УДК 519.7

DOI: 10.15827/0236-235X.118.180-186

Дата подачи статьи: 06.10.16

2017. Т. 30. № 2. С. 180–186

ПРИНЦИП МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА НА ОСНОВЕ АГРЕГИРУЮЩИХ ФУНКЦИЙ СРЕДНИХ ПОТЕРЬ ДЛЯ РЕШЕНИЯ ЗАДАЧ РЕГРЕССИИ

З.М. Шибзухов, д.ф.-м.н., ведущий научный сотрудник, szport@gmail.ru;

Д.П. Димитриченко, к.т.н., старший научный сотрудник, dimdp@rambler.ru;

М.А. Казаков, младший научный сотрудник, f_wolfgang@mail.ru

*(Институт прикладной математики и автоматизации,
ул. Шортанова, 89а, г. Нальчик, 360000, Россия)*

В настоящей работе предлагается расширенный вариант принципа минимизации эмпирического риска для решения задачи регрессии.

Он строится на основе применения усредняющих агрегирующих функций для вычисления эмпирического риска вместо среднего арифметического. Это оправданно, если распределение потерь имеет выбросы или существенно искажено, отчего оценка риска как средних потерь с самого начала является смещенной. Поэтому в таких случаях при оптимизации параметров в задаче регрессии изначально следует использовать робастную оценку среднего риска.

Подобные оценки среднего риска можно построить, используя усредняющие агрегирующие функции, которые являются решением задачи минимизации штрафной функции за отклонение от своего среднего значения. Такой подход для представления агрегирующих функций среднего позволяет, с одной стороны, определить значительно более широкий класс функций среднего, а с другой, определить дифференцируемые функции среднего, которые аппроксимируют недифференцируемые функции среднего, такие как медиана или квантиль. В результате появляется возможность построить градиентные методы решения задачи регрессии, в определенном смысле аппроксимирующие робастные методы, такие как Least Median и Least Quantile.

В настоящей работе предлагается новая градиентная схема для решения задачи минимизации среднего риска. Она является аналогом схемы, применяемой в алгоритме SAG в случае, когда риск вычисляется при помощи среднего арифметического.

Приведен иллюстративный пример построения робастной процедуры оценки параметров в задаче линейной регрессии на базе использования усредняющей функции среднего, аппроксимирующей медиану.

Ключевые слова: агрегирующая функция, агрегирующая операция, эмпирический риск, регрессия, штрафная функция, процедура градиентного спуска.

Метод минимизации эмпирического риска [1] является признанным методом решения задач параметрической регрессии.

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 среднего значения, в то время как медиана позволяет оценивать эмпирическое среднее при наличии выбросов. Поэтому для построения параметрических регрессионных зависимостей также используются эмпирические оценки среднего при помощи медианы, несмотря на то, что использование медианы делает процедуру настройки параметров регрессионной зависимости более медленной.

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.

Классический метод эмпирического риска

Задача поиска параметрической регрессионной зависимости $y = f(\mathbf{x}, \mathbf{w})$ между входами $\mathbf{x}$ и скалярным выходом y является одной из классических задач машинного обучения. Имеются конечный набор входов $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_k : k = 1, \dots, N\}$ и набор известных значений на выходе: $\tilde{\mathbf{Y}} = \{\tilde{y}_k : k = 1, \dots, N\}$. Требуется найти такой набор параметров $\mathbf{w}^*$, при котором функция $f^*(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}^*)$ адекватно представляет зависимость между y и $\mathbf{x}$ на множестве $\tilde{\mathbf{X}}$.

