Вестник КРАУНЦ. Физ.-мат. науки. 2016. № 4-1(16). С. 112-125. ISSN 2079-6641

DOI: 10.18454/2079-6641-2016-16-4-1-112-125

УДК 519.7

АЛГОРИТМ СТОХАСТИЧЕСКОГО УСРЕДНЕННОГО ГРАДИЕНТА НА БАЗЕ АГРЕГИРУЮЩИХ ФУНКЦИИ *

З. М. Шибзухов, М. А. Казаков

Институт прикладной математики и автоматизации, 36000, г. Нальчик, ул. Шортанова, 89а

E-mail: szport@gmail.com, f_wolfgang@mail.ru

В работе предлагается новая градиентная схема для решения задачи минимизации усредненных потерь. Она является аналогом схемы, применяемой в алгоритме SAG в случае, когда риск вычисляется при помощи среднего арифметического. Приведен иллюстративный пример построения робастной классификации на основе максимизации суррогата медианы от отступов.

Ключевые слова: Эмпирический риск, задача классификации, усредняющая агрегирующая функция, градиентная схема.

MSC 68T27

STOCHASTIC GRADIENT ALGORITHM BASED ON THE AVERAGE AGGREGATE FUNCTIONS

Z. M. Shibzukhov, M. A. Kazakov

Institute of Applied Mathematics and Automation, 360000, KBR, Nalchik, st. Shortanova 89a, Russia

E-mail: szport@gmail.com, f_wolfgang@mail.ru

The paper proposes a new scheme for the gradient solution to minimize losses averaged problem. It is an analog circuit used in the SAG algorithm in the case when the risk is calculated using the arithmetic mean. An illustrative example of the construction of robust classification based on the maximization of the surrogate median indentation.

Key words: Empirical risk, classification problem, averaging aggregation function, gradient based algorithm.

*Работа выполнена при поддержке гранта РФФИ 15-01-03381 и гранта ОНИТ РАН

Введение

Метод минимизации эмпирического риска [1] является признанным методом решения задач параметрической классификации. Эмпирический риск обычно вычисляется как среднее арифметическое от значений параметрической **функции потерь**. Эмпирическая оценка средних потерь, как среднее арифметическое, адекватна со статистической точки зрения если потери распределены по нормальному закону. Однако даже для нормального закона среднее арифметическое не является робастной оценкой. В то время, как медиана позволяет оценивать эмпирическое среднее при наличии выбросов. Поэтому для построения параметрических регрессионных зависимостей также используется эмпирические оценки среднего при помощи медианы, несмотря на то, что использование медианы делает делает процедуру настройки параметров регрессионной зависимости более сложной с вычислительной точки зрения. В условиях выбросов также используют оценки квантилей, когда искажения в распределении потерь составляют меньше 50%. Это позволяет при настройке параметров при помощи медианы не терять полезную часть распределения потерь, которая расположена выше значения медианы, разделяющей упорядоченный по возрастанию набор потерь на две равные части.

Среднее арифметическое, медиана и квантили являются примерами **усредняющих агрегирующих функций**, к которым относятся практически все известные функции вычисления среднего значения. В настоящей работе рассматривается подход, когда для оценки средних потерь может использоваться произвольная усредняющая агрегирующая функция и рассматривается метод **стохастически усредненного градиента** для настройки параметров искомой зависимости на основе эмпирических оценок средних потерь, вычисляемых в этих условиях. Этот метод здесь применяется в случае, когда для вычисления приближенного значения медианы или квантиля используются агрегирующие функции, которые в определенном смысле апроксимируют медиану или квантиль и являются дифференцируемыми функциями, что позволяет в принципе использовать градиентные методы поиска параметров искомой зависимости для решения задачи классификации.

Классический метод эмпирического риска в задачах классификации

Задача поиска параметрического закона $y = R \circ A(\mathbf{x}, \mathbf{w})$ для разбиения на классы между входами $\mathbf{x}$ и скалярным выходом y является одной из классических задач машинного обучения. Здесь $A\colon \mathbf{X} \times \mathbf{W} \to \mathbf{U} \subseteq \mathbb{R}^m$ – это преобразование, которое вычисляет скалярную или векторную оценку. По ней при помощи решающего правила $R\colon \mathbf{U} \to \mathbf{Y}$ находится ответ. В случае задачи классификации на 2 класса или идентификации класса $\mathbf{Y} = \{0,1\}$ или $\mathbf{Y} = \{-1,0,1\}$ преобразование A является скалярной функцией. В случае задачи класификации на несколько классов $\mathbf{Y} = \{1,\ldots,q\}$ или $\mathbf{Y} = \{0,1,\ldots,q\}$ преобразование A, как правило, является векторным, т.е. $A = (A_1,\ldots,A_q)$, A_j вычисляет оценку «за класс ». Имеется конечный набор входов $\widetilde{\mathbf{X}} = \{\widetilde{\mathbf{x}}_k\colon k = 1..N\}$ и набор известных значений на выходе: $\widetilde{\mathbf{Y}} = \{\tilde{y}_k\colon k = 1..N\}$. Требуется найти такой набор параметров $\mathbf{w}^*$, что преобразование $R \circ A^*(\mathbf{x}) = R \circ A(\mathbf{x}, \mathbf{w}^*)$ адекватно относит элементы множества $\widetilde{\mathbf{X}}$ к соответствующим классам.

