего прогнозирования: алгоритм следования за возмущённым лидером, агрегирующий алгоритм Вовка. Кроме того, рассматривается алгоритм усиления классификаторов (бустинг), природа которого сходна природе смешивающих алгоритмов. Затем рассматриваются понятие прогнозной стратегии и примеры детерминированной и вероятностной прогнозных стратегий. Содержание статьи основано на материале из книги [1], в настоящем тексте представлены новые, более простые доказательства соответствующих теорем из [1].

[1] *Миронов Андрей Михайлович* — доцент каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: amironov66@gmail.com.

Mironov Andrew Mikhaylovich — associate professor, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

1. Задача прогнозирования временны́х рядов

Под **временны́м рядом** понимается последовательность $y = (y_1, y_2, \ldots)$ элементов некоторого множества Y, называемых **исходами**. Мы будем рассматривать случай, когда $Y = \{0, 1\}$ или $[0, 1]$ (вместо 0 м.б. -1).

Прогнозируемый временной ряд y может быть бесконечным или иметь конечную длину, которую мы будем обозначать символом T, т.е. во втором случае $y = (y_1, \ldots, y_T)$.

Задача прогнозирования временного ряда y заключается в построении **прогнозирующего алгоритма (ПА)** A, который на каждом шаге прогнозирования $t = 1, 2, \ldots$ выдаёт значение $\gamma_t \in Y$, называемое **прогнозом** временного ряда y на шаге t. После того, как A выдал γ_t, становится известным значение $y_t \in Y$ **исхода** в момент t.

На каждом шаге прогнозирования t, который выполняет ПА A, определена **потеря** $l_t \in [0, 1]$ ПА A, связанная с несовпадением прогноза γ_t и исхода y_t. Будем считать, что $l_t = 0$ тогда и только тогда, когда $\gamma_t = y_t$. Как правило, потеря l_t является значением некоторой **функции потерь** λ на паре (γ_t, y_t), например $l_t = [\![\gamma_t \neq y_t]\!]$. Напомним, что для каждого логического утверждения φ запись $[\![\varphi]\!]$ обозначает число 1, если φ истинно, и 0, если φ ложно. Если $y = (y_1, \ldots, y_T)$, то будем называть **кумулятивной потерей** ПА A величину $L_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} l_t$.

2. Смешивающее прогнозирование

Один из методов построения прогнозирующих алгоритмов заключается в следующем. Пусть имеется несколько ПА A_1, …, A_N. Для каждого $i = 1, \ldots, N$ в каждый момент времени $t = 1, \ldots, T$ ПА A_i выдаёт прогноз γ_t^i. Будем обозначать записью L_T^i кумулятивную потерю ПА A_i. Используя алгоритмы $A_1, \ldots, A_N$ можно построить ПА A, называемый **смешивающим** ПА. На каждом шаге прогнозирования t прогноз γ_t, выдаваемый алгоритмом A, определяется как некоторая функция от прогнозов $\gamma_t^1, \ldots, \gamma_t^N$, называемая **функцией смешивания**.

Алгоритмы $A_1, \ldots, A_N$, участвующие в определении смешивающего ПА A, будем называть **экспертами**. Экспертов можно сравнивать по качеству их прогнозов: эксперт A_i лучше эксперта A_j, если $L_T^i < L_T^j$. Будем обозначать экспертов просто их номерами $1, \ldots, N$, и множество экспертов $\{1, \ldots, N\}$ будем обозначать символом I.

Величина $R_T = L_T - \min_{i \in I} L_T^i$ называется **регретом** смешивающего ПА A. Данная величина выражает собой отличие кумулятивной потери ПА A от кумулятивной потери наилучшего эксперта. При построении смешивающих ПА функция смешивания должна выбираться так, чтобы регрет смешивающего ПА был как можно меньше.

3. Алгоритм большинства

В этом пункте излагается простейший смешивающий ПА, который может использоваться лишь в ситуации, когда среди экспертов из множества $I = \{1, \ldots, N\}$ существует эксперт i_0, в каждый момент времени выдающий правильный прогноз, т.е. такой, что $\forall\, t \geq 1 \quad \gamma_t^{i_0} = y_t$. Этот алгоритм называется **алгоритмом большинства (Majority Algorithm, MA)**. Прогнозы данного ПА определяются следующим образом:

$$\forall\, t \geq 1 \quad \gamma_t := [\![|\{i \in B_t \mid \gamma_t^i = 1\}| \geq \tfrac{|B_t|}{2}]\!], \tag{1}$$

где для каждого конечного множества X запись $|X|$ обозначает число элементов в X, и $\forall\, t \geq 1$ множество B_t определяется следующим образом:

$$B_t = \{i \in I \mid \forall\, t' = 1, \ldots, t-1 \;\; \gamma_{t'}^i = y_{t'}\}. \tag{2}$$

Будем говорить, что ПА A **делает ошибку** на шаге t, если $\gamma_t \neq y_t$.

Теорема 1.
МА делает не более $\log_2 N$ ошибок.

Доказательство.
Нетрудно видеть, что последовательность множеств (2) обладает свойством $B_1 \supseteq B_2 \supseteq \ldots$, и каждое из множеств B_t непусто, т.к. $i_0 \in B_t$.
Если МА делает ошибку на шаге t, т.е. $\gamma_t \neq y_t$, то

- либо $\gamma_t = 1$ и $y_t = 0$, в этом случае, согласно (1),

$$|\{i \in B_t \mid \gamma_t^i = 1\}| \geq \tfrac{|B_t|}{2}, \tag{3}$$

  согласно (2), $B_{t+1} = \{i \in B_t \mid \gamma_t^i = 0\}$, и из (3) следует $|B_{t+1}| \leq \frac{|B_t|}{2}$,

- либо $\gamma_t = 0$ и $y_t = 1$, в этом случае, согласно (1),

$$|\{i \in B_t \mid \gamma_t^i = 1\}| < \tfrac{|B_t|}{2}, \tag{4}$$

  согласно (2), $B_{t+1} = \{i \in B_t \mid \gamma_t^i = 1\}$, и из (4) следует $|B_{t+1}| < \frac{|B_t|}{2}$.

В обоих случаях $|B_{t+1}| \leq \frac{|B_t|}{2}$.
Пусть МА делает k ошибок, и t – момент, в который делается k–я ошибка, тогда, по установленному выше,

$$N \geq |B_1| \geq 2^k |B_{t+1}|. \tag{5}$$

Учитывая $|B_{t+1}| \geq 1$, из (5) получаем $N \geq 2^k$, или $k \leq \log_2 N$. ■

4. Алгоритм взвешенного большинства

В этом пункте излагается смешивающий алгоритм, называемый **алгоритмом взвешенного большинства (Weighted Majority Algorithm, WMA)**, впервые он был изложен в работе [2]. Данный алгоритм может использоваться в том случае, когда среди экспертов из $I = \{1, \ldots, N\}$ может не быть эксперта, выдающего в каждый момент времени правильный прогноз.

В каждый момент времени t данный алгоритм сопоставляет каждому эксперту $i \in I$ некоторое число $w_t^i \in [0, 1]$, называемое **весом** этого эксперта. В начальный момент $t = 1$ вес каждого эксперта равен 1. Считаем, что потери имеют вид $l_t = |\gamma_t - y_t|$, $l_t^i = |\gamma_t^i - y_t|$.

Прогнозы данного ПА и изменения весов определяются следующим образом: выбирается параметр $\varepsilon \in (0, 1)$, и

$$\forall\, t = 1, \ldots, T \quad \left\{ \begin{array}{l} \gamma_t := [\![\sum_{i:\gamma_t^i = 0} w_t^i \leq \sum_{i:\gamma_t^i = 1} w_t^i]\!] \\ w_{t+1}^i := w_t^i (1 - \varepsilon l_t^i). \end{array} \right. \tag{6}$$

Теорема 2.

Для кумулятивных потерь WMA верно неравенство

$$L_T \leq \tfrac{2}{1-\varepsilon} \min_{i \in I} L_T^i + \tfrac{2}{\varepsilon} \ln N \tag{7}$$

(т.е. WMA ошибается примерно не более чем в $\frac{2}{1-\varepsilon}$ раз, чем наилучший эксперт).

Доказательство.

Будем использовать следующие обозначения:

$$M = L_T, \quad m = \min_{i \in I} L_T^i, \quad |\vec{w}_t| = \sum_{i \in I} w_t^i.$$

Пусть i – номер наилучшего эксперта.

w_t^i корректируется $\leq m$ раз, поэтому

$$|\vec{w}_T| \geq w_T^i \geq (1 - \varepsilon)^m. \tag{8}$$

Нетрудно проверить (это делается так же, как в доказательстве предыдущей теоремы), что если WMA делает ошибку на шаге t, то

$$\sum_{i:\gamma_t^i \neq y_t} w_t^i \geq \sum_{i:\gamma_t^i = y_t} w_t^i. \tag{9}$$

Прибавив к обеим частям (9) слагаемое $\sum_{i:\gamma_t^i \neq y_t} w_t^i$, получаем

$$\sum_{i:\gamma_t^i \neq y_t} w_t^i \geq \tfrac{|\vec{w}_t|}{2}. \tag{10}$$

Из (10) и из определения w_{t+1}^i в (6) следует, что если WMA делает ошибку на шаге t, то

$$\begin{aligned}|\vec{w}_{t+1}| &= \sum_{i\in I} w_t^i(1-\varepsilon l_t^i) = \\ &= |\vec{w}_t| - \varepsilon \sum_{i:\gamma_t^i \neq y_t} w_t^i \leq |\vec{w}_t|(1-\tfrac{\varepsilon}{2}),\end{aligned}$$

т.е. если WMA делает ошибку на шаге t, то $\frac{|\vec{w}_{t+1}|}{|\vec{w}_t|} \leq 1 - \frac{\varepsilon}{2}$.

Из определения весов в (6) следует, что для каждого $t = 1, \ldots, T-1$ $\frac{|\vec{w}_{t+1}|}{|\vec{w}_t|} \leq 1$. Следовательно,

$$\frac{|\vec{w}_T|}{|\vec{w}_1|} = \prod_{t=1}^{T-1} \frac{|\vec{w}_{t+1}|}{|\vec{w}_t|} \leq (1-\tfrac{\varepsilon}{2})^M,$$

откуда, учитывая (8) и равенство $|\vec{w}_1| = N$, получаем неравенство

$$\frac{(1-\varepsilon)^m}{N} \leq (1-\tfrac{\varepsilon}{2})^M,$$

логарифмируя которое, и учитывая неравенство

$$\ln(1+x) \leq x \quad \text{при } x \in (-1,1) \tag{11}$$

получаем:

$$m\ln(1-\varepsilon) - \ln N \leq M\ln(1-\tfrac{\varepsilon}{2}) \leq -\tfrac{\varepsilon}{2}M$$

откуда следует неравенство

$$\tfrac{\varepsilon}{2}M \leq m\ln\tfrac{1}{1-\varepsilon} + \ln N. \tag{12}$$

Применяя (11) для $x = \frac{\varepsilon}{1-\varepsilon}$, получаем соотношения

$$\ln\tfrac{1}{1-\varepsilon} = \ln(1+\tfrac{\varepsilon}{1-\varepsilon}) \leq \tfrac{\varepsilon}{1-\varepsilon} \tag{13}$$

Из (12) и (13) следует неравенство

$$\tfrac{\varepsilon}{2}M \leq m\tfrac{\varepsilon}{1-\varepsilon} + \ln N,$$

которое эквивалентно доказываемому неравенству (7). ■

5. Алгоритм оптимального распределения потерь

В этом пункте рассматривается другая постановка задачи построения смешивающего ПА: как и выше, задано множество экспертов $I = \{1, \ldots, N\}$, но для каждого шага прогнозирования t вместо прогнозов экспертов γ_t^i известны лишь потери $l_t^i \in [0,1]$, которые несут эксперты на шаге t. Требуется построить ПА, кумулятивные потери которого были бы как можно ближе к кумулятивным потерям наилучшего из этих экспертов (т.е. к $\min_{i\in I} L_T^i$).