В качестве меры адекватности f^* часто используют эмпирический риск. 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 $Q(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^N \ell_k(\mathbf{w})$, где $\ell_k(\mathbf{w}) = \ell_k(r_k(\mathbf{w}))$, где $\ell(r)$ – функция потерь; $r_k(\mathbf{w}) = r(f(\tilde{\mathbf{x}}_k, \mathbf{w}), \tilde{y}_k)$ – функция невязки между значением функции f и ожидаемым значением в k -й точке. Например:

- разность: $r(\mathbf{w}) = f(\mathbf{x}, \mathbf{w}) - y$;

- абсолютная разность: $r(\mathbf{w}) = |f(\mathbf{x}, \mathbf{w}) - y|$;
- несимметричная абсолютная разность: $|f(\mathbf{x}, \mathbf{w}) - y|_a$, где $|r|_a = \begin{cases} \alpha r, & \text{если } r \geq 0, \\ (\alpha - 1)r, & \text{если } r < 0; \end{cases}$
- относительная разность: $r(\mathbf{w}) = \frac{1}{|y|} |f(\mathbf{x}, \mathbf{w}) - y|$

при условии, что значения y отделены от нуля, или

$$r(\mathbf{w}) = \frac{1}{1 + |y|} |f(\mathbf{x}, \mathbf{w}) - y|.$$

Функция потерь – это неотрицательная функция, которая имеет единственный минимум, так что $\ell(0) = \min \ell(r) = 0$. Например:

- абсолютная: $\ell(r) = |r|$;
- квадратичная: $\ell(r) = r^2$;
- Хьюбера: $\ell(r) = \begin{cases} c(2|r| - c), & \text{если } |r| \geq c, \\ r^2, & \text{если } |r| < c; \end{cases}$
- Тьюки (для простоты приведена производная функции):

$$\ell'(r) = \begin{cases} r \left(1 - \left(\frac{r}{c} \right)^2 \right), & \text{если } |r| \leq c, \\ 0, & \text{если } |r| > c; \end{cases}$$

- несимметричная абсолютная:

$$\ell(r) = \begin{cases} \alpha r, & \text{если } r \geq c, \\ (\alpha - 1)r, & \text{если } r < c; \end{cases}$$

- несимметричная квадратичная:

$$\ell(r) = \begin{cases} \alpha r^2, & \text{если } r \geq c, \\ (1 - \alpha)r^2, & \text{если } r < c. \end{cases}$$

Здесь $c > 0$, $0 < \alpha < 1$.

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

Среднее арифметическое, медиана и квантили – примеры усредняющей агрегирующей функции, поэтому в общем случае средние потери можно вычислять при помощи усредняющих агрегирующих функций.

Усредняющие агрегирующие функции

Пусть $\mathbb{I} \subseteq \mathbb{R}$ – сегмент $\mathbb{R}$; $\mathbb{I}^*$ – множество всех конечных последовательностей $\{z_1, \dots, z_N\} \in \mathbb{I}$.

Определение. Агрегирующая функция – это отображение $M: \mathbb{I}^* \rightarrow \mathbb{I}$, которое удовлетворяет следующим требованиям:

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

Последнее требование – требование монотонности агрегирующей функции.

Агрегирующая функция M симметричная, если $M\{z_1, \dots, z_N\} = M\{z_{\pi(1)}, \dots, z_{\pi(N)}\}$ для любой перестановки π ряда чисел $1, \dots, N$.

Усредняющие агрегирующие функции, по определению, удовлетворяют дополнительному требованию $\min\{z_1, \dots, z_N\} \leq M\{z_1, \dots, z_N\} \leq \max\{z_1, \dots, z_N\}$.

Основные понятия и основные свойства агрегирующих функций подробно описаны в [2–4].

Существует универсальный способ определения усредняющих агрегирующих функций [5]. Для их определения используются штрафные функции.

Определение. Функция $P\{z_1, \dots, z_N, u\}$ является штрафной, если удовлетворяет следующим требованиям:

- $P\{z_1, \dots, z_N, u\} \geq 0$ 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 $M_{z_1, \dots, z_N} = \{u: P(z_1, \dots, z_N, u) = P_{\min}(z_1, \dots, z_N)\}$, где $P_{\min}(z_1, \dots, z_N) = \min_u P(z_1, \dots, z_N, u)$, является синглтоном или связным сегментом.