В качестве меры адекватности зависимости часто используют **эмпирический риск (empirical risk)**. Набор параметров $\mathbf{w}^*$, задающий адекватную параметрическую зависимость, должен минимизировать величину эмпирического риска.

Эмпирический риск обычно вычисляется как среднее арифметическое от значений параметрической **функции потерь (loss function)**:

$$\mathscr{Q}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N} \ell_k(\mathbf{w}), \tag{1}$$

где $\ell_k(\mathbf{w}) = \ell\big(\mu_k(\mathbf{w})\big)$, где $\ell\colon \mathbb{R} \to \mathbb{R}_+$ – функция потерь, $\mu_k(\mathbf{w}) = \mu(A(\widetilde{\mathbf{x}}_k, \mathbf{w}), \tilde{y}_k)$ – **функция отступа (margin function)** k-го примера из обучающего множества. Она вычисляет степень «удаленности» примера от неверных классов или, наоборот, степени «погруженности» в верный класс.

В случае 2-х классов $\mu(\mathbf{w}) = yA(\mathbf{x}, \mathbf{w})$. Величина потерь вычисляется при помощи функции $L\colon \mathbb{R} \to \mathbb{R}$ – монотонно невозрастающая функция потерь, така что $\lim\limits_{v\to\infty} L_1(v) = 0$.

В случае q классов

$$\mu(\mathbf{w}) = \begin{cases} (u_1 - u_2, \ldots, u_1 - u_q), & \text{если } \tilde{y} = 1 \\ (u_{\tilde{y}} - u_1, \ldots, u_{\tilde{y}} - u_{\tilde{y}-1}, u_{\tilde{y}} - u_{\tilde{y}+1}, \ldots, u_q), & \text{если } 1 < \tilde{y} < q \\ u_q - u_1, \ldots, u_q - u_{q-1}), & \text{если } \tilde{y} = q, \end{cases}$$

где $u_j = A_j(\tilde{\mathbf{x}}_k, \mathbf{w})$. Величина потерь вычисляется при помощи функция $L(v_1, \ldots, v_{q-1})$ – монотонно невозрастающая функция, такая что

1) если $v_1 \geqslant 0, \ldots, v_{q-1} \geqslant 0$, то $L(v_1, \ldots, v_{q-1}) \geqslant 0$;

2) если $v_1 > 0, \ldots, v_{q-1} > 0$, то $L(v_1, \ldots, v_{q-1}) > 0$.

Например,

$$L(v_1, \ldots, v_{q-1}) = \sum_{j=1}^{q-1} L(v_j).$$

Функция потерь – это неотрицательная невозрастающая функция, которая имеет единственный минимум, такой что $\lim\limits_{r\to+\infty} L(r) = \min L(r) = 0$. Например, функция Хинжа: $L(r) = (1-r)_+$.

Со статистической точки зрения потерь при помощи среднего арифметического является адекватной, если потери распределены по нормальному закону. Однако, если в действительности потери распределены по другому закону, то оценка средних потерь должна осуществляться другим способом. Но даже в случае нормально распределенных потерь среднее арифметическое не является устойчивой по отношению к выбросам в распределении. В этом случае существенно более адекватной оценкой является, например, медиана.

Среднее арифметическое и медиана являются примерами **усредняющей агрегирующей функции (averaged aggregation function)**. Поэтому в общем случае средние потери можно вычислять при помощи усредняющих агрегирующих функций.

Усредняющие агрегирующие функции

Пусть $\mathbb{I} \subseteq \mathbb{R}$ – сегмент $\mathbb{R}$, $\mathbb{I}^\star$ – множество всех конечных последовательностей $\{z_1,\dots,z_N\} \subset \mathbb{I}$, т.е.

$$\mathbb{I}^\star = \bigcup_{N=1}^{\infty} \mathbb{I}^N.$$

Агрегирующая функция (aggregation function) это отображение $\mathsf{M}\colon \mathbb{I}^\star \to \mathbb{I}$, которое удовлетворяет следующим требованиям:

- $\mathsf{M}\{z\} = z$;
- если $z'_1 \leqslant z''_1,\dots,z'_N \leqslant z''_N$, то $\mathsf{M}\{z'_1,\dots,z'_N\} \leqslant \mathsf{M}\{z''_1,\dots,z''_N\}$.

Последнее требование – требование монотонности агрегирующей функции.

Агрегирующая функция M – симметричная, если

$$\mathsf{M}\{z_1,\dots,z_N\} = \mathsf{M}\{z_{\pi(1)},\dots,z_{\pi(N)}\}$$

для любой перестановки π ряда чисел $1,\dots,N$.

Усредняющие агрегирующие функции (averaging aggregation function), по определению, удовлетворяют дополнительному требованию:

$$\min\{z_1,\dots,z_N\} \leqslant \mathsf{M}\{z_1,\dots,z_N\} \leqslant \max\{z_1,\dots,z_N\}.$$

Подробное изложение основных понятий и основных свойств агрегирующих функций можно найти в [4, 5, 6].

Существует универсальный способ определения усредняющих агрегирующих функций [7]. Для их определения используются штрафные функции (penalty function).