Будем использовать понятие **вероятностного распределения (ВР)** на множестве $I = \{1, \ldots, N\}$, которое представляет собой произвольный вектор $\vec{p} = (p^1, \ldots, p^N)$ неотрицательных действительных чисел, удовлетворяющих условию $\sum_{i \in I} p^i = 1$. Множество всех ВР на I будем обозначать записью $I^{\triangle}$. Вектор из $I^{\triangle}$, все компоненты которого совпадают (т.е. равны $\frac{1}{N}$) будем называть **равномерно распределённым (р.р.)**.

Также будем использовать следующее обозначение – если $\vec{w}$ – вектор неотрицательных действительных чисел вида $(w^1, \ldots, w^N)$, то $norm(\vec{w})$ – это ВР $(p^1, \ldots, p^N)$, где

$$\forall\, i \in I \;\; p^i = \frac{w^i}{|\vec{w}|}, \quad \text{где } |\vec{w}| = \sum_{i=1}^{N} w^i.$$

Предлагается следующее решение описанной выше задачи: на каждом шаге прогнозирования t определяется ВР $\vec{p}_t = (p_t^1, \ldots, p_t^N) \in I^{\triangle}$, и прогноз искомого алгоритма A на шаге t полагается равным прогнозу эксперта i, номер которого выбран из I случайным образом, в соответствии с распределением $\vec{p}_t$ (т.е. с вероятностью p_t^1 выбран эксперт 1, с вероятностью p_t^2 выбран эксперт 2, и т.д.). Потери ПА A в момент t совпадают с потерями выбранного эксперта i в момент t.

Нетрудно видеть, что математическое ожидание l_t потери ПА A в момент t совпадает со **скалярным произведением** $\langle \vec{p}_t, \vec{l}_t \rangle = \sum_{i \in I} p_t^i l_t^i$, где $\vec{l}_t = (l_t^1, \ldots, l_t^N)$.

$\forall\, t = 1, \ldots, T$ определяем $\vec{p}_t$ как $norm(\vec{w}_t)$, где

$$\begin{array}{l} \vec{w}_1 \text{ – р.р.} \\ \forall\, t = 1, \ldots, T-1, \forall\, i \in I \;\; w_{t+1}^i := w_t^i \beta^{l_t^i} \end{array} \tag{14}$$

где $\beta \in (0, 1)$ – параметр.

Описанный выше алгоритм называется **алгоритмом оптимального распределения потерь**, и обозначается записью $Hedge(\beta)$. Впервые он был изложен в [3].

Лемма 1.

Средняя кумулятивная потеря $L_T = \sum_{t=1}^{T} l_t$ данного алгоритма удовлетворяет неравенству

$$\ln |\vec{w}_{T+1}| \leq -(1-\beta) L_T. \tag{15}$$

Доказательство.

Докажем эквивалентное неравенство:

$$|\vec{w}_{T+1}| \leq e^{-(1-\beta) L_T}. \tag{16}$$

Согласно определению (14), $\forall\, t = 1, \ldots, T$

$$|\vec{w}_{t+1}| = \sum_{i\in I} w^i_{t+1} = \sum_{i\in I} w^i_t \beta^{l^i_t} \le \sum_{i\in I} w^i_t (1-(1-\beta)l^i_t) \tag{17}$$

(в (17) используем неравенство

$$\beta^{l^i_t} \le 1-(1-\beta)l^i_t, \tag{18}$$

которое следует из выпуклости функции $y = \beta^x$).

Правая часть (17) равна

$$|\vec{w}_t| - (1-\beta)\sum_{i\in I} w^i_t l^i_t = |\vec{w}_t|(1-(1-\beta)l_t) \tag{19}$$

Из неравенства $1 + x \le e^x$ $(\forall\, x \in \mathbb{R})$, где $\mathbb{R}$ обозначает множество действительных чисел, следует неравенство

$$1-(1-\beta)l_t \le e^{-(1-\beta)l_t} \tag{20}$$

Из (17), (19) и (20) следует, что $\forall\, t = 1, \ldots, T$

$$|\vec{w}_{t+1}| \le |\vec{w}_t| e^{-(1-\beta)l_t} \tag{21}$$

Перемножая неравенства (21) для $t = 1, \ldots, T$, производя сокращения, и учитывая $|\vec{w}_1| = 1$, получаем искомое неравенство (16). ■

Перепишем (15) в виде

$$L_T \le -\tfrac{1}{1-\beta} \ln |\vec{w}_{T+1}|. \tag{22}$$

$\forall\, i \in I$ из неравенства $w^i_{T+1} \le |\vec{w}_{T+1}|$ следует неравенство

$$-\tfrac{1}{1-\beta} \ln |\vec{w}_{T+1}| \le -\tfrac{1}{1-\beta} \ln w^i_{T+1} \tag{23}$$

Из (14) следует, что

$$w^i_{T+1} = w^i_1 \beta^{L^i_T} = \tfrac{1}{N} \beta^{L^i_T}. \tag{24}$$

Из (22), (23) и (24) следует, что

$$L_T \le -\tfrac{1}{1-\beta}(\ln \tfrac{1}{N} + L^i_T \ln \beta). \tag{25}$$

Поскольку $\forall\, i \in I$ верно (25), то получаем соотношение

$$L_T \le \tfrac{1}{1-\beta} \ln \tfrac{1}{\beta} \min_{i\in I} L^i_T + \tfrac{\ln N}{1-\beta}. \tag{26}$$

Неравенство (26) означает, что средние кумулятивные потери ПА $Hedge(\beta)$ не превосходят кумулятивных потерь наилучшего эксперта, умноженных на константу $\frac{1}{1-\beta}\ln\frac{1}{\beta}$, к которым добавлен регрет (т.е. ошибка обучения) $\frac{\ln N}{1-\beta}$.

Теорема 3.
Если в ПА $Hedge(\beta)$ значение параметра β равно $\frac{1}{1+\sqrt{\frac{2}{T/\ln N}}}$, то

$$L_T \le \min_{i\in I} L_T^i + \sqrt{2T\ln N} + \ln N. \tag{27}$$

Доказательство.
Сначала докажем утверждение: если действительные числа L, L', R, R' удовлетворяют неравенствам $L' > L \ge 0$, $R' \ge R > 0$, и $\beta = \frac{1}{1+\sqrt{\frac{2}{L'/R'}}}$, то

$$\frac{1}{1-\beta}\ln\frac{1}{\beta}L + \frac{1}{1-\beta}R \le L + \sqrt{2L'R'} + R. \tag{28}$$

Обозначим $\sigma = \sqrt{2R'/L'}$.
Имеет место неравенство

$$\ln\frac{1}{\beta} \le \frac{1-\beta^2}{2\beta} \tag{29}$$

т.к. производная функции $f(\beta) = \frac{1-\beta^2}{2\beta} - \ln\frac{1}{\beta}$ равна $-\frac{(\beta-1)^2}{2\beta^2}$, и поскольку $f(1) = 0$, то если бы неравенство (29) было бы неверно, т.е. $f(\beta) < 0$, то, по теореме Лагранжа, $\exists\, \beta' \in (0,1) : f'(\beta') > 0$, что невозможно (производная функции f отрицательна во всех точках интервала $(0,1)$).
Из (29) следует, что $\frac{1}{1-\beta}\ln\frac{1}{\beta} \le \frac{1+\beta}{2\beta}$, поэтому

$$\begin{aligned}&\text{левая часть (28)} \le \frac{1+\beta}{2\beta}L + \frac{1}{1-\beta}R = \\ &= \frac{1}{2}(1+\frac{1}{\beta})L + \frac{1}{1-\beta}R = L + \frac{1}{2}L\sigma + \frac{1}{1-\frac{1}{1+\sigma}}R \le \\ &\le L + \sqrt{\frac{L'R'}{2}} + R + \frac{R}{\sigma} \le \text{правая часть (28)}\ .\end{aligned}$$

(мы используем равенство $\frac{1}{1-\frac{1}{1+\sigma}} = 1 + \frac{1}{\sigma}$).

Рассматривая (28) для $L = \min_{i\in I} L_T^i$, $L' = T$, $R = R' = \ln N$, и учитывая (26), получаем (27). ■

6. Бустинг

В этом параграфе рассматривается задача построения сильных алгоритмов машинного обучения. Для ее описания приведем необходимые определения.

Пусть заданы множества X и Y, элементы которых называются **объектами** и **ответами** соответственно, как правило, $Y = \{0,1\}$. **Обучающей выборкой (ОВ)** будем называть совокупность S вида

$$S = \{(x^i, y^i, p^i) \mid i \in I\} \tag{30}$$

где $I = \{1, \ldots, N\}$, $\forall\, i \in I \;\; x^i \in X$, $y^i \in Y$, $(p^1, \ldots, p^N) \in I^{\triangle}$. Запись $|S|$ обозначает число компонентов в S (т.е. N).

Каждая тройка (x, y, p) из ОВ S интерпретируется как утверждение о том, что объекту x соответствует ответ y с мерой уверенности p.

Под **алгоритмом машинного обучения (АМО)** понимаем алгоритм, получающий на вход ОВ S, и выдающий функцию $h : X \to Y$, называемую **классификатором**. **Ошибка** классификатора h на ОВ S – это число

$$Err(h, S) = \textstyle\sum_{i \in I} p^i [\![h(x^i) \neq y^i]\!].$$

АМО называется

- **сильным**, если для каждой ОВ S и $\forall\, \varepsilon, \delta \in (0,1)$ он выдает с вероятностью $> 1 - \delta$ за время, полиномиально зависящее от $\frac{1}{\varepsilon}, \frac{1}{\delta}, |S|$, классификатор h, такой, что $Err(h, S) \leq \varepsilon$,

- **слабым**, если для каждой ОВ S $\exists\, \varepsilon \in (0, \frac{1}{2})$: $\forall\, \delta \in (0,1)$ верно то же свойство, что и для сильного АМО.

Ниже решается следующая задача: пусть имеется слабый АМО, требуется на базе него построить сильный АМО. Метод преобразования слабого АМО в сильный АМО называется **бустингом**. Излагаемый ниже бустинг называется **AdaBoost** (**адаптивное усиление**). Впервые он был изложен в [3]. Говоря неформально, в основе данного бустинга лежит выделение таких элементов ОВ S, на которых классификатор h, получаемый при помощи слабого АМО делает наибольшую ошибку, и коррекция h на именно этих элементах. Входными данными для алгоритма AdaBoost являются ОВ (30) и слабый АМО $WeakLearn$.

Алгоритм AdaBoost имеет следующий вид. Выбирается натуральное число T, и выполняется нижеследующая последовательность из T шагов. На каждом шаге $t = 1, \ldots, T$ определяются следующие объекты:

- вектора $\vec{w}_t = (w_t^1, \ldots, w_t^N)$ и $\vec{p}_t = (p_t^1, \ldots, p_t^N)$, где

  $$\begin{array}{l} \vec{p}_t = norm(\vec{w}_t), \\ \vec{w}_1 = (p^1, \ldots, p^N) \;\; (p^i \text{ – компоненты исходной ОВ } S, \end{array}$$

- классификатор h_t, получаемый применением исходного слабого АМО $WeakLearn$ к ОВ $S(\vec{p}_t)$, где $S(\vec{p}_t)$ получается из S заменой в каждой входящей в неё тройке (x^i, y^i, p^i) компоненты p^i на p_t^i,

- $\varepsilon_t := Err(h_t, S(\vec{p}_t))$ $(< \frac{1}{2})$, $\beta_t := \frac{\varepsilon_t}{1-\varepsilon_t}$,
- $w^i_{t+1} := w^i_t \beta_t^{l^i_t}$, где $l^i_t = [\![h_t(x^i) = y^i]\!]$.