Всякую усредняющую агрегирующую функцию можно определить на основе некоторой штрафной функции P следующим образом: $M_P\{z_1, \dots, z_N\} = \arg \min_u P(z_1, \dots, z_N, u)$, если $M_{z_1, \dots, z_N}$ – синглетон и $M_P\{z_1, \dots, z_N\} = \frac{a+b}{2}$, если $M_{z_1, \dots, z_N}$ – 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 разнovidность штрафных функций, которые являются суммами функций несходства:

$$P(z_1, \dots, z_N, u) = \sum_{k=1}^N p(z_k, u), \quad (1)$$

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

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

- $p(z, u) = 0 \Leftrightarrow z = u$;
- $p(z_1, u) \geq p(z_2, u)$, когда $z_1 \geq z_2 \geq u$ или $z_1 \leq z_2 \leq u$.

Агрегирующую функцию, определенную на базе штрафной функции вида (1), будем обозначать M_p .

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

Уникальность минимума $P_{z_1, \dots, z_N}(u) = P(z_1, \dots, z_N, u)$ и монотонность $M_p\{z_1, \dots, z_N\}$ гарантированы, когда

$$p(z, u) = G(h(z) - h(u)), \quad (2)$$

где $G: \mathbb{R} \rightarrow \mathbb{R}$ – непрерывная строго выпуклая функция; $h(u)$ – строго монотонная функция [4, 5].

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

- Среднее арифметическое получается при

$$p(z, u) = (z_k - u)^2: M\{z_1, \dots, z_N\} = \arg \max_u \sum_{k=1}^N (z_k - u)^2.$$

- Медиана

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

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

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

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

- Квантиль α $Q_\alpha\{z_1, \dots, z_N\}$ получается при $p(z, u) = |z_k - u|_\alpha$:

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

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

- Экспектиль α

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

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

- Среднее по Колмогорову

$$M_g\{z_1, \dots, z_N\} = g^{-1}\left(\frac{1}{N} \sum_{k=1}^N g(z_k)\right) \quad \text{получается}$$

при $p(u, z_k) = (g(u) - g(z_k))^2$:

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

- Масштабированная медиана

$$\text{med}_g\{z_1, \dots, z_N\} = g^{-1}\left(\text{med}\{g(z_k) : k = 1, \dots, N\}\right)$$

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

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

Поиск значения $M_p\{z_1, \dots, z_N\}$ можно осуществлять методом полного градиента или методом Ньютона. В первом случае на каждом шаге текущая оценка искомого значения обновляется по следующему правилу: $u_{t+1} = u_t - h_t P'_u(u_t, z_1, \dots, z_N)$,

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

Во втором случае обновление осуществляется по правилу

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

$$\text{где } \frac{P'_u(u_t, z_1, \dots, z_N)}{P''_{uu}(u_t, z_1, \dots, z_N)} = \frac{\sum_{k=1}^N P'_u(z_k, u_t)}{\sum_{k=1}^N P''_{uu}(z_k, u_t)}.$$

Параметр темпа обучения h_t в этих методах может быть постпостоянным 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 [6, 7].

В первом случае обновление будет осуществляться по правилу $u_{t+1} = u_t - h_t \bar{g}_t$, где $\bar{g}_t = \frac{1}{N} \sum_{k=1}^N g_{t,k}$, $k(t)$ – номер случайно выбранного значения из $z_1, \dots, z_N$ на шаге t . При этом

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

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

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

Во втором случае обновление осуществляется по следующему правилу: $u_{t+1} = u_t - h_t \frac{G_t}{H_t}$, где

$$G_t = \sum_{k=1}^N G_{t,k} \quad \text{и} \quad H_t = \sum_{k=1}^N H_{t,k}.$$

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

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

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

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

Псевдокод алгоритма – Алг. 1.

Algorithm 1. Алгоритм типа SAG для вычисления значения $M_g\{z_1, \dots, z_N\}$.