Функция $P(z_1,\dots,z_N,u)$ является штрафной функцией, если она удовлетворяет следующим требованиям:

- $P(z_1,\dots,z_N,u) \geqslant 0$ для всех u и $z_1,\dots,z_N$;
- $P(z_1,\dots,z_N,u) = 0$ только если $z_1 = \dots = z_N = u$;
- для всех $z_1,\dots,z_N$ множество

$$\mathbf{M}_{z_1,\dots,z_N} = \big\{u\colon P(z_1,\dots,z_N,u) = P_{\min}(z_1,\dots,z_N)\big\},$$

где

$$P_{\min}(z_1,\dots,z_N) = \min_u P(z_1,\dots,z_N,u),$$

является синглетоном или связным сегментом.

Всякую усредняющую агрегирующую функцию можно определить на основе некоторой штрафной функции P следующим образом:

$$\mathsf{M}_P\{z_1,\dots,z_N\} = \arg\min_u P(z_1,\dots,z_N,u),$$

если $\mathbf{M}_{z_1,\dots,z_N}$ – синглетон и

$$\mathsf{M}_P\{z_1,\dots,z_N\} = \frac{a+b}{2},$$

если $\mathbf{M}_{z_1,\ldots,z_N}$ – сегмент с концами a и b. Заметим, что формально в последнем случае можно было бы выбрать любое значение из интервала (a,b) или некоторое значение из (a,b), зависящее от P.

Далее рассмотрим разновидность штрафных функций, которые являются суммами функций несходства:

$$P(z_1,\ldots,z_N,u)=\sum_{k=1}^{N}p(u,z_k), \tag{2}$$

где $p(u,z)$ – **функция несходства (dissimilarity function).** Функция несходства определяется следующим образом.

Функция $p(z,u)$ является функцией несходства, если она удовлетворяет следующим условиям:

- $p(u,z)=0 \iff u=z$;
- $p(u,z_1)\geqslant p(u,z_2)$, когда $z_1\geqslant z_2\geqslant u$ или $z_1\leqslant z_2\leqslant u$.

Агрегирующую функцию, определенную на базе штрафной функции вида (2) будем обозначать M_p.

Статистическая интерпретация $\mathsf{M}_p\{z_1,\ldots,z_N\}$ на основе принципа максимума правдоподобия следующая: если случайная величина z распределена по вероятностному закону $e^{-p(\tilde{z},z)}$, где $\tilde{z}$ – среднее значение, то $\mathsf{M}_P\{z_1,\ldots,z_N\}$ является эмпирической оценкой $\tilde{z}$.

Уникальность минимума $P_{z_1,\ldots,z_N}(u)=P(z_1,\ldots,z_N,u)$ и монотонность $\mathsf{M}_p\{z_1,\ldots,z_N\}$ гарантированы, когда

$$p(u,z)=G\big(h(u)-h(z)\big), \tag{3}$$

где $G\colon \mathbb{R}\to\mathbb{R}$ – непрерывная неотрицательная выпуклая функция, $h(u)$ – обратимая монотонная функция [7, 6].

Приведем примеры известных усредняющих агрегирующих функций, которые можно определить таким образом.

- **Среднее арифметическое** получается при $p(u,z)=(u-z)^2$:

$$\mathsf{M}\{z_1,\ldots,z_N\}=\arg\min_u\sum_{k=1}^{N}(u-z_k)^2.$$

- **Медиана**

$$\operatorname{med}\{z_1,\ldots,z_N\}=\begin{cases}z_{(k)}, & \text{если } N=2k+1\\ (z_{(k)}+z_{(k+1)})/2, & \text{если } N=2k\end{cases}$$

получается при $p(u,z_k)=|u-z_k|$:

$$\operatorname{med}\{z_1,\ldots,z_N\}=\arg\min_u\sum_{k=1}^{N}|u-z_k|$$

где $z_{(1)},\ldots,z_{(N)}$ – множество $z_1,\ldots,z_N$, упорядоченное в порядке неубывания.

- α-квантиль $\mathsf{Q}_\alpha\{z_1,\dots,z_N\}$ получается при $p(u,z)=|u-z|_\alpha$:

$$\mathsf{Q}_\alpha\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^N|u-z_k|_\alpha,$$

где

$$|u|_\alpha=\begin{cases}\alpha|u|, & \text{если } u\geqslant 0\\ (1-\alpha)|u|, & \text{если } u<0.\end{cases}$$

- α-экспектиль

$$\mathsf{E}_\alpha\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^N|u-z_k|_\alpha^2,$$

где

$$|u|_\alpha^2=\begin{cases}\alpha u^2, & \text{если } u\geqslant 0\\ (1-\alpha)u^2, & \text{если } u<0.\end{cases}$$

- Среднее по Колмогорову

$$\mathsf{M}_g\{z_1,\dots,z_N\}=g^{-1}\Big(\frac{1}{N}\sum_{k=1}^N g(z_k)\Big);$$

получается при $p(u,z_k)=(g(u)-g(z_k))^2$:

$$\mathsf{M}_g\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^N(g(u)-g(z_k))^2.$$

- Масштабированная медиана

$$\mathrm{med}_g\{z_1,\dots,z_N\}=g^{-1}\big(\mathrm{med}\{g(z_k)\colon k=1,\dots,N\}\big)$$

получается при $p(u,z_k)=|g(u)-g(z_k)|$:

$$\mathrm{med}_g\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^N|g(u)-g(z_k)|.$$

Приведенные агрегирующие функции являются примерами усреднения величин, которые более устойчивы по отношению к выбросам.