Затем определяется искомый классификатор h:

$$h(x) = [\![\langle \vec{q}, \vec{h}(x) \rangle \geq \tfrac{1}{2}]\!] \tag{31}$$

где $\vec{q} = norm(\ln \frac{1}{\beta_1}, \ldots, \ln \frac{1}{\beta_T})$, $\vec{h}(x) = (h_1(x), \ldots, h_T(x))$.

Теорема 4.

Ошибка результирующего классификатора (31) удовлетворяет неравенству

$$Err(h, S) \leq 2^T \textstyle\prod_{t=1}^T \sqrt{\varepsilon_t(1-\varepsilon_t)}. \tag{32}$$

Доказательство.

Согласно определениям, $\forall\, t = 1, \ldots, T$ верны равенства

$$\varepsilon_t = \textstyle\sum_{i\in I} p^i_t(1 - l^i_t) = 1 - \sum_{i\in I} p^i_t l^i_t = 1 - \frac{1}{|\vec{w}_t|}\sum_{i\in I} w^i_t l^i_t$$

Следовательно, $\sum_{i\in I} w^i_t l^i_t = |\vec{w}_t|(1-\varepsilon_t)$, откуда, учитывая неравенство (18) для $\beta = \beta_t$, получаем:

$$\begin{aligned} &|\vec{w}_{t+1}| = \textstyle\sum_{i\in I} w^i_t \beta_t^{l^i_t} \leq \sum_{i\in I} w^i_t(1 - (1-\beta_t)l^i_t) = \\ &= |\vec{w}_t| - (1-\beta_t)\textstyle\sum_{i\in I} w^i_t l^i_t = \\ &= |\vec{w}_t| - (1-\beta_t)|\vec{w}_t|\textstyle\sum_{i\in I} p^i_t l^i_t = \\ &= |\vec{w}_t|(1 - (1-\beta_t)(1-\varepsilon_t)) = |\vec{w}_t| 2\varepsilon_t. \end{aligned} \tag{33}$$

Таким образом, $\forall\, t = 1, \ldots, T$

$$|\vec{w}_{t+1}| \leq |\vec{w}_t| 2\varepsilon_t. \tag{34}$$

Перемножая неравенства (34) для $t = 1, \ldots, T$, и учитывая $|\vec{w}_1| = 1$, получаем:

$$|\vec{w}_{T+1}| \leq 2^T \textstyle\prod_{t=1}^T \varepsilon_t. \tag{35}$$

Отметим, что $\forall\, i \in I$ из $h(x_i) \neq y_i$ следует, что

$$\textstyle\prod_{t=1}^T \beta_t^{l^i_t} \geq (\prod_{t=1}^T \beta_t)^{1/2} \tag{36}$$

Действительно, $l^i_t = 1 - |h_t(x_i) - y_i|$, и

- если $y_i = 0$ и $h(x_i) = 1$, то $\forall\, t = 1, \ldots, T$

  $$\begin{aligned} &\beta_t^{l^i_t} = \beta_t^{1-|h_t(x_i)-y_i|} = \beta_t^{1-h_t(x_i)} \\ &\textstyle\sum_{t=1}^T \ln \frac{1}{\beta_t} h_t(x_i) \geq \frac{1}{2} \sum_{t=1}^T \ln \frac{1}{\beta_t} \end{aligned}$$

  откуда следует (36) для данного случая, и

- если $y_i = 1$ и $h(x_i) = 0$, то $\forall\, t = 1, \ldots, T$

$$\begin{array}{l} \beta_t^{l_t^i} = \beta_t^{1-|h_t(x_i)-y_i|} = \beta_t^{h_t(x_i)} \\ \sum_{t=1}^T \ln \frac{1}{\beta_t} h_t(x_i) < \frac{1}{2} \sum_{t=1}^T \ln \frac{1}{\beta_t} \end{array}$$

откуда следует (36) для данного случая.

Учитывая (36), получаем:

$$\begin{array}{l} |\vec{w}_{T+1}| \geq \sum_{i:h(x_i)\neq y_i} w_{T+1}^i = \sum_{i:h(x_i)\neq y_i} p^i \prod_{t=1}^T \beta_t^{l_t^i} \geq \\ \geq (\sum_{i:h(x_i)\neq y_i} p^i)(\prod_{t=1}^T \beta_t)^{1/2} = Err(h,S)(\prod_{t=1}^T \beta_t)^{1/2}, \end{array}$$

откуда, учитывая (35), получаем (32). ■

Следствие 1.

Пусть $\forall\, t = 1, \ldots, T$ ошибка ε_t классификатора h_t из алгоритма AdaBoost удовлетворяет условию $\varepsilon_t \leq \frac{1}{2} - \gamma_t$, где $\gamma_t > 0$. Тогда ошибка результирующего классификатора (31) удовлетворяет условию

$$Err(h,S) \leq e^{-2\sum_{t=1}^T \gamma_t^2}. \tag{37}$$

Доказательство.

В данном случае правая часть (32) равна

$$\prod_{t=1}^T \sqrt{1-4\gamma_t^2} = e^{\sum_{t=1}^T \frac{1}{2}\ln(1-4\gamma_t^2)} \;,$$

откуда, учитывая неравенство $\ln(1-4\gamma_t^2) \leq -4\gamma_t^2$, получаем (37). ■

В частности, если $\forall\, t = 1, \ldots, T$ $\gamma_t = \gamma$, то (37) будет иметь вид

$$Err(h,S) \leq e^{-2T\gamma^2}$$

откуда получаем оценку на число итераций AdaBoost, достаточных для выполнения условия $Err(h,S) < \varepsilon$:

$$T > \frac{1}{2\gamma^2} \ln \frac{1}{\varepsilon} \;.$$

7. Алгоритм следования за возмущённым лидером

В этом пункте рассматривается другой подход к решению задачи прогнозирования, описанной в пункте 5. ПА, построенный в соответствии с данным подходом, называется **алгоритмом следования за возмущённым лидером (Follow the Perturbed Leader, FPL)**, описания данного алгоритма и его разновидностей впервые было изложено в работах [4], [5], [6], [7]. Данный алгоритм является вероятностной модификацией обычного ПА A следования за лидером, который имеет следующий вид: на каждом шаге прогнозирования t определяется лидер, т.е. такой эксперт i, кумулятивные потери L^i_{t-1} которого минимальны, и на шаге t прогноз A полагается равным прогнозу лидера i. Потери ПА A в момент t совпадают с потерями эксперта i в момент t. Такой ПА может привести к потерям, существенно превышающим потери каждого из экспертов. Например, пусть число экспертов равно двум, и последовательности их потерь на шагах $1,\dots,7$ имеют вид

$$\begin{array}{ll} l^1_{1,\dots,7} = & (0.5,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1), \\ l^2_{1,\dots,7} = & \quad (0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0). \end{array}$$

Последовательности соответствующих кумулятивных потерь имеют вид

$$\begin{array}{ll} L^1_{1,\dots,7} = & (0.5,\ 0.5,\ 1.5,\ 1.5,\ 2.5,\ 2.5,\ 3.5), \\ L^2_{1,\dots,7} = & \quad (0,\quad 1,\quad 1,\quad 2,\quad 2,\quad 3,\quad 3). \end{array}$$

Нетрудно видеть, что в данном случае лидерами на шагах $2,\dots,7$ являются соответственно 2, 1, 2, 1, 2, 1, и каждый раз, следуя за лидером на текущем шаге, ПА следования за лидером будет нести потерю 1, и его кумулятивные потери на шагах $2,\dots,7$ будет иметь вид

$$L_{2,\dots,7} = \ (1,\ 2,\ 3,\ 4,\ 5,\ 6)$$

т.е. его кумулятивные потери на каждом шаге примерно вдвое больше кумулятивных потерь каждого из экспертов.

Излагаемый ниже ПА FPL отличается от детерминированного ПА следования за лидером лишь в изменении понятия лидера: на каждом шаге t лидером среди экспертов $1,\dots,N$ является тот эксперт i (называемый **возмущённым лидером**), у которого минимальной является величина

$$L^i_{t-1} - \tfrac{1}{\varepsilon_t}\xi^i,$$

где ε_t – параметр, и $\xi^1,\dots,\xi^N$ – независимые одинаково распределенные **случайные величины (СВ)** , с экспоненциальным законом распределения, т.е. их плотность имеет вид $p(x) = e^{-x}$ $(x \geq 0)$.

$\forall\, t = 1,\dots,T$ обозначим

$$\begin{array}{l} i_t = \ \text{СВ} \ \arg\min_{i\in I}(L^i_{t-1} - \tfrac{1}{\varepsilon_t}\xi^i) \\ l_t = \mathbf{E} l^{i_t}_t = \sum_{i\in I} l^i_t \mathbf{P}\{i_t = i\}, \ \ L_T = \sum_{t=1}^T l_t, \end{array}$$

где $\mathbf{P}\{\varphi\}$ обозначает вероятность события φ, а $\mathbf{E}\xi$ обозначает математическое ожидание СВ ξ.

Теорема 5.

Если параметр ε_t из ПА FPL имеет вид $\sqrt{\frac{2\ln N}{t}}$, то

$$L_T \le \min_{i\in I} L_T^i + 3\sqrt{2T\ln N}. \qquad (38)$$

Доказательство.

$\forall\, t = 1,\ldots,T$ обозначим

$$\begin{array}{l} i'_t = \text{ CB } \arg\min_{i\in I}(L_t^i - \frac{1}{\varepsilon_t}\xi^i) \\ l'_t = \mathbf{E} l_t^{i'_t} = \sum_{i\in I} l_t^i \mathbf{P}\{i'_t = i\}, \;\; L'_T = \sum_{t=1}^T l'_t. \end{array}$$

Неравенство (38) следует из доказываемых ниже соотношений (39) и (49).

1) Докажем неравенства

$$L_T \le L'_T + \sum_{t=1}^T \varepsilon_t \le L'_T + 2\sqrt{2T\ln N}. \qquad (39)$$

Второе неравенство следует из определения ε_t и свойства

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \le 1 + \int_1^T \frac{dt}{\sqrt{t}} < 2\sqrt{T}, \qquad (40)$$

а первое неравенство следует из свойства

$$\forall\, t = 1,\ldots,T \quad l_t - l'_t \le \varepsilon_t l_t \;\; (\le \varepsilon_t, \text{ т.к. } l_t \in [0,1]). \qquad (41)$$

(41) следует из неравенств

$$l'_t \ge e^{-\varepsilon_t} l_t \ge (1-\varepsilon_t) l_t. \qquad (42)$$

Второе неравенство в (42) следует из свойства

$$\forall\, x \in \mathbb{R} \;\; e^{-x} \ge 1 - x,$$

а первое неравенство в (42) можно переписать в виде

$$\sum_{i\in I} l_t^i \mathbf{P}\{i_t = i\} \le e^{\varepsilon_t} \sum_{i\in I} l_t^i \mathbf{P}\{i'_t = i\} \qquad (43)$$

(43) следует из свойства $\forall\, i \in I$

$$\mathbf{P}\{i_t = i\} \le e^{\varepsilon_t} \mathbf{P}\{i'_t = i\}. \qquad (44)$$

(44) следует из соответствующих неравенств для условных вероятностей: $\forall\, c_1,\ldots,c_N \ge 0$

$$\begin{array}{r} \mathbf{P}\{i_t = i \mid \forall\, j \ne i \;\; \xi^j = c_j\} \le \\ \le e^{\varepsilon_t} \mathbf{P}\{i'_t = i \mid \forall\, j \ne i \;\; \xi^j = c_j\} \end{array} \qquad (45)$$

Докажем (45). Обозначим условие $\forall j \neq i \;\; \xi^j = c_j$ символом φ, и определим

$$\begin{aligned}
&m_i = \min_{j\neq i}(L^j_{t-1} - \tfrac{1}{\varepsilon_t}c_j),\\
&m'_i = \min_{j\neq i}(L^j_t - \tfrac{1}{\varepsilon_t}c_j) = \min_{j\neq i}(L^j_{t-1} + l^j_t - \tfrac{1}{\varepsilon_t}c_j).
\end{aligned}$$

Нетрудно видеть, что $m_i \leq m'_i$.