Инициализировать u_0

$$\mathbf{g}_k \leftarrow P'(z_k, u_0), k = 1, \dots, N$$

$$G_k \leftarrow P'_u(z_k, u_0), k = 1, \dots, N$$

$$G \leftarrow G_1 + \dots + G_N$$

If используется схема Ньютона **then**

$$H_k \leftarrow p_{uu}''(z_k, u_0), k = 1, \dots, N,$$

$$H \leftarrow H_1 + \dots + H_N$$

end if

$$t \leftarrow 0$$

repeat

$$k \leftarrow k(t)$$

$$G \leftarrow G + p_u'(z_k, u_t) - G_k$$

$$G_k \leftarrow p_u'(z_k, u_t)$$

If используется схема Ньютона **then**

$$H \leftarrow H + p_{uu}''(z_k, u_t) - H_k$$

$$H_k \leftarrow p_{uu}''(z_k, u_t)$$

$$\bar{g} = \frac{G}{H}$$

else

$$\bar{g} = \frac{G}{N}$$

end if

$$u_{t+1} \leftarrow u_t - h_t \bar{g}$$

$$t \leftarrow t + 1$$

until значение u_t не стабилизируется.

От эмпирического риска к агрегированному

Усредняющие агрегирующие функции уже использовались в [8, 9] для построения функционалов потерь в контексте задачи построения операций над алгоритмами классификации и регрессии, которые сохраняют свойство корректности алгоритмов. Применим их теперь для оценки средних потерь: $\mathcal{Q}_p(\mathbf{w}) = M_p\{\ell_k(\mathbf{w}) : k = 1, \dots, N\}$, где усредняющая агрегирующая функция M_p определяется на основе штрафной функции вида (1):

$$M_p\{\ell_1(\mathbf{w}), \dots, \ell_N(\mathbf{w})\} = \arg \min_u \sum_{k=1}^N p(\ell_k(\mathbf{w}), u).$$

Оптимальный набор параметров $\mathbf{w}^*$ доставляет минимум $\mathcal{Q}_p(\mathbf{w})$: $\mathcal{Q}_p(\mathbf{w}^*) = \min_u \mathcal{Q}_p(\mathbf{w})$.

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

$$\frac{\partial M_p\{z_1, \dots, z_N\}}{\partial z_k} = - \frac{p_{uz}^*(z_k, \bar{z})}{\sum_{l=1}^N p_{uu}^*(z_k, \bar{z})},$$

где $\bar{z} = M_p\{z_1, \dots, z_N\}$.

Тогда $\text{grad } M_p\{\ell_1(\mathbf{w}), \dots, \ell_N(\mathbf{w})\} =$

$$= \frac{\sum_{k=1}^N -p_{uz}^*(\ell_k(\mathbf{w}), \bar{z}) \text{grad } \ell_k(\mathbf{w})}{\sum_{k=1}^N p_{uu}^*(\ell_k(\mathbf{w}), \bar{z})},$$

где $\bar{z} = M_p\{\ell_1(\mathbf{w}), \dots, \ell_N(\mathbf{w})\}$.

Поиск оптимального набора $\mathbf{w}$ можно осуществлять при помощи следующего варианта про-

цедуры полного градиента. Правило обновления вектора параметров имеет вид

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \text{grad } M_p\{\ell_1(\mathbf{w}_t), \dots, \ell_N(\mathbf{w}_t)\}.$$

Обновления вектора параметров осуществляются до тех пор, пока значения $\mathbf{w}_t$ и $M_p\{\ell_1(\mathbf{w}_{t+1}), \dots, \ell_N(\mathbf{w}_{t+1})\}$ не стабилизируются.