Поиск значения $\mathsf{M}_p\{z_1,\dots,z_N\}$ можно осуществлять методом полного градиента или методом Ньютона. В первом случае на каждом шаге текущая оценка искомого значения обновляется по следующему правилу:

$$u_{t+1}=u_t-h_tP'_u(u_t,z_1,\dots,z_N),$$

где

$$P'_u(u_t,z_1,\dots,z_N)=\sum_{k=1}^N p'_u(u_t,z_k).$$

Во втором случае обновление осуществляется по следующему правилу:

$$u_{t+1}=u_t-h_t\frac{P'_u(u_t,z_1,\dots,z_N)}{P''_{uu}(u_t,z_1,\dots,z_N)},$$

где

$$\frac{P'_u(u_t, z_1, \ldots, z_N)}{P''_{uu}(u_t, z_1, \ldots, z_N)} = \frac{\sum\limits_{k=1}^{N} p'_u(u_t, z_k)}{\sum\limits_{k=1}^{N} p''_{uu}(u_t, z_k)}.$$

Параметр темпа обучения h_t в этих методах может быть постоянным или выбираться при помощи одного из методов поиска типа line search.

При больших N удобнее применять стохастические варианты этих алгоритмов. Например, такие алгоритмы, в основе которых лежит такая же схема, как и в основе SAG [10, 11].

В первом случае обновление будет осуществляться по правилу:

$$u_{t+1} = u_t - h_t \bar{g}_t,$$

где

$$\bar{g}_t = \frac{1}{N} \sum_{k=1}^{N} g_{t,k},$$

где $k(t)$ – номер случайно выбранного значения из $z_1, \ldots, z_N$ на шаге t. При этом,

$$g_{t+1,k} = \begin{cases} p'_u(u_t, z_k), & \text{если } k = k(t) \\ g_{t,k}, & \text{иначе.} \end{cases}$$

Среднее значение производной $\bar{g}_t$ можно обновлять на каждом шаге по простому правилу:

$$\bar{g}_{t+1} = \bar{g}_t + \tfrac{1}{N}(p'_u(u_t, z_k) - g_{t,k}).$$

Во втором случае обновление осуществляется по следующему правилу:

$$u_{t+1} = u_t - h_t \frac{G_t}{H_t},$$

где

$$G_t = \sum_{k=1}^{N} G_{t,k}$$

и

$$H_t = \sum_{k=1}^{N} H_{t,k}.$$

При этом,

$$G_{t+1,k} = \begin{cases} p'_u(u_t, z_k), & \text{если } k = k(t) \\ G_{t,k}, & \text{иначе,} \end{cases}$$

а

$$H_{t+1,k} = \begin{cases} p''_{uu}(u_t, z_k), & \text{если } k = k(t) \\ H_{t,k}, & \text{иначе.} \end{cases}$$

Значение отношения G_t/H_t обновляется на каждом шаге по простому правилу:

$$\frac{G_{t+1}}{H_{t+1}} = \frac{G_t + p'_u(u_t, z_k) - G_{t,k}}{H_t + p''_{uu}(u_t, z_k) - H_{t,k}}.$$

Algorithm 1 Алгоритм типа SAG для вычисления значения $\mathsf{M}_p\{z_1,\ldots,z_N\}$.

Инициализировать u_0
$g_k \leftarrow p'(u_0,z_k)$, $k=1,\ldots,N$
$G_k \leftarrow p'_u(u_0,z_k)$, $k=1,\ldots,N$
$G \leftarrow G_1+\cdots+G_N$
if используется схема Ньютона **then**
  $H_k \leftarrow p''_{uu}(u_0,z_k)$, $k=1,\ldots,N$
  $H \leftarrow H_1+\cdots+H_N$
end if
$t \leftarrow 0$
repeat
  $k \leftarrow k(t)$
  $G \leftarrow G+p'_u(u_t,z_k)-G_k$
  $G_k \leftarrow p'_u(u_t,z_k)$
  if используется схема Ньютона **then**
    $H \leftarrow H+p''_{uu}(u_t,z_k)-H_k$
    $H_k \leftarrow p''_{uu}(u_t,z_k)$
    $\bar{g}=\dfrac{G}{H}$
  else
    $\overline{g} \leftarrow G/N$
  end if
  $u_{t+1} \leftarrow u_t-h_t\overline{g}$
  $t \leftarrow t+1$
until значение u_t не стабилизируется

От эмпирического риска к агрегированному риску

Усредняющие агрегирующие функции уже использовались для построения функционалов потерь в [8, 9] в контексте задачи построения операций над алгоритмами классификации и регрессии, которые сохраняют свойство корректности алгоритмов. Применим их теперь для оценки средних потерь:

$$\mathscr{Q}_p(\mathbf{w}) = \mathsf{M}_p\big\{\ell_k(\mathbf{w})\colon k=1,\ldots,N\big\},$$