Используя введенные выше обозначения, неравенство (45) можно переписать в виде неравенства условных вероятностей:

$$\begin{gathered}
\mathbf{P}\{L^i_{t-1} - \tfrac{1}{\varepsilon_t}\xi^i \leq m_i \mid \varphi\} \leq \\
\leq e^{\varepsilon_t}\mathbf{P}\{L^i_{t-1} + l^i_t - \tfrac{1}{\varepsilon_t}\xi^i \leq m'_i \mid \varphi\}.
\end{gathered} \tag{46}$$

Если в неравенстве в правой части (46) заменить l^i_t на 1, а m'_i на m_i, то данное неравенство усилится, поэтому (46) следует из неравенства

$$\begin{gathered}
\mathbf{P}\{L^i_{t-1} - \tfrac{1}{\varepsilon_t}\xi^i \leq m_i \mid \varphi\} \leq \\
\leq e^{\varepsilon_t}\mathbf{P}\{L^i_{t-1} + 1 - \tfrac{1}{\varepsilon_t}\xi^i \leq m_i \mid \varphi\},
\end{gathered} \tag{47}$$

которое эквивалентно неравенству

$$\begin{gathered}
\mathbf{P}\{\xi^i \geq \varepsilon_t(L^i_{t-1} - m_i) \mid \varphi\} \leq \\
\leq e^{\varepsilon_t}\mathbf{P}\{\xi^i \geq \varepsilon_t(L^i_{t-1} - m_i + 1) \mid \varphi\}.
\end{gathered} \tag{48}$$

(48) обосновывается следующими свойствами экспоненциально распределенной СВ ξ: $\forall\, a, b \geq 0$

$$\begin{aligned}
&\mathbf{P}\{\xi \geq a\} = e^{-a},\\
&\mathbf{P}\{\xi \geq a + b\} = e^{-b}\mathbf{P}\{\xi \geq a\}.
\end{aligned}$$

2) Докажем, что

$$L'_T \leq \min_{i\in I} L^i_T + \tfrac{\ln N}{\varepsilon_T}. \tag{49}$$

Будем использовать следующие обозначения:

$$\begin{aligned}
&\vec{l}_t := (l^1_t, \ldots, l^N_t),\ \vec{L}_t := (L^1_t, \ldots, L^N_t),\\
&\vec{\xi} := (\xi^1, \ldots, \xi^N),\\
&\tilde{l}_t = \vec{l}_t - \vec{\xi}(\tfrac{1}{\varepsilon_t} - \tfrac{1}{\varepsilon_{t-1}}),\ \tilde{L}_t = \vec{L}_t - \vec{\xi}\tfrac{1}{\varepsilon_t}
\end{aligned} \tag{50}$$

(полагаем $\varepsilon_0 = 1$).

Нетрудно доказать, что $\tilde{L}_T = \tilde{L}_{T-1} + \tilde{l}_T$.

Пусть $E = \{\vec{e}_1, \ldots, \vec{e}_N\} \subseteq \mathbb{R}^N$, где $\forall\, i = 1, \ldots, N$ $\vec{e}_i$ имеет вид $(0, \ldots, 1, \ldots 0)$ (единица – на i-м месте).

$\forall \vec{l} = (l^1, \dots, l^N) \in \mathbb{R}$ обозначим

$$M(\vec{l}) = \arg\min_{\vec{e}_i \in E} \langle \vec{e}_i, \vec{l} \rangle,$$

Нетрудно видеть, что

$$\begin{array}{l} \langle M(\vec{l}), \vec{l} \rangle = \min_{i \in I} l^i, \\ l_t^{i'_t} = \langle M(\tilde{L}_t), \vec{l}_t \rangle, \quad L'_T = \mathbf{E} \sum_{t=1}^T \langle M(\tilde{L}_t), \vec{l}_t \rangle, \\ \langle M(\tilde{L}_{T-1}), \tilde{L}_{T-1} \rangle \leq \langle M(\tilde{L}_T), \tilde{L}_{T-1} \rangle \end{array} \tag{51}$$

(неравенство в третьей строчке (51) следует из того, что его левая часть – минимальная компонента $\tilde{L}_{T-1}$, а правая – некоторая компонента $\tilde{L}_{T-1}$).

Индукцией по T докажем неравенство

$$\sum_{t=1}^T \langle M(\tilde{L}_t), \tilde{l}_t \rangle \leq \langle M(\tilde{L}_T), \tilde{L}_T \rangle. \tag{52}$$

При $T = 1$ (52) имеет вид $\langle M(\tilde{l}_1), \tilde{l}_1 \rangle \leq \langle M(\tilde{l}_1), \tilde{l}_1 \rangle$.

Индуктивный переход (от $T - 1$ к T): используя индуктивное предположение, и учитывая неравенство в третьей строчке (51), получаем:

$$\begin{array}{l} \text{левая часть (52)} \leq \\ \leq \langle M(\tilde{L}_{T-1}), \tilde{L}_{T-1} \rangle + \langle M(\tilde{L}_T), \tilde{l}_T \rangle \leq \\ \leq \langle M(\tilde{L}_T), \tilde{L}_{T-1} \rangle + \langle M(\tilde{L}_T), \tilde{l}_T \rangle = \langle M(\tilde{L}_T), \tilde{L}_T \rangle = \\ = \text{ правая часть (52)}. \end{array}$$

Используя определение $\tilde{l}_t$ (см. третью строчку в (50)), неравенство (52) можно переписать так:

$$\begin{array}{l} \sum_{t=1}^T \langle M(\tilde{L}_t), \vec{l}_t \rangle \leq \\ \leq \langle M(\tilde{L}_T), \tilde{L}_T \rangle + \sum_{t=1}^T \langle M(\tilde{L}_t), \vec{\xi} \rangle (\frac{1}{\varepsilon_t} - \frac{1}{\varepsilon_{t-1}}). \end{array} \tag{53}$$

Из определения $\tilde{L}_T$, следует неравенство

$$\begin{array}{l} \langle M(\tilde{L}_T), \tilde{L}_T \rangle \leq \langle M(\vec{L}_T), \vec{L}_T - \vec{\xi} \frac{1}{\varepsilon_T} \rangle = \\ = \min_{i \in I} L_T^i - \langle M(\vec{L}_T), \xi \rangle \frac{1}{\varepsilon_T}. \end{array} \tag{54}$$

Т.к. $\langle M(\vec{L}_T), \xi \rangle = \xi^k$ для некоторого k, и $\mathbf{E}\xi^k = 1$, то

$$\mathbf{E} \langle M(\vec{L}_T), \xi \rangle \frac{1}{\varepsilon_T} = \frac{1}{\varepsilon_T} \mathbf{E} \xi^k = \frac{1}{\varepsilon_T}. \tag{55}$$

Оценим второй член в (53):

$$\begin{array}{l}\sum_{t=1}^{T}\langle M(\tilde{L}_t),\vec{\xi}\rangle(\frac{1}{\varepsilon_t}-\frac{1}{\varepsilon_{t-1}})) \le \\ \le \sum_{t=1}^{T}\max_{i\in I}\xi^i(\frac{1}{\varepsilon_t}-\frac{1}{\varepsilon_{t-1}}) \le \frac{1}{\varepsilon_T}\max_{i\in I}\xi^i.\end{array}$$

Нетрудно доказать, что

$$\mathbf{E}\max_{i=1,\dots,N}\xi^i \le 1+\ln N. \tag{56}$$

Действительно, поскольку СВ $\xi^1,\dots,\xi^N$ независимы, и функция распределения показательно распределённой СВ имеет вид $1-e^{-x}$, то функция распределения СВ $\max_{i=1,\dots,N}\xi^i$ имеет вид $(1-e^{-x})^N$, поэтому её плотность равна $N(1-e^{-x})^{N-1}$, и следовательно её мат. ожидание равно

$$N\int_0^\infty(1-e^{-x})^{N-1}e^{-x}xdx. \tag{57}$$

Обозначим (57) записью a_N. Поскольку

$$\begin{array}{l}a_N=N\int_0^\infty(1-e^{-x})^{N-1}e^{-x}xdx= \\ =N\int_0^\infty(1-e^{-x})(1-e^{-x})^{N-2}e^{-x}xdx= \\ =\frac{N}{N-1}a_{N-1}-N\int_0^\infty e^{-x}(1-e^{-x})^{N-2}e^{-x}xdx= \\ =\frac{N}{N-1}a_{N-1}-\frac{N}{N-1}\int_0^\infty e^{-x}xd(1-e^{-x})^{N-1}= \\ \text{(применяем интегрирование по частям)} \\ =\frac{N}{N-1}a_{N-1}+\frac{N}{N-1}\int_0^\infty(1-e^{-x})^{N-1}de^{-x}x= \\ =\frac{N}{N-1}a_{N-1}+\frac{N}{N-1}\int_0^\infty(1-e^{-x})^{N-1}e^{-x}(1-x)dx= \\ =\frac{N}{N-1}a_{N-1}+\frac{1}{N-1}(1-a_N),\end{array}$$

откуда получаем: $a_N=a_{N-1}+\frac{1}{N}$, следовательно

$$\mathbf{E}\max_{i=1,\dots,N}\xi^i=a_N=1+\frac{1}{2}+\ldots+\frac{1}{N}, \tag{58}$$

откуда следует (56), поэтому

$$\begin{array}{l}\mathbf{E}\sum_{t=1}^{T}\langle M(\tilde{L}_t),\vec{\xi}\rangle(\frac{1}{\varepsilon_t}-\frac{1}{\varepsilon_{t-1}}) \le \\ \le \frac{\mathbf{E}\max_{i=1,\dots,N}\xi^i}{\varepsilon_T} \le \frac{1+\ln N}{\varepsilon_T}\end{array} \tag{59}$$

Таким образом, согласно второй строчке в (51), а также (53), (54), (55) и (59)

$$L'_T=\mathbf{E}\sum_{t=1}^{T}\langle M(\tilde{L}_t),\vec{l}_t\rangle \le \min_{i\in I}L_T^i-\frac{1}{\varepsilon_T}+\frac{1+\ln N}{\varepsilon_T}$$

откуда следует (49). ■

8. Агрегирующий алгоритм В.Г.Вовка

В этом пункте описывается агрегирующий алгоритм В.Г.Вовка, существенной особенностью которого является зависимость регрета $R_T = L_T - \min_{i\in I} L^i_T$ только от количества экспертов N и независимость регрета от величины периода наблюдения T. Данный алгоритм был впервые описан в [8].

8.1. Смешиваемые функции потерь

Напомним некоторые введённые ранее понятия и обозначения:

- $Y = \{0, 1\}$ – множество **исходов**,
- $\Gamma = [0, 1]$, или $[-1, 1]$, или $Y^\triangle$ – множество **прогнозов**,
- $\eta > 0$ – **параметр обучения**, $\beta = e^{-\eta}$,
- $\lambda : Y \times \Gamma \to \mathbb{R}_{\geq 0}$ – **функция потерь (ФП)**, она предполагается непрерывной по второму аргументу,
- $I = 1, \ldots, N$ – множество **экспертов**,
- $\forall\, t \geq 1$, $\forall\, i \in I$
  - γ^i_t – **прогноз** эксперта i на шаге t,
  - y_t – **исход** на шаге t,
  - $l^i_t = \lambda(\gamma^i_t, y_t)$ – **потери** эксперта i на шаге t, $L^i_t = \sum_{t'=1}^t l^i_{t'}$ – кумулятивные потери эксперта i на шаге t,
  - $m_t = \log_\beta \sum_{i\in I} \beta^{l^i_t} p^i_{t-1}$ – **средние потери** на шаге t, $M_t = \sum_{t'=1}^t m_{t'}$ – кумулятивные средние потери.