Если $p(z, u) = G(z - u)$ – частный случай (2), то $\text{grad } M_p\{\ell_1(\mathbf{w}), \dots, \ell_N(\mathbf{w})\} = \sum_{k=1}^N \alpha_k(\mathbf{w}) \text{grad } \ell_k(\mathbf{w})$,

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

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

Нетрудно заметить, 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 $\ell_1(\mathbf{w}) - \bar{z}, \dots, \ell_N(\mathbf{w}) - \bar{z}$ – отклонений между агрегированным средним от потерь и текущими потерями. Если $G(z - u) = (z - u)^2/2$, то $\alpha_k(\mathbf{w}) = 1/N$, что соответствует среднему арифметическому от потерь или значению эмпирического риска.

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

Algorithm PBFSG. Алгоритм полного градиентного спуска на базе агрегирующей функции

$$t \leftarrow 0$$

Инициализировать $\mathbf{w}_0$

$$u_0 \leftarrow M_p\{\ell_1(\mathbf{w}_0), \dots, \ell_N(\mathbf{w}_0)\}$$

repeat

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \text{grad } M_p\{\ell_1(\mathbf{w}_t), \dots, \ell_N(\mathbf{w}_t)\}$$

$$u_{t+1} \leftarrow M_p\{\ell_1(\mathbf{w}_{t+1}), \dots, \ell_N(\mathbf{w}_{t+1})\}$$

$$t \leftarrow t + 1$$

until $\{u\}$ и $\{\mathbf{w}_t\}$ не стабилизируются.

Алгоритм стохастического усредненного градиента на базе агрегирующих функций

Поскольку данный усредненный градиент является взвешенной суммой градиентов от соответствующих потерь, можно применить метод, лежащий в основе алгоритма SAG (Stochastic Average Gradient) [6, 7]. Построим на основе этого метода алгоритм PBSAG (Penalty Based Stochastic Average Gradient) стохастически усредненного градиента

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

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

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

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

$$\bar{q}_t = \frac{1}{N} \sum_{k=1}^N q_{k,t} \quad \text{или} \quad \bar{q}_t = \frac{\sum_{k=1}^N q_{k,t}}{\sum_{k=1}^N \mathbf{g}_{k,t}} \quad \text{в зависимости от используемого метода: градиентного спуска или Ньютона. Векторы из набора } \{\mathbf{g}_{k,t} : k = \overline{1, N}\}$$

обновляются по следующему правилу:

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

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

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

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

Algorithm PBSAG. Алгоритм стохастически усредненного градиента на базе усредняющей функции

$t \leftarrow 0$

Инициализировать $\mathbf{w}_0$

for $k \in \{1, \dots, N\}$ **do**

$$\mathbf{G}_k \leftarrow p''_{uz}(\ell_k(\mathbf{w}_0), u_0) \text{grad } \ell_k(\mathbf{w}_0)$$

$$H_k \leftarrow p''_{uu}(\ell_k(\mathbf{w}_0), u_0)$$

$$Q_k \leftarrow p'_u(\ell_k(\mathbf{w}_0), u_0)$$

end for

$$\mathbf{G} \leftarrow \mathbf{G}_1 + \dots + \mathbf{G}_N$$

$$H \leftarrow H_1 + \dots + H_N$$

$$Q \leftarrow Q_1 + \dots + Q_N$$

repeat

$$k = k(t)$$

$$\mathbf{G} \leftarrow \mathbf{G} - \mathbf{G}_k + p''_{uz}(\ell_k(\mathbf{w}_0), u_t) \text{grad } \ell_k(\mathbf{w}_t)$$

$$\mathbf{G}_k \leftarrow p''_{uz}(\ell_k(\mathbf{w}_0), u_t) \text{grad } \ell_k(\mathbf{w}_t)$$

$$H \leftarrow H - H_k + p''_{uu}(\ell_k(\mathbf{w}_0), u_t)$$

$$H_k \leftarrow p''_{uu}(\ell_k(\mathbf{w}_0), u_t)$$

$$Q \leftarrow Q - Q_k + p'_u(\ell_k(\mathbf{w}_0), u_t)$$

$$Q_k \leftarrow p'_u(\ell_k(\mathbf{w}_0), u_t)$$

$$\bar{\mathbf{g}} = \frac{\mathbf{G}_1}{\mathbf{G}_2}$$

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \bar{\mathbf{g}}$$

if используется схема Ньютона **then**

$$\bar{q} \leftarrow Q / \mathbf{G}_2$$

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 стабилизируются

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

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

$$\mathbf{w}_{t+1} = \mathbf{w}_t - h_t \mathbf{g}_t, \quad \text{где } \mathbf{g}_t = \frac{1}{N} \sum_{k=1}^N \mathbf{g}_{k,t},$$