где усредняющая агрегирующая функция M_p определяется на основе штрафной функции вида (2):

$$\mathsf{M}_p\big\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\big\} = \arg\min_u \sum_{k=1}^{N} p\big(u,\ell_k(\mathbf{w})\big).$$

Оптимальный набор параметров $\mathbf{w}^\star$ доставляет минимум $\mathscr{Q}_p(\mathbf{w})$:

$$\mathscr{Q}_p(\mathbf{w}^\star) = \min_{\mathbf{w}} \mathscr{Q}_p(\mathbf{w}).$$

Если $p(z,u)$ имеет частные производные до второго порядка включительно, то

$$\frac{\partial \mathsf{M}_p\{z_1,\ldots,z_N\}}{\partial z_k} = -\frac{p''_{uz}(\bar{z},z_k)}{\sum_{\ell=1}^{N} p''_{uu}(\bar{z},z_k)},$$

где $\bar{z} = \mathsf{M}_p\{z_1,\ldots,z_N\}$. Тогда

$$\operatorname{grad}\mathsf{M}_p\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\} = \frac{\sum\limits_{k=1}^{N} -p''_{uz}\big(\bar{z},\ell_k(\mathbf{w})\big)\operatorname{grad}\ell_k(\mathbf{w})}{\sum\limits_{k=1}^{N} p''_{uu}\big(\bar{z},\ell_k(\mathbf{w})\big)}, \tag{4}$$

где $\bar{z} = \mathsf{M}_p\big\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\big\}$.

Поиск оптимального набора $\mathbf{w}$ можно осуществлять при помощи следующего варианта процедуры полного градиента. Правило обновления вектора параметров имеет вид:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t\operatorname{grad}\mathsf{M}_p\{\ell_1(\mathbf{w}_t),\ldots,\ell_N(\mathbf{w}_t)\}.$$

Обновление вектора параметров осуществляется до тех пор, пока значения $\mathbf{w}_t$ и $\mathsf{M}_p\{\ell_1(\mathbf{w}_{t+1}),\ldots,\ell_N(\mathbf{w}_{t+1})\}$ не стабилизируются.

Заметим, что если $p(u,z) = G(u-z)$ – частный случай (3), то

$$\operatorname{grad}\mathsf{M}_p\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\} = \sum_{k=1}^{N}\alpha_k(\mathbf{w})\operatorname{grad}\ell_k(\mathbf{w}),$$

где

$$\alpha_k(\mathbf{w}) = \frac{G''\big(\bar{z}-\ell_k(\mathbf{w})\big)}{G''\big(\bar{z}-\ell_1(\mathbf{w})\big)+\cdots+G''\big(\bar{z}-\ell_N(\mathbf{w})\big)},$$

причем $\alpha_1(\mathbf{w})+\cdots+\alpha_N(\mathbf{w}) = 1$.

Нетрудно увидеть, что в этом случае процедура градиентного спуска похожа на процедуру поиска минимума взвешенного среднего от потерь с числовыми весами. Однако, в данном случае веса являются функциями от $\bar{z}-\ell_1(\mathbf{w}),\ldots,\bar{z}-\ell_N(\mathbf{w})$ – отклонений между агрегированным средним от потерь и текущими потерями. Если $G(u-z) = (u-z)^2/2$, то $\alpha_k(\mathbf{w}) = \dfrac{1}{N}$, что соответствует среднему арифметическому от потерь или значению эмпирического риска.

Псевдокод алгоритма настройки параметров $\mathbf{w}$ на основе метода полного градиента – Алгоритм 2. Приведенный алгоритм не является оптимальным с вычислительной точки зрения, так как на каждом шаге итерации необходимо решать задачу на поиск минимума функции для вычисления значения агрегированного среднего значения. Поэтому построим другой итерационный алгоритм, который ищет значения $\mathbf{w}^*$ и $\mathsf{M}_p\{\ell_1(\mathbf{w}^*),\ldots,\ell_N(\mathbf{w}^*)\}$ одновременно.

Алгоритм стохастически усредненного градиента на базе агрегирующей функции

Поскольку градиент (4) является взвешенной суммой градиентов от соответствующих потерь, то можно применить метод, который лежит в основе алгоритма SAG (Stochastic Average Gradient) [10, 11]. Построим на основе этого метода алгоритм PBSAG – Penalty Based Stochastic Average Gradient – стохастически усредненного

Algorithm 2 Алгоритм полного градиента на базе агрегирующей функции.

$t \leftarrow 0$
Инициализировать $\mathbf{w}_0$
$u_0 \leftarrow \mathsf{M}_p\{\ell_1(\mathbf{w}_0),\ldots,\ell_N(\mathbf{w}_0)\}$
repeat
$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \mathrm{grad}\mathsf{M}_p\{\ell_1(\mathbf{w}_t),\ldots,\ell_N(\mathbf{w}_t)\}$
$u_{t+1} \leftarrow \mathsf{M}_p\{\ell_1(\mathbf{w}_{t+1}),\ldots,\ell_N(\mathbf{w}_{t+1})\}$
$t \leftarrow t+1$
until $\{u_t\}$ и $\{\mathbf{w}_t\}$ не стабилизируется