Алгоритм обучения: это построение последовательностей $\vec{w}_0, \vec{w}_1, \ldots$ и $\vec{p}_0, \vec{p}_1, \ldots$ векторов из $\mathbb{R}^I$, где

- $\vec{w}_0 = \vec{p}_0 \in I^\triangle$,
- $\forall\, t \geq 1, \forall\, i \in I$ $\; w^i_t = \beta^{l^i_t} w^i_{t-1} = \beta^{L^i_t} p^i_0$,
- $\forall\, t \geq 1$ $\; \vec{p}_t = norm(\vec{w}_t) \in I^\triangle$.

Отметим, что $M_t = \log_\beta \sum_{i\in I} \beta^{L^i_t} p^i_0$. Действительно,

$$\begin{aligned} m_t &= \log_\beta \textstyle\sum_{i\in I} \beta^{l^i_t} p^i_{t-1} = \log_\beta \sum_{i\in I} \beta^{l^i_t} \frac{w^i_{t-1}}{\sum_{j\in I} w^j_{t-1}} = \\ &= \log_\beta \frac{\sum_{i\in I} \beta^{l^i_t} w^i_{t-1}}{\sum_{j\in I} w^j_{t-1}} = \log_\beta \frac{\sum_{i\in I} \beta^{L^i_t} p^i_0}{\sum_{j\in I} \beta^{L^j_{t-1}} p^j_0} = \\ &= \log_\beta \textstyle\sum_{i\in I} \beta^{L^i_t} p^i_0 - \log_\beta \sum_{i\in I} \beta^{L^i_{t-1}} p^i_0 \end{aligned}$$

откуда непосредственно следует доказываемое равенство.

Примеры ФП:

$$\lambda(\gamma, y) = \begin{cases} c(y-\gamma)^2 \text{ (квадратичная, } c \text{ – константа),} \\ c|y-\gamma| \text{ (абсолютная, } c \text{ – константа),} \\ [\![y \neq \gamma]\!] \text{ (простая),} \\ -\ln\gamma(y) \text{ (логарифмическая, } \Gamma = Y^{\triangle}). \end{cases}$$

В последнем случае можно отождествить распределение

$$\gamma = (\gamma(1), \gamma(0)) \in Y^{\triangle}$$

с числом $\gamma(1) \in [0, 1]$, которое будем обозначать тем же символом γ, и значение $\lambda(\gamma, y)$ в данном случае можно записать в виде $-\ln|1 - y - \gamma|$.

ФП λ называется **смешиваемой ФП (СФП)**, если

$$\mathcal{U}_\lambda = \bigcup_{\gamma\in\Gamma}[0, \beta^{\lambda(\gamma,0)}] \times [0, \beta^{\lambda(\gamma,1)}] \qquad (60)$$

является выпуклым подмножеством $\mathbb{R}^2$ (это будет, например, когда ФП λ – квадратичная или логарифмическая.

В настоящей статье рассматривается задача вычисления прогнозов в том случае, когда ФП λ является СФП.

Теорема 6.

Если λ – СФП, то $\forall\, t \geq 1\ \ \exists\, \gamma_t^* \in \Gamma : \forall\, y \in Y$

$$\lambda(\gamma_t^*, y) \leq m_t. \qquad (61)$$

Доказательство.

Совокупность точек

$$\{(\beta^{(\lambda(\gamma_t^i,0)}, \beta^{\lambda(\gamma_t^i,1)}) \mid i \in I\} \qquad (62)$$

принадлежит выпуклому множеству $\mathcal{U}_\lambda$.

Согласно определению, m_t – выпуклая комбинация вида

$$m_t = \log_\beta \textstyle\sum_{i\in I} \beta^{\lambda(\gamma_t^i,y)} p_{t-1}^i.$$

Т.к. $\mathcal{U}_\lambda$ выпукло, то выпуклая комбинация

$$\textstyle\sum_{i\in I}(\beta^{(\lambda(\gamma_t^i,0)}, \beta^{\lambda(\gamma_t^i,1)} p_{t-1}^i \qquad (63)$$

точек из множества (62) тоже принадлежит $\mathcal{U}_\lambda$.

Построим луч с началом в $0 = (0, 0)$, проходящий через точку (63). Этот луч пересекает границу

$$\{(\beta^{\lambda(\gamma,0)}, \beta^{\lambda(\gamma,1)}) \mid \gamma \in \Gamma\}$$

множества $\mathcal{U}_\lambda$ в некоторой точке

$$u = (\beta^{\lambda(\gamma_t^*,0)}, \beta^{\lambda(\gamma_t^*,1)}). \tag{64}$$

Поскольку точка (63) принадлежит отрезку $[0, u]$, то её абсцисса и ордината не превосходят абсциссы и ординаты соответственно точки u, т.е.

$$\begin{array}{l} \sum_{i \in I} \beta^{\lambda(\gamma_t^i,0)} p_{t-1}^i \le \beta^{\lambda(\gamma_t^*,0)}, \\ \sum_{i \in I} \beta^{\lambda(\gamma_t^i,1)} p_{t-1}^i \le \beta^{\lambda(\gamma_t^*,1)}, \end{array}$$

т.е. верно утверждение

$$\forall\, y \in Y \ \ \sum_{i \in I} \beta^{\lambda(\gamma_t^i,y)} p_{t-1}^i \le \beta^{\lambda(\gamma_t^*,y)}. \tag{65}$$

Логарифмируя неравенство в (65), получаем (61). ■

L_T^{AA} обозначает кумулятивную потерю $\sum_{t=1}^T \lambda(\gamma_t^*, y_t)$, где γ_t^* – прогнозы, определяемые в доказательстве теоремы 6 (AA является аббревиатурой словосочетания «агрегирующий алгоритм»). Из теоремы 6 следует неравенство

$$L_T^{AA} \le M_T.$$

Отметим, что если $\vec{p}_0$ – р.р., то $\forall\, i \in I$

$$L_t^{AA} \le M_t = \log_\beta(\textstyle\sum_{i \in I} \beta^{L_t^i} \frac{1}{N}) \le \log_\beta(\beta^{L_t^i} \frac{1}{N}),$$

поэтому

$$L_T^{AA} \le \min_{i \in I} L_T^i + \tfrac{\ln N}{\eta}. \tag{66}$$

8.2. Смешиваемость квадратичной функции потерь

Будем считать, что $\Gamma = [-1, 1]$, $Y = \{-1, 1\}$ (м.б. и $[-1, 1]$). В этом случае $\mathcal{U}_\lambda$ имеет вид

$$\textstyle\bigcup_{\gamma \in \Gamma} [0, \beta^{\lambda(\gamma,-1)}] \times [0, \beta^{\lambda(\gamma,1)}]$$

Лемма 2.

ФП $\lambda(\gamma, y) = (y - \gamma)^2$ является η–смешиваемой тогда и только тогда, когда $\eta \le \frac{1}{2}$.

Доказательство.

Нетрудно доказать, что множество $\mathcal{U}_\lambda$ выпукло тогда и только тогда, когда его граница – кривая

$$\begin{array}{l} \hat{\mathcal{U}}_\lambda \overset{\text{def}}{=} \{(\beta^{\lambda(\gamma,-1)}, \beta^{\lambda(\gamma,1)}) \mid \gamma \in [-1, 1]\} = \\ = \{(e^{-\eta(-1-\gamma)^2}, e^{-\eta(1-\gamma)^2}) \mid \gamma \in [-1, 1]\} \end{array} \tag{67}$$

обладает следующим свойством: при увеличении γ от -1 до 1 абсцисса соответствующей точки кривой уменьшается и кривая поворачивает налево, что эквивалентно свойству

$$\forall\,\gamma \in [-1,1] \quad \begin{cases} \gamma_x < 0, \\ y_{xx} = \gamma_x \frac{x_\gamma y_{\gamma\gamma} - x_{\gamma\gamma} y_\gamma}{x_\gamma^2} \leq 0. \end{cases} \tag{68}$$

Из первого неравенства в (68) следует, что второе неравенство в (68) равносильно неравенству

$$x_\gamma y_{\gamma\gamma} \geq x_{\gamma\gamma} y_\gamma. \tag{69}$$

Нетрудно видеть, что

$$\begin{aligned} &x_\gamma = -2\eta(1+\gamma)e^{-\eta(1+\gamma)^2} \\ &x_{\gamma\gamma} = 2\eta(-1+2\eta(1+\gamma)^2)e^{-\eta(1+\gamma)^2} \\ &y_\gamma = 2\eta(1-\gamma)e^{-\eta(1-\gamma)^2} \\ &y_{\gamma\gamma} = 2\eta(-1+2\eta(1-\gamma)^2)e^{-\eta(1-\gamma)^2} \end{aligned}$$

откуда следует, что (69) можно переписать в виде

$$-(1+\gamma)(-1+2\eta(1-\gamma)^2) \geq (1-\gamma)(-1+2\eta(1+\gamma)^2).$$

Последнее неравенство эквивалентно утверждению

$$\eta(1-\gamma^2) \leq \tfrac{1}{2},$$

которое должно быть верным для каждого $\gamma \in [-1,1]$, что равносильно неравенству $\eta \leq \frac{1}{2}$. ■

Рассмотрим задачу вычисления оптимального прогноза γ_t^* для квадратичной ФП в случае $\eta = \frac{1}{2}$.

Точка (63) в данном случае имеет вид (A,B), где

$$A = \textstyle\sum_i \beta^{\lambda(\gamma_t^i,-1)} p_{t-1}^i, \quad B = \sum_i \beta^{\lambda(\gamma_t^i,1)} p_{t-1}^i,$$

и точка (64) ищется из уравнения

$$\tfrac{B}{A} = \tfrac{\beta^{\lambda(\gamma_t^*,1)}}{\beta^{\lambda(\gamma_t^*,1)}} = \beta^{(1-\gamma_t^*)^2-(-1-\gamma_t^*)^2} = \beta^{-4\gamma_t^*}.$$

Поэтому $\gamma_t^* = \frac{1}{4}\log_\beta \frac{A}{B}$.

8.3. Супермартингалы

8.3.1. Понятие супермартингала

Пусть $Y = \{0, 1\}$, $\Gamma = [0, 1]$. Ниже запись u_n обозначает произвольную последовательность из $(\Gamma \times Y)^n$ $(n \geq 0)$:

$$u_n = ((\gamma_1, y_1), \ldots, (\gamma_n, y_n)).$$

Последовательность u_0 пуста и обозначается ε.