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

Таким образом, схема алгоритма PBSAG является естественным обобщением схемы алгоритма SAG [6, 7], когда для вычисления средних потерь используется усредняющая агрегирующая функция, основанная на штрафной, вместо среднего арифметического.

Примеры применения PBSAG

Рассмотрим применение PBSAG с использованием аппроксимированного варианта LMS [10, 11] для поиска линейной регрессии в условиях выбросов. В стандартном алгоритме LMS ищется минимум медианы квадрата ошибки:

$$\mathcal{E}(\mathbf{w}) = \text{med} \left\{ (f(\mathbf{x}_k, \mathbf{w}) - y_k)^2 : k = 1, \dots, N \right\}.$$

PBSAG нельзя применить, когда M является медианой. Однако ее можно заменить на суррогат медианы, который асимптотически эквивалентен ей.

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

$$\lim_{\alpha \rightarrow \alpha^*} p_\alpha(z-u) = |z-u|,$$

$$\lim_{\alpha \rightarrow \alpha^*} p'_\alpha(z-u) = \text{sign}(z-u).$$

Рассмотрим пример:

$$p_\alpha(z-u) = \sqrt{\alpha^2 + |z-u|^2} - \alpha,$$

где $\alpha^* = 0$,

$$p'_\alpha(z-u) = \frac{z-u}{(\alpha^2 + |z-u|^2)^{1/2}},$$

$$p''_\alpha(u-z) = \frac{\alpha^2}{(\alpha^2 + |z-u|^2)^{3/2}}.$$

Соответствующую ей усредняющую агрегирующую функцию будем называть α -медианой:

$$\text{med}_\alpha \{z_1, \dots, z_N\} = \arg \min_u \sum_{k=1}^N \left(\sqrt{\alpha^2 + |z_k - u|^2} - \alpha \right).$$

Другой пример суррогата можно построить на основе следующей функции несходства:

$$p_\alpha(z-u) = |z-u| - \alpha \ln(\alpha + |z-u|) + \alpha \ln \alpha. \quad (3)$$

На рисунке представлены примеры применения алгоритма PBSAG. Они показывают способность метода и алгоритма PBSAG на базе усредняющего верного агрегирующего функционала, аппроксимирующего медиану (3), восстанавливать линейную регрессионную зависимость в случае, когда исходные данные содержат выбросы.

Заключение

В данной работе эти функции используются для оценки средних потерь, где усредняющая агрегиру-

ющая функция определяется на основе штрафной функции. В результате процедура градиентного спуска становится аналогичной процедуре поиска минимума взвешенного среднего от потерь с числовыми весами. Это позволяет построить алгоритм PBSAG – стохастически усредненного градиента на базе усредняющей верной агрегирующей функции. Предложенный алгоритм, реализующий метод минимизации эмпирического риска, позволяет справляться с задачей восстановления линейной регрессионной зависимости в случае, когда исходные данные содержат выбросы. Данное свойство алгоритма продемонстрировано на соответствующих примерах.

(Работа выполнена при поддержке гранта РФФИ № 15-01-03381 и гранта ОННТ РАН).