градиента на базе усредняющей верной агрегирующей функции. Схема адаптации параметров $\mathbf{w}$ и u имеет вид:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - h_t\overline{\mathbf{g}}_t,$$
$$u_{t+1} = u_t - \tau_t\overline{q}_t,$$

где

$$\overline{\mathbf{g}}_t = \frac{\sum\limits_{k=1}^{N}\mathbf{g}_{k,t}}{\sum\limits_{k=1}^{N}g_{k,t}}.$$

Значение $\overline{q}_t$ для поиска значения усредняющей агрегирующей функции может обновляться в соответствии с одним из следующих правил

$$\overline{q}_t = \frac{1}{N}\sum_{k=1}^{N}q_{k,t}$$

или

$$\overline{q}_t = \frac{\sum\limits_{k=1}^{N}q_{k,t}}{\sum\limits_{k=1}^{N}g_{k,t}}$$

в зависимости от того используется метод градиентного спуска или метод Ньютона для поиска минимального значения усредняющей агрегирующей функции M_p. Векторы из набора $\{\mathbf{g}_{k,t}\colon k=\overline{1,N}\}$ обновляются по следующему правилу:

$$\mathbf{g}_{k,t+1} = \begin{cases} -p''_{uz}\big(u_t,\ell_k(\mathbf{w}_t)\big)\mathrm{grad}\ell_k(\mathbf{w}_t), & \text{если } k=k(t) \\ \mathbf{g}_{k,t}, & \text{иначе.} \end{cases}$$

Значения из наборов $\{g_{k,t}\colon k=\overline{1,N}\}$ и $\{q_{k,t}\colon k=\overline{1,N}\}$ обновляются по следующим правилам:

$$g_{k,t+1} = \begin{cases} p''_{uu}\big(u_t,\ell_k(\mathbf{w}_t)\big), & \text{если } k=k(t) \\ g_{k,t}, & \text{иначе,} \end{cases}$$

$$q_{k,t+1} = \begin{cases} p'_{u}\big(u_t,\ell_k(\mathbf{w}_t)\big), & \text{если } k=k(t) \\ q_{k,t}, & \text{иначе.} \end{cases}$$

Algorithm 3 Алгоритм стохастически усредненного градиента на базе усредняющей агрегирующей функции.

$t \leftarrow 0$
Инициализировать $\mathbf{w}_0$
for $k \in \{1,\dots,N\}$ **do**
  $\mathbf{G}_k \leftarrow p''_{uz}\big(u_0, \ell_k(\mathbf{w}_0)\big)\operatorname{grad}\ell_k(\mathbf{w}_0)$
  $H_k \leftarrow p''_{uu}\big(u_0, \ell_k(\mathbf{w}_0)\big)$
  $Q_k \leftarrow p'_u\big(u_0, \ell_k(\mathbf{w}_0)\big)$
end for
$\mathbf{G} \leftarrow \mathbf{G}_1 + \cdots + \mathbf{G}_N$
$H \leftarrow H_1 + \cdots + H_N$
$Q \leftarrow Q_1 + \cdots + Q_N$
repeat
  $k = k(t)$
  $\mathbf{G} \leftarrow \mathbf{G} - \mathbf{G}_k + p''_{uz}\big(u_t, \ell_k(\mathbf{w}_t)\big)\operatorname{grad}\ell_k(\mathbf{w}_t)$
  $\mathbf{G}_k \leftarrow p''_{uz}\big(u_t, \ell_k(\mathbf{w}_t)\big)\operatorname{grad}\ell_k(\mathbf{w}_t)$
  $H \leftarrow H - H_k + p''_{uu}\big(u_t, \ell_k(\mathbf{w}_t)\big)$
  $H_k \leftarrow p''_{uu}\big(u_t, \ell_k(\mathbf{w}_t)\big)$
  $Q \leftarrow Q - Q_k + p'_u(u_t, \ell_k(\mathbf{w}_t))$
  $Q_k \leftarrow p'_u\big(u_t, \ell_k(\mathbf{w}_t)\big)$
  $\bar{\mathbf{g}} = \dfrac{\mathbf{G}_1}{G_2}$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t\bar{\mathbf{g}}$
  if используется схема Ньютона **then**
    $\bar{q} \leftarrow Q/G_2$
  else
    $\bar{q} \leftarrow Q/N$
  end if
  $u_{t+1} \leftarrow u_t - \tau_t\bar{q}$
  $t \leftarrow t+1$
until $\{u_t\}$ и $\{\mathbf{w}_t\}$ не стабилизируется

Алгоритму PBSAG на каждом шаге необходимо хранить по одному градиентному вектору и два значения на каждый пример из обучающего набора данных, т.е. $N(m+2)$ вещественных чисел, где m – ранг вектора параметров $\mathbf{w}$. Поэтому его следует применять, если есть память для хранения такого объема данных.

Нетрудно заметить, что если $p(u,z) = (u-z)^2/2$, то схема алгоритма PBSAG редуцируется к схеме алгоритма SAG:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - h_t\mathbf{g}_t,$$

где

$$\mathbf{g}_t = \frac{1}{N}\sum_{k=1}^{N}\mathbf{g}_{k,t},$$

$$\mathbf{g}_{k,t+1} = \begin{cases}\operatorname{grad}\ell_k(\mathbf{w}_t), & \text{если } k = k(t)\\ \mathbf{g}_{k,t}, & \text{иначе.}\end{cases}$$

Таким образом схема алгоритма PBSAG является естественным обобщением схемы алгоритма SAG [10, 11], когда для вычисления средних потерь используется усредняющая агрегирующая функция, основанная на штрафной, вместо среднего арифметического.