Супермартингал (СМ) – это семейство функций

$$Q = \{Q_n : (\Gamma \times Y)^n \to \mathbb{R}_{\geq 0} \mid n \geq 0\} \tag{70}$$

таких, что

- $Q_0(\varepsilon) \leq 1$,
- $\forall\, n \geq 0, \forall\, u_n \in (\Gamma \times Y)^n, \forall\, y \in Y$ функция

  $$Q_{n+1}(u_n, (\cdot, y)) : \Gamma \to \mathbb{R}_{\geq 0}$$

  непрерывна, и верно свойство

  $$\begin{aligned} \forall\, \gamma \in \Gamma \quad & \gamma Q_{n+1}(u_n, (\gamma, 1)) + \\ & + (1 - \gamma) Q_{n+1}(u_n, (\gamma, 0)) \leq Q_n(u_n). \end{aligned} \tag{71}$$

Теорема 7.
Пусть задан СМ Q вида (70).
Тогда $\forall\, n \geq 0$, $\forall\, u_n \in (\Gamma \times Y)^n$ $\exists\, \gamma^*$:

$$\forall\, y \in Y \;\; Q_n(u_n) \geq Q_{n+1}(u_n, (\gamma^*, y)). \tag{72}$$

Доказательство.
Определим функцию $f_{u_n} : \Gamma \times Y \to \mathbb{R}_{\geq 0}$:

$$f_{u_n}(\gamma, y) = Q_{n+1}(u_n, (\gamma, y)) - Q_n(u_n).$$

f_{u_n} непрерывна по γ, и из (71) следует, что

$$\forall\, \gamma \in \Gamma \;\; \gamma f_{u_n}(\gamma, 1) + (1 - \gamma) f_{u_n}(\gamma, 0) \leq 0 \tag{73}$$

поэтому $f_{u_n}(1, 1) \leq 0$ и $f_{u_n}(0, 0) \leq 0$.

Докажем, что

$$\exists\, \gamma^* : f_{u_n}(\gamma^*, 0) \leq 0 \text{ и } f_{u_n}(\gamma^*, 1) \leq 0. \tag{74}$$

Отметим, что из (74) следует (72).

- Если $f_{u_n}(1,0) \le 0$, то $\gamma^* = 1$, и если $f_{u_n}(0,1) \le 0$, то $\gamma^* = 0$,
- иначе $f_{u_n}(1,0) > 0$ и $f_{u_n}(0,1) > 0$, в этом случае рассмотрим непрерывную функцию

$$f(\gamma) = f_{u_n}(\gamma,1) - f_{u_n}(\gamma,0).$$

Поскольку $f(0) > 0$, $f(1) < 0$, и f непрерывна, то

$$\exists\, \gamma^* \in (0,1) : f(\gamma^*) = 0,$$

т.е. $f_{u_n}(\gamma^*,1) = f_{u_n}(\gamma^*,0)$. По (73), отсюда следует (74). ■

8.3.2. Пример супермартингала

Определим $\forall\, i \in I = \{1,\ldots,N\}$

$$R_n^i(u_n) = \sum_{t=1}^n (\lambda(\gamma_t, y_t) - \lambda(\gamma_t^i, y_t)).$$

В некоторых случаях $Q_n^i(u_n) = e^{\eta R_n^i} = e^{\eta(L_n - L_n^i)}$ будет СМ. Неравенство в (71) равносильно неравенству

$$\gamma e^{\eta(\lambda(\gamma,1)-\lambda(\gamma_{n+1}^i,1))} + (1-\gamma)e^{\eta(\lambda(\gamma,0)-\lambda(\gamma_{n+1}^i,0))} \le 1 \tag{75}$$

1) Если $\lambda(\gamma,y) = -\ln|1-y-\gamma|$, то (75) имеет вид

$$\gamma e^{\eta(-\ln\gamma+\ln\gamma_{n+1}^i)} + (1-\gamma)e^{\eta(-\ln(1-\gamma)+\ln(1-\gamma_{n+1}^i))} \le 1 \tag{76}$$

что равносильно неравенству

$$\gamma^{1-\eta}(\gamma_{n+1}^i)^\eta + (1-\gamma)^{1-\eta}(1-\gamma_{n+1}^i)^\eta \le 1. \tag{77}$$

Если $\eta = \frac{1}{2}$, то (77) следует из неравенства Коши- Буняковского для векторов

$$(\gamma^{1-\eta}, (1-\gamma)^{1-\eta}), \quad ((\gamma_{n+1}^i)^\eta, (1-\gamma_{n+1}^i)^\eta).$$

Левая часть (77) – скалярное произведение этих векторов, а их норма в случае $\eta = \frac{1}{2}$ равна 1.

2) $\lambda(\gamma,y) = (y-\gamma)^2 : y \in \{0,1\}, \gamma \in [0,1], \eta \in (0,2]$.

В этом случае (75) равносильно неравенству

$$\gamma e^{\eta((\gamma-1)^2-(1-\gamma_{n+1}^i)^2)} + (1-\gamma)e^{\eta(\gamma^2-(\gamma_{n+1}^i)^2)} \le 1. \tag{78}$$

Представим γ^i_{n+1} в виде $\gamma + x$ и перепишем (78) в виде

$$\gamma e^{2\eta(1-\gamma)x} + (1-\gamma)e^{-2\eta\gamma x} \le e^{\eta x^2} \tag{79}$$

(79) вытекает из следующего утверждения: если значения СВ ξ лежат в отрезке $[a, b]$, то $\forall\, s \in \mathbb{R}$

$$\ln \mathbf{E} e^{s\xi} \le s\mathbf{E}\xi + \tfrac{s^2(b-a)^2}{8}. \tag{80}$$

Если СВ ξ принимает значение 1 с вероятностью γ и значение 0 с вероятностью $1-\gamma$, то полагая $a = 0$, $b = 1$ получаем: (80) имеет вид

$$\gamma e^{s(1-\gamma)} + (1-\gamma)e^{-s\gamma} \le e^{s^2/8}.$$

Если $s := 2\eta x$, то Л.Ч. (79) $\le e^{\eta^2 x^2/2} \le$ Пр.Ч. (79), ибо $\eta \le 2$.

8.3.3. Применение теоремы 7

Пусть $Y = \{0, 1\}$, $\Gamma = [0, 1]$, ФП $\lambda(\gamma, y)$ – η-смешиваемая, $w_0 \in I^{\triangle}$, где $I = \{1, \ldots, N\}$, построены прогнозы $\gamma_1, \ldots, \gamma_{T-1}$ и имеются прогнозы экспертов $\gamma^i_1, \ldots, \gamma^i_T$ $(i \in I)$.

(66) будет выполнено, если $\forall\, t$ верно (65), что равносильно следующему: $\forall\, y \in \{0, 1\}$ верно неравенство

$$\textstyle\sum_{i\in I} w^i_{t-1} \ge \sum_{i\in I} w^i_{t-1} e^{-\eta(\lambda(\gamma^i_t, y) - \lambda(\gamma_t, y))}$$

которое эквивалентно неравенству

$$\textstyle\sum_{i\in I} p^i_0 e^{-\eta L^i_{t-1}} \ge \sum_{i\in I} p^i_0 e^{-\eta L^i_{t-1}} e^{\eta(\lambda(\gamma_t, y) - \lambda(\gamma^i_t, y))},$$

после домножения обоих частей которого на $e^{\eta L_{t-1}}$ получаем

$$\textstyle\sum_{i\in I} p^i_0 Q^i_{t-1} \ge \sum_{i\in I} p^i_0 Q^i_{t-1} e^{\eta(\lambda(\gamma_t, y) - \lambda(\gamma^i_t, y))} = \sum_{i\in I} p^i_0 Q^i_t(y).$$

Таким образом утверждение (65) равносильно тому, что последовательность $\{\sum_{i\in I} p^i_0 Q^i_t \mid t \ge 1\}$ не возрастает с ростом t.

Нетрудно доказать, что семейство функций

$$\{\textstyle\sum_{i\in I} p^i_0 Q^i_n \mid n \ge 1\}$$

тоже является СМ, и поэтому для него верна теорема 7, т.е. $\forall\, t > 0$, $\forall\, u \in (\Gamma \times Y)^{t-1}$ $\exists\, \gamma_t : \forall\, y \in Y$

$$\textstyle\sum_{i\in I} p^i_0 Q^i_t(u, (\gamma_t, y)) \le \sum_{i\in I} p^i_0 Q^i_{t-1}(u) \le 1. \tag{81}$$

$\forall\, t \ge 1$ из свойства

$$\textstyle\sum_{i\in I} p^i_0 Q^i_t = \sum_{i\in I} p^i_0 e^{\eta(L_t - L^i_t)} \le 1$$

следует, что $\forall\, i \in I$ $p^i_0 e^{\eta(L_t - L^i_t)} \le 1$, откуда следует (66), если $p^i_0 = 1/N$.

9. Прогнозные стратегии

В этом пункте рассматривается задача прогнозирования временного ряда в ситуации, когда эксперты отсутствуют. Для определения качества алгоритма прогнозирования в данном случае используется понятие калибруемости алгоритма, впервые введённое в [10]. Излагаемый в данном пункте вероятностный алгоритм вычисления калибруемых прогнозов впервые был описан в [11], см. также [9] и [6].

9.1. Понятие прогнозной стратегии

Пусть множество исходов Y имеет вид $\{0, 1\}$. Будем обозначать произвольную последовательность из $Y^n = \underbrace{Y \times \ldots \times Y}_{n}$ записью y^n, и последний элемент последовательности y^n – записью y_n. Обозначим записью Y^* множество $\bigcup_{n\geq 0} Y^n$, где Y^0 состоит из пустой последовательности y^0 (которая обозначается ε). Символом y будем обозначать неограниченную последовательность элементов множества Y, и если y – такая последовательность,то $\forall\, n \geq 1$ записи y^n и y_n обозначают префикс последовательности y длины n и n-й элемент y соответственно.

Прогнозная стратегия (ПС) – это функция

$$\begin{array}{l} f : Y^* \to [0, 1], \text{ где } f(\varepsilon) = 1 \text{ и } \forall\, n \geq 0, \forall\, y^n \in Y^n \\ f(y^n) = \mathbf{P}\{y^n = \text{ последовательность первых } n \text{ исходов}\} \end{array}$$

Для каждой ПС f и $\forall\, n \geq 0$ верно равенство

$$\textstyle\sum_{y^n\in Y^n} f(y^n) = 1.$$

Будем обозначать $f(y_n \mid y^{n-1}) = \frac{f(y^n)}{f(y^{n-1})}$.

Пусть $\mathcal{F}$ – некоторый класс ПС. $\forall\, y \in Y^*$ будем обозначать записью $\mathcal{F}(y)$ число $\sum_{\varphi\in\mathcal{F}} \varphi(y)$.

Ниже будем опускать «$\in \mathcal{F}$» в $\sum_{\varphi\in\mathcal{F}}, \sup_{\varphi\in\mathcal{F}}, \inf_{\varphi\in\mathcal{F}}$.

Потери ПС f определяются следующим образом: $\forall\, n \geq 1$

$$\begin{array}{l} l^f_n(y) = -\ln f(y_n \mid y^{n-1}), \\ L^f_n(y) = \sum_{i=1}^n l^f_i(y) = -\sum_{i=1}^n \ln f(y_i \mid y^{i-1}) = -\ln f(y^n). \end{array}$$

Регрет ПС f относительно класса ПС $\mathcal{F}$:

$$\begin{array}{l} R^f_n(y) = L^f_n(y) - \inf_\varphi L^\varphi_n(y) = \sup_\varphi \ln \frac{\varphi(y^n)}{f(y^n)} = \ln \frac{\sup_\varphi \varphi(y^n)}{f(y^n)} \\ R_n(y) = \inf_\varphi R^\varphi_n(y). \end{array}$$

9.2. Примеры прогнозных стратегий

9.2.1. Смешивающая и минимаксная прогнозные стратегии

Пусть задан конечный класс $\mathcal{F}$ ПС. **Смешивающая ПС** для класса $\mathcal{F}$ определяется следующим образом:

$$f(y_n) = \tfrac{1}{|\mathcal{F}|}\mathcal{F}(y^n).$$

Нетрудно видеть, что

$$\begin{aligned} R_n^f(y) = \sup_\varphi \ln \tfrac{\varphi(y^n)}{f(y^n)} = \sup_\varphi \ln \tfrac{\varphi(y^n)}{\frac{1}{|\mathcal{F}|}\mathcal{F}(y^n)} = \\ = \ln|\mathcal{F}| + \sup_\varphi \ln \tfrac{\varphi(y^n)}{\mathcal{F}(y^n)} \leq \ln|\mathcal{F}|. \end{aligned}$$

Минимаксная ПС для класса $\mathcal{F}$ определяется следующим образом:

$$f(y^n) = \tfrac{\sup_\varphi \varphi(y^n)}{\sum_{u^n\in Y^n}\sup_\varphi \varphi(u^n)}.$$

Ниже будем опускать «$\in Y^n$» в записях $\sum_{u^n\in Y^n}$.

Минимаксный регрет определяется следующим образом:

$$\begin{aligned} V_n^f &= \sup_{y^n} R_n^f(y) = \sup_{y^n} \ln \tfrac{\sup_\varphi \varphi(y^n)}{f(y^n)} \\ V_n &= \inf_{\varphi\in\mathcal{F}} V_n^\varphi. \end{aligned}$$

Нетрудно видеть, что

$$\begin{aligned} R_n^f(y) = \ln \tfrac{sup_\varphi \varphi(y^n)}{f(y^n)} = \ln \tfrac{\sup_\varphi \varphi(y^n)}{\frac{\sup_\varphi \varphi(y^n)}{\sum_{u^n}\sup_\varphi \varphi(u^n)}} = \\ = \ln \textstyle\sum_{u^n} \sup_\varphi \varphi(u^n). \end{aligned} \tag{82}$$

Отметим, что правая часть (82) не зависит от y и f, поэтому можно обозначить её R_n или V_n.

ПС f называется **оптимальной**, если $\forall\, n \geq 0$ $V_n^f = V_n$.

Докажем оптимальность минимаксной ПС f, т.е. следующее свойство: для каждой ПС $f' \neq f$ $V_n^{f'} \geq V_n$:

- если $\forall\, y^n \in Y^n$ $f'(y^n) = f(y^n)$, то

$$V_n^{f'} = \sup_{y^n} \ln \tfrac{\sup_\varphi \varphi(y^n)}{f'(y^n)} = \sup_{y^n} \ln \tfrac{\sup_\varphi \varphi(y^n)}{f(y^n)} = V_n^f,$$

- иначе, если $\exists\, y^n \in Y^n : f'(y^n) \neq f(y^n)$, то, поскольку

$$\textstyle\sum_{y^n} f'(y^n) = \sum_{y^n} f(y^n) = 1,$$

то $\exists\, y^n : f'(y^n) < f(y^n)$, поэтому

$$R_n^{f'}(y) = \ln \tfrac{\sup_\varphi \varphi(y^n)}{f'(y^n)} > \ln \tfrac{\sup_\varphi \varphi(y^n)}{f(y^n)} = V_n,$$

откуда следует: $V_n^{f'} = \sup_{y^n} R_n^{f'}(y) > V_n$. ■

Отметим, что

$$V_n = \ln \sum_{u^n} \sup_\varphi \varphi(u^n) \leq \ln \sum_{u^n} \sum_\varphi \varphi(u^n) =$$
$$= \ln \sum_\varphi \sum_{u^n} \varphi(u^n) = \ln \sum_\varphi 1 \leq \ln |\mathcal{F}|.$$

9.2.2. Прогнозная стратегия Лапласа

ПС Лапласа, применяется в ситуации, когда $\forall\, n \geq 1$ y_n генерируется независимо, и $\forall\, n \geq 1$ $\mathbf{P}\{y_n = 1\}$ равно одному и тому же числу p.

Ниже $\forall\, n \geq 0$ и $\forall\, y^n \in Y^n$ n_1 и n_2 обозначают число единиц и нулей соответственно в последовательности y^n.

Нетрудно доказать, что вероятность того, что y^n является последовательностью первых n исходов, равна $p^{n_1}(1-p)^{n_2}$.

Определим класс ПС $\mathcal{F}$ как класс функций, каждая из которых соответствует некоторому числу $p \in [0,1]$ и сопоставляет последовательности $y^n \in Y^n$ вероятность $p^{n_1}(1-p)^{n_2}$ того, что y^n – последовательность первых n исходов.

ПС Лапласа f имеет следующий вид:

$$f(y^n) \stackrel{\text{def}}{=} \int_0^1 p^{n_1}(1-p)^{n_2} dp, \tag{83}$$

т.е. $f(y^n)$ равно мат. ожиданию вероятности того, что y^n является последовательностью первых n исходов. Нетрудно доказать, что данное значение равно $\frac{1}{(n+1)C_n^{n_1}}$. Это доказывается обратной индукцией по n_1:

- для $n_1 = n$ имеем $\int_0^1 p^n dp = \frac{1}{n+1}$, что верно, и
- если верно

  $$\int_0^1 p^{n_1+1}(1-p)^{n_2-1} dp = \frac{1}{(n+1)C_n^{n_1+1}} =: A$$

  то, интегрируя по частям, получаем:

  $$\int_0^1 p^{n_1}(1-p)^{n_2} dp = \frac{n-n_1}{n_1+1} A = \frac{1}{(n+1)C_n^{n_1}}.$$

Нетрудно видеть, что

$$f(y_{n+1} = 1 \mid y^n) = \frac{f(y^n 1)}{f(y^n)} = \frac{1}{(n+2)C_{n+1}^{n_1+1}} \Big/ \frac{1}{(n+1)C_n^{n_1}} = \frac{n_1+1}{n+2},$$
$$f(y_{n+1} = 0 \mid y^n) = \frac{n_2+1}{n+2}.$$

$$\forall\, y \;\; R_n^f(y) = \ln \frac{\sup_{0 \leq p \leq 1} p^{n_1}(1-p)^{n_2}}{\int_0^1 p^{n_1}(1-p)^{n_2} dp} = \ln \frac{(\frac{n_1}{n})^{n_1}(\frac{n_2}{n})^{n_2}}{\frac{1}{(n+1)C_n^{n_1}}} \leq \ln(n+1).$$

Для оптимальной ПС оценка V_n имеет следующий вид:

$$V_n = \frac{1}{2}\ln n + \frac{1}{2}\ln\frac{\pi}{2} + \varepsilon_n \;\; (\text{где } \varepsilon_n \to 0). \tag{84}$$

Действительно,

$$V_n = \ln \sum_{u^n} \sup_{\varphi} \varphi(u^n) = \\ = \ln \sum_{u^n} \sup_{0 \le p \le 1} p^{n_1}(1-p)^{n_2} = \\ = \ln \sum_{u^n} (\tfrac{n_1}{n})^{n_1} (\tfrac{n_2}{n})^{n_2} = \\ = \ln \sum_{n_1=0}^{n} C_n^{n_1} (\tfrac{n_1}{n})^{n_1} (\tfrac{n_2}{n})^{n_2}. \tag{85}$$

Из формулы Стирлинга

$$\sqrt{2\pi n}(\tfrac{n}{e})^n e^{\frac{1}{12n}} \le n! \le \sqrt{2\pi n}(\tfrac{n}{e})^n e^{\frac{1}{12n+\varepsilon}}$$

(где $\varepsilon \in (0,1)$ – некоторая константа) следует неравенство

$$\tfrac{1}{\sqrt{2\pi}} \sqrt{\tfrac{n}{n_1 n_2}} e^{\frac{1}{12n}} \le C_n^{n_1} (\tfrac{n_1}{n})^{n_1} (\tfrac{n_2}{n})^{n_2} \le \tfrac{1}{\sqrt{2\pi}} \sqrt{\tfrac{n}{n_1 n_2}} e^{\frac{1}{12n+1}}$$

из которого, учитывая (85), получаем верхнюю оценку

$$V_n \le \ln \left((1+o(1)) \sqrt{\tfrac{n}{2\pi}} e^{\frac{1}{12n+1}} \sum_{n_1=1}^{n-1} \tfrac{1}{\sqrt{n_1 n_2}} \right)$$

Однако

$$\sum_{n_1=1}^{n-1} \tfrac{1}{\sqrt{n_1 n_2}} = \sum_{n_1=1}^{n-1} \tfrac{1}{n} \tfrac{1}{\sqrt{\frac{n_1}{n}(1-\frac{n_1}{n})}} \approx \int_0^1 \tfrac{dx}{\sqrt{x(1-x)}} = \pi,$$

поэтому $V_n \le \ln((1+o(1))\sqrt{\frac{n\pi}{2}}$ = правая часть (84).

Нижняя оценка для n устанавливается аналогично. ■

Из (84) следует, что для оптимальной ПС f, т.е. такой ПС f, что $V_n = V_n^f$, будет выполнено свойство

$$\forall\, y\ L_n^f(y) - \inf_{\varphi} L_n^{\varphi}(y) \le \text{ Пр.Ч. (84)},$$

или: $\forall\, y,\ \forall\, \varphi \in \mathcal{F}\ \ L_n^f(y) \le L_n^{\varphi}(y) +$ Пр.Ч. (84).

9.3. Детерминированное прогнозирование

В этом и следующем пунктах мы рассмотрим задачу прогнозирования временного ряда в условиях, когда эксперты отсутствуют. Мы рассмотрим детерминированный и вероятностный алгоритмы прогнозирования.

Ниже используется следующее обозначение: если a – некоторый объект, и m – сообщение, то запись $a!m$ обозначает действие, которое заключается в том, что объект a посылает в окружающую среду сообщение m.

Алгоритм детерминированного прогнозирования (АДП) заключается в выполнении следующих действий:

$$\forall\, n \ge 1 \begin{cases} \text{прогнозист } !\ \gamma_n \in [0,1] \\ \text{природа } !\ y_n \in [0,1] \end{cases}$$

Будем говорить, что АДП **калибруется**, если для каждой последовательности исходов $(y_1, y_2, \ldots)$ и каждого связного подмножества $I \subseteq [0,1]$ последовательность прогнозов $(\gamma_1, \gamma_2, \ldots)$ обладает следующим свойством:

$$\frac{\sum_{i=1}^n [\![\gamma_i \in I]\!](y_i - \gamma_i)}{\sum_{i=1}^n [\![\gamma_i \in I]\!]} \to 0 \text{ при } \sum_{i=1}^n [\![\gamma_i \in I]\!] \to \infty. \tag{86}$$

АДП не калибруется, для обоснования этого определим

$$I_1 = [0, \tfrac{1}{2}], \quad I_2 = [\tfrac{1}{2}, 1], \quad \forall\, n \geq 1 \;\; y_n = [\![\gamma_n < \tfrac{1}{2}]\!]$$

откуда следует, что $\forall\, n \geq 1$ $|y_n - \gamma_n| \geq \frac{1}{2}$.

Нетрудно установить, что (86) нарушается когда $I = I_1$ или $I = I_2$. Действительно, в один из отрезков I_1, I_2 попадает бесконечное число точек γ_i, и свойство $\sum_{i=1}^n [\![\gamma_i \in I]\!] \to \infty$ для этого отрезка верно, однако модуль дроби в (86) больше или равен $\frac{1}{2}$.

9.4. Вероятностное прогнозирование

Алгоритм вероятностного прогнозирования (АВП) имеет следующий вид:

$$\forall\, n \geq 1 \begin{cases} \text{прогнозист ! } p_n \in [0,1]^{\triangle} \\ \text{природа ! } y_n \in \{0,1\} \\ \text{ГСЧ ! } \gamma_n \sim p_n \end{cases} \tag{87}$$

где ГСЧ – генератор случайных чисел, третье действие в (87) заключается в случайном порождении значения $\gamma_n \in [0,1]$ в соответствии с распределением p_n.