Примеры применения PBSAG

Рассмотрим применение PBSAG с использованием «апроксимированного» варианта медианы для построения робастного аналога SVM для решения задачи в условиях выбросов. Напомним, что

$$med\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N} |u-z_k|.$$

Непостредственно PBSAG применить нельзя. Однако ее можно заменить на суррогат медианы, который в определенном смысле эквивалентен ей.

$p_\alpha(z-u)$ определяет суррогат медианы, ассимптотически эквивалентный медиане, если для некоторого α^*:

- $\lim\limits_{\alpha\to\alpha^*} p_\alpha(u-z) = |u-z|$;
- $\lim\limits_{\alpha\to\alpha^*} p'_\alpha(u-z) = \operatorname{sign}(u-z)$;
- $\lim\limits_{\alpha\to\alpha^*} p''_\alpha(u-z) = \delta(u-z)$.

Рассмотрим пример:

$$p_\alpha(u-z) = |u-z| - \alpha\ln(\alpha+|u-z|) + \alpha\ln\alpha. \tag{5}$$

При $\alpha = 0$ суррогат медианы совпадает с обычной медианой. Заметим, что

$$p'_\alpha(u-z) = \frac{u-z}{\alpha+|u-z|}, \qquad p''_\alpha(u-z) = \frac{\alpha}{(\alpha+|u-z|)^2}.$$

Рассматриваем линейное разделение на 2 класса при помощи линейной функции $A(\mathbf{x},\mathbf{w})$ по правилу:

$$y(\mathbf{x}) = \begin{cases} 1, & \text{если } A(\mathbf{x},\mathbf{w}) > d \\ 0, & \text{если } |A(\mathbf{x},\mathbf{w})| \leqslant d \\ -1, & \text{если } A(\mathbf{x},\mathbf{w}) < -d. \end{cases}$$

В случае полного разделения $A(\widetilde{\mathbf{x}}_k,\mathbf{w})\tilde{y}_k > 0$ для всех точек $\widetilde{\mathbf{x}}_1,\dots,\widetilde{\mathbf{x}}_N$ с заданными метками классов $\tilde{y}_1,\dots,\tilde{y}_N$ $(\tilde{y}_k \in \{-1,1\})$.

Для разделения на 2 класса используем функции потерь как в методе SVM: $\ell_k(\mathbf{w}) = (1-m_k(\mathbf{w}))_+$, где $m_k(\mathbf{w}) = A(\widetilde{\mathbf{x}}_k,\mathbf{w})\tilde{y}_k$,

$$(S)_+ = \begin{cases} S, & \text{если } S > 0 \\ 0, & \text{иначе.} \end{cases}$$

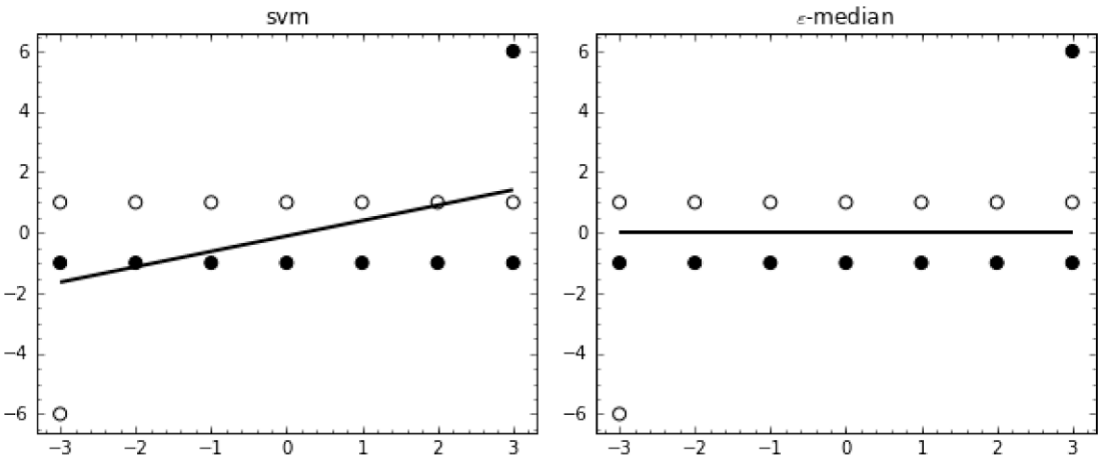

Рисунок. Примеры восстановления линейной разделяющей линии между двумя классами, содержащими выбросы: svm – при помощи SVM; ε-median – при помощи PBSAG.

Поиск по методу SVM здесь сводим к решению задачи минимизации функции:

$$\mathscr{E}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N}(1 - m_k(\mathbf{w}))_+.$$

Для сравнения рассматриваем задачу поиска разделяющей линии на базе минимизации функции:

$$\mathscr{E}(\mathbf{w}) = \mathrm{med}_\alpha\big\{(1 - m_k(\mathbf{w}))_+ \colon k = 1..N\big\}$$

при помощи алгоритма PBSAG при $\alpha = 0.01$. Результаты представлены на рисунке.

Данный пример демонстрирует робастность процедуры поиска путем минимизации med_α от потерь и способность алгоритма PBSAG находить решение.

Приведенный пример показывает способность метода минимизации среднего риска на базе усредняющего агрегирующего функционала, апроксимирующего медиану, и алгоритма PBSAG по разделению двух классов в случае, когда исходные данные содержат выбросы, которые не может преодолеть метод классификации на основе стандартного SVM.

Список литературы/References

[1] Vapnik V., *The Nature of Statistical Learning Theory. Information Science and Statistics*, Springer-Verlag, 2000.

[2] Rousseeuw P. J., “Least Median of Squares Regression”, *Journal of the American Statistical Association*, 1984, № 79, 871–880.

[3] Rousseeuw P. J., Leroy A. M., *Robust Regression and Outlier Detection*, John Wiley and Sons, New York, 1987.

[4] Mesiar R., Komornikova M., Kolesarova A., Calvo T., “Aggregation functions”, *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, eds. H. Bustince, F. Herrera, J. Montero, Springer, Berlin, Heidelberg, 2008.

[5] Grabich M., Marichal J.-L., Pap E., *Aggregation Functions.*, Series: Encyclopedia of Mathematics and its Applications,. V. 127, Cambridge University Press, 2009.

[6] Beliakov G. , Sola H., Calvo T. A, *Practical Guide to Averaging Functions*, Springer, 2016, 329 pp.

[7] Calvo T., Beliakov G., "Aggregation functions based on penalties", *Fuzzy Sets and Systems*, **161**:10 (2010), 1420-1436

[8] Shibzukhov Z. M., "Correct Aggregate Operations with Algorithms", *Pattern Recognition and Image Analysis*, **24**:3 (2014), 377–382.

[9] Shibzukhov Z. M., "Aggregation correct operations on algorithms", *Doklady Mathematics*, **91**:3 (2015), 391-393.

[10] Le Roux N., Schmidt M., Bach F. A, "Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets", 2012, arXiv: abs/1202.6258.

[11] Schmidt M., Le Roux N., Bach F., "Minimizing Finite Sums with the Stochastic Average Gradient", 2013, arXiv: abs/1309.2388.

[12] Shalev-Shwartz, Zhang T. Stochastic dual coordin ate ascent methods for regularized loss minimization, *Journal of Machine Learning Research*, 2013, № 14. 2013, 567–599.

Список литературы (ГОСТ)

[1] Vapnik V. The Nature of Statistical Learning Theory. Information Science and Statistics. 2000. Springer-Verlag.

[2] Rousseeuw P.J. Least Median of Squares Regression // Journal of the American Statistical Association. 1984. No.79. PP.871–880.

[3] Rousseeuw P.J., Leroy A.M. Robust Regression and Outlier Detection. New York:John Wiley and Sons, 1987.

[4] Mesiar R., Komornikova M., Kolesarova A., Calvo T. Aggregation functions: A revision. In H. Bustince, F. Herrera, J. Montero, editors, Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. Springer, Berlin, Heidelberg, 2008.

[5] Grabich M., Marichal J.-L., Pap E. Aggregation Functions. Series: Encyclopedia of Mathematics and its Applications, No.127. Cambridge University Press. 2009.

[6] Beliakov G. , Sola H., Calvo T. A Practical Guide to Averaging Functions. 2016. Springer. 329 p.

[7] Calvo T., Beliakov G. Aggregation functions based on penalties // Fuzzy Sets and Systems. 2010. Vol.161, No.10, PP.1420-1436.

[8] Shibzukhov Z.M. Correct Aggregate Operations with Algorithms // Pattern Recognition and Image Analysis. 2014. Vol. 24. No. 3. PP. 377–382.

[9] Shibzukhov Z.M. Aggregation correct operations on algorithms // Doklady Mathematics. 2015. Vol. 91. No. 3. PP. 391-393.

[10] Le Roux N., Schmidt M., Bach F. A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets. nips.org, 2012. http://arxiv.org/abs/1202.6258

[11] Schmidt M., Le Roux N., Bach F. Minimizing Finite Sums with the Stochastic Average Gradient. arXiv.org, 2013. http://arxiv.org/abs/1309.2388

[12] Shalev-Shwartz and T. Zhang. Stochastic dual coordin ate ascent methods for regularized loss minimization // Journal of Machine Learning Research 14. 2013. PP. 567–599.

Для цитирования: Шибзухов З. М., Казаков М. А. Алгоритм стохастического усредненного градиента на базе агрегирующих функции // *Вестник КРАУНЦ. Физ.-мат. науки.* 2016. № 4-1(16). С. 112-125. DOI: 10.18454/2079-6641-2016-16-4-1-112-125

For citation: Shibzukhov Z. M., Kazakov M. A. Stochastic gradient algorithm based on the average aggregate functions, *Vestnik KRAUNC. Fiz.-mat. nauki.* 2016, **16**: 4-1, 112-125. DOI: 10.18454/2079-6641-2016-16-4-1-112-125

Поступила в редакцию / Original article submitted: 15.11.2016