Свойство **калибруемости** АВП имеет следующий вид:

- для каждого $\delta > 0$, и
- для каждой последовательности исходов $(y_1, y_2, \ldots)$

последовательность прогнозов $(\gamma_1, \gamma_2, \ldots)$, которую порождает АВП, удовлетворяет условию: $\forall\, I \subseteq [0,1]$

$$\models \left(\lim_{n\to\infty} \sup \left| \tfrac{1}{n} \sum_{i=1}^n [\![\gamma_i \in I]\!](y_i - \gamma_i) \right| \right) \leq \delta \tag{88}$$

где обозначение $\models A$ имеет следующий смысл: событие A выполняется с вероятностью 1.

Теорема 8.
Существует калибруемый АВП.

Доказательство.
Представим доказательство в виде последовательности этапов.

1) $\forall\, k \geq 1$ обозначим $V_k = \{v_0, \ldots, v_k\}$, где $v_i = \frac{i}{k}$.

$\forall\, c \in [0,1]$ определён отрезок $[v_{i-1}, v_i]$, который содержит c.

Число c можно представить в виде суммы

$$c = \lambda v_{i-1} + (1-\lambda) v_i, \quad \text{где } \lambda \in [0,1].$$

$$\forall\, v \in V_k \text{ определим } w_v(c) \overset{\text{def}}{=} \begin{cases} \lambda & \text{при } v = v_{i-1}, \\ 1-\lambda & \text{при } v = v_i, \\ 0, & \text{иначе.} \end{cases}$$

Таким образом, $c = \sum_{v \in V_k} w_v(c) v$.

2) Определим индуктивно последовательность $c_1, c_2, \ldots$ чисел из $[0,1]$ следующим образом. Полагаем $c_1 = 0$.

Пусть для некоторого n определены числа

$$c_1, \ldots, c_{n-1}. \tag{89}$$

$\forall\, v \in V_k$ будем использовать обозначение

$$\mu_n(v) = \textstyle\sum_{i=1}^{n} w_v(c_i)(y_i - c_i), \tag{90}$$

где $c_1, \ldots, c_{n-1}$ – числа из (89), и y_n, c_n – переменные.

Из (90) следует, что

$$\begin{array}{l} \mu_n(v)^2 = \mu_{n-1}(v) + w_v(c_n)(y_n - c_n))^2 = \\ = \mu_{n-1}(v)^2 + 2\mu_{n-1}(v) w_v(c_n)(y_n - c_n) + w_v(c_n)^2 (y_n - c_n)^2, \end{array}$$

поэтому $\sum_{v \in V_k} \mu_n(v)^2 = A + 2(y_n - c_n)B + C$, где

$$\begin{cases} A = \sum_{v \in V_k} \mu_{n-1}(v)^2 \\ B = \sum_{v \in V_k} w_v(c_n) \mu_{n-1}(v) \\ C = (y_n - c_n)^2 \sum_{v \in V_k} w_v(c_n)^2 \end{cases}$$

Заметим:

$$\begin{array}{l} B = \sum_{v \in V_k} w_v(c_n) \sum_{i=1}^{n-1} w_v(c_i)(y_i - c_i) = \\ = \sum_{i=1}^{n-1} \Big(\sum_{v \in V_k} w_v(c_n) w_v(c_i) \Big)(y_i - c_i) = \\ = \sum_{i=1}^{n-1} \langle \vec{w}(c_n), \vec{w}(c_i) \rangle (y_i - c_i) = \\ = \sum_{i=1}^{n-1} K(c_n, c_i)(y_i - c_i), \end{array}$$

где $\vec{w}(c) = (w_{v_0}(c), \ldots w_{v_k}(c))$, $K(c_n, c_i) = \langle \vec{w}(c_n), \vec{w}(c_i) \rangle$.

Определяем c_n следующим образом:

- если уравнение

$$B(c) = \sum_{i=1}^{n-1} K(c, c_i)(y_i - c_i) = 0$$

имеет корень в $[0, 1]$, то полагаем c_n равным этому корню,

- иначе $c_n = 1$ или 0, если $\forall\, c \in [0, 1]$ $B(c) > 0$ или $B(c) < 0$ соответственно.

3) Отметим, что $2(y_n - c_n)B \le 0$ и $C \le \sum_{v \in V_k} w_v(c_n) = 1$.

Таким образом,

$$\begin{aligned} &\sum_{v \in V_k} \mu_n(v)^2 = A + 2(y_n - c_n)B + C \le \\ &\le \sum_{v \in V_k} \mu_{n-1}(v)^2 + 1, \end{aligned}$$

откуда, учитывая равенство $\mu_0(v) = 0$, получаем:

$$\sum_{v \in V_k} \mu_n(v)^2 \le n. \tag{91}$$

4) Определим $p_n \in V_k^{\triangle} : \forall\, v \in V_k \;\; p_n(v) = w_v(c_n)$.

$\forall\, i = 1, \ldots, n$ рассмотрим СВ

$$\xi_i = [\![p_i \in I]\!](y_i - p_i),$$

где $I \subseteq [0, 1]$ – заданное подмножество. Используя ξ_i, перепишем условие (88) в виде

$$\models \left(\lim_{n\to\infty} \sup |\tfrac{1}{n} \sum_{i=1}^{n} \xi_i| \right) \le \delta. \tag{92}$$

Нетрудно видеть, что

$$\mathbf{E}\xi_i = \sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - v). \tag{93}$$

По усиленному закону больших чисел,

$$\models \lim_{n\to\infty} \left(\tfrac{1}{n} \sum_{i=1}^{n} \xi_i - \tfrac{1}{n} \sum_{i=1}^{n} \mathbf{E}\xi_i \right) = 0. \tag{94}$$

Поскольку $\forall\, i = 1, \ldots, n$

$$\begin{aligned} &|\mathbf{E}\xi_i - \sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - c_i)| = \\ &= |\sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - v) - \\ &- \sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - c_i)| = \\ &= |\sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](c_i - v)| < \delta \end{aligned} \tag{95}$$

то

$$|\sum_{i=1}^{n} \mathbf{E}\xi_i| \le |\sum_{i=1}^{n} \sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - c_i)| + \delta n. \tag{96}$$

Обозначим записями $\vec{\mu}_n$ и $\vec{I}$ вектора

$$(\mu_n(v_0), \ldots, \mu_n(v_k)) \text{ и } ([\![v_0 \in I]\!], \ldots, [\![v_k \in I]\!])$$

соответственно. Из (91) следует, что $||\vec{\mu}_n|| \leq \sqrt{n}$.

Используя неравенство Коши-Буняковского, оценим первое слагаемое в правой части (96):

$$\begin{aligned} & |\sum_{i=1}^{n} \sum_{v \in V_k} w_v(c_i)[\![v \in I]\!](y_i - c_i)| = \\ & = |\sum_{v \in V_k} (\sum_{i=1}^{n} w_v(c_i)(y_i - c_i)[\![v \in I]\!]| = \\ & = |\sum_{v \in V_k} \mu_n(v)[\![v \in I]\!]| = \\ & = |\langle \vec{\mu}_n, \vec{I} \rangle| \leq ||\vec{\mu}_n|| \cdot ||\vec{I}|| \leq \sqrt{n}\sqrt{k+1} \end{aligned} \tag{97}$$

Из (4.15) и (4.16) следует:

$$|\sum_{i=1}^{n} \mathbf{E}\xi_i| \leq \sqrt{n}\sqrt{k+1} + \delta n. \tag{98}$$

Из (98) и (94) получаем соотношение (92). ■

Нетрудно доказать более сильное утверждение: существует АВП, такой, что для каждой последовательности исходов $(y_1, y_2, \ldots)$ последовательность прогнозов $(\gamma_1, \gamma_2, \ldots)$ обладает свойством:

$$\forall\, I \subseteq [0,1] \quad \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i = 0.$$

Для обоснования этого утверждения в процессе конструирования чисел c_n нужно в определенные моменты времени n_s $(s \geq 1)$ изменять δ, т.е. вместо фиксированного δ рассматривать последовательность δ_s $(s \geq 1)$, стремящуюся к 0.

10. Заключение

В работе были изложены основные понятия смешивающего прогнозирования, и приведены доказательства основных свойств изложенных алгоритмов смешивающего прогнозирования. Развитие изложенных результатов может заключаться, например, путем нового определения меры качества алгоритма прогнозирования, и построения алгоритма смешивающего прогнозирования, оптимального относительно новой меры качества. Например, в качестве такой меры качества можно выбрать долю ошибочных предсказаний алгоритма смешивающего прогнозирования не на всём периоде наблюдения, а на некоторой его части, на которой предсказания экспертов имеют высокую точность.

Список литературы

[1] Вьюгин В.В., *Математические основы машинного обучения и прогнозирования*, МЦНМО, Москва, 2018, 384 pp.

[2] Littlestone N., Warmuth M., "The weighted majority algorithm", *Information and Computation*, **108** (1994), 212–261

[3] Freund Y., Schapire R.E., "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, **55** (1997), 119–139

[4] Hannan J., "Approximation to Bayes risk in repeated plays", *Contributions to the Theory of Games (ed. by M.Dresher, A.W.Tucker, and P. Wolfe)*, **3** (1957), 97–139

[5] Kalai A., Vempala S., "Efficient algorithms for online decisions", *Journal of Computer and System Sciences*, **71** (2005), 291–307

[6] G. Lugosi, N. Cesa-Bianchi, *Prediction, Learning and Games*, Cambridge University Press, New York, 2006

[7] M. Hutter, J. Poland, "Adaptive online prediction by following the perturbed leader", *Journal of Machine Learning Research*, **6** (2005), 639–660

[8] V. Vovk, "Aggregating strategies", *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (M. Fulk and J.Case, editors)*, 1990, 371–383

[9] Cover Thomas M., Thomas Joy A., *Elements of Information Theory, 2nd ed.*, John Wiley and Sons, Inc., 2006, 748 pp.

[10] A.P. Dawid, "Calibration-based empirical probability", *Ann. Statist.*, **13** (1985), 1251–1285

[11] A. Chernov, Y. Kalnishkan, F. Zhdanov, V. Vovk, "Supermartingales in Prediction with Expert Advice", *Theoretical Computer Science*, **411**:29-30 (2010), 2647–2669

Mathematical foundations of time series prediction

Mironov A.M.

The article outlines the basic concepts and methods of prediction time series. Various mixing prediction algorithms are considered and assessments of the quality of these algorithms are provided.

Keywords: time series, prediction algorithms, mixing prediction

References

[1] Vyugin V.V., *Mathematical foundations of machine learning and prediction*, MCNMO, Moskva, 2018 (In Russian), 384 pp.

[2] Littlestone N., Warmuth M., "The weighted majority algorithm", *Information and Computation*, **108** (1994), 212–261

[3] Freund Y., Schapire R.E., "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, **55** (1997), 119–139

[4] Hannan J., "Approximation to Bayes risk in repeated plays", *Contributions to the Theory of Games (ed. by M.Dresher, A.W.Tucker, and P. Wolfe)*, **3** (1957), 97–139

[5] Kalai A., Vempala S., "Efficient algorithms for online decisions", *Journal of Computer and System Sciences*, **71** (2005), 291–307

[6] G. Lugosi, N. Cesa-Bianchi, *Prediction, Learning and Games*, Cambridge University Press, New York, 2006

[7] M. Hutter, J. Poland, "Adaptive online prediction by following the perturbed leader", *Journal of Machine Learning Research*, **6** (2005), 639–660

[8] V. Vovk, "Aggregating strategies", *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (M. Fulk and J.Case, editors)*, 1990, 371–383

[9] Cover Thomas M., Thomas Joy A., *Elements of Information Theory, 2nd ed.*, John Wiley and Sons, Inc., 2006, 748 pp.

[10] A.P. Dawid, "Calibration-based empirical probability", *Ann. Statist.*, **13** (1985), 1251–1285

[11] A. Chernov, Y. Kalnishkan, F. Zhdanov, V. Vovk, "Supermartingales in Prediction with Expert Advice", *Theoretical Computer Science*, **411**:29-30 (2010), 2647–2669