Литература

1. Vapnik V. The nature of statistical learning theory (Information Science and Statistics). Springer-Verlag. NY, 2000, 314 p.
2. Mesiar R., Komornikova M., Kolesarova A., Calvo T. Aggregation functions: a revision. In H. Bustince, F. Herrera, J. Montero, eds. (Fuzzy Sets and Their Extensions: Representation, Aggregation and Models). Springer, Berlin, Heidelberg, 2008.
3. Grabich M., Marichal J.-L., Pap E. Aggregation functions (Encyclopedia of Mathematics and its Applications), Cambridge Univ. Press, 2009, no. 127.
4. Beliakov G., Sola H., Calvo T. A practical guide to averaging functions. Springer, 2016, 329 p.

5. Calvo T., Beliakov G. Aggregation functions based on penalties. *Fuzzy Sets and Systems*. 2010, vol. 161, no. 10, pp. 1420–1436.
6. Le Roux N., Schmidt M., Bach F. A stochastic gradient method with an exponential convergence rate for finite training sets. 2012. URL: <http://arxiv.org/abs/1202.6258> (дата обращения: 05.10.2016).
7. Schmidt M., Le Roux N., Bach F. Minimizing finite sums with the stochastic average gradient. 2013. URL: <http://arxiv.org/abs/1309.2388> (дата обращения: 05.10.2016).
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. *Dokl. Math.* 2015, vol. 91, no. 3, pp. 391–393.
10. Rousseeuw P.J. Least median of squares regression. *Jour. of the American Statistical Association*, 1984, no. 79, pp. 871–880.
11. Rousseeuw P.J., Leroy A.M. *Robust regression and outlier detection*. NY, John Wiley and Sons, 1987.

Software & Systems
DOI: 10.15827/0236-235X.118.180-186

Received 06.10.16
2017, vol. 30, no. 2, pp. 180–186

THE EMPIRICAL RISK MINIMIZATION PRINCIPLE BASED ON AVERAGE LOSS AGGREGATING FUNCTIONS FOR REGRESSION PROBLEMS

Z.M. Shibzukhov¹, *Dr.Sc. (Physics and Mathematics), Leading Researcher, szport@gmail.ru*
D.P. Dimitrichenko¹, *Ph.D. (Engineering), Senior Researcher, dimdp@rambler.ru*
M.A. Kazakov¹, *Junior Researcher, f_wolfgang@mail.ru*

¹ *Institute of Applied Mathematics and Automation, Shortanova St. 89a, Nalchik, 360000, Russian Federation*

Abstract. The paper proposes an extended principle of empirical risk minimization to solve the regression problem. It is based on using aggregate functions instead of arithmetic mean to calculate risk. This can be justified if the loss distribution of emissions is significant or distorted, causing a shift in the risk assessment of the average loss from the very beginning. Therefore, in such cases, when optimizing characteristics in the regression problem the robust estimate of average value-at-risk should be initially used.

Such intermediate risk assessment can be constructed using avg functions, which are the solution to the problem of penalty function minimization in case of mean deviation. This approach allows, on one hand, to determine a much broader class of secondary functions, and, on the other hand, to determine the average differentiable functions that approximate the average non-differentiable functions, such as a median or quintile. As a result, it is possible to construct gradient methods for solving the regression problem that, in a sense, can approximate robust techniques such as Least Median and Least Quintile.

This paper proposes a new gradient scheme for solving the minimization problem of the intermediate risk. It is an analog of the used in the SAG algorithm circuit when the risk is calculated by arithmetic mean. An illustrative example presents the construction of robust procedures for characteristics assessment in a linear regression based on the use of the avg function, which approximates the median.

Keywords: aggregation function/operation, empirical risk, regression, penalty function, gradient descent procedure.

Acknowledgements. This work is supported by RFBR grant no. 15-01-03381 and RAS Onita grant.

References

1. Vapnik V. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer-Verlag Publ., 2000.
2. Mesiar R., Komornikova M., Kolesarova A., Calvo T. Aggregation functions: A revision. H. Bustince, F. Herrera, J. Montero (Eds.). *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. Springer, Berlin, Heidelberg Publ., 2008.
3. Grabich M., Marichal J.-L., Pap E. *Aggregation Functions*. Series: Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, 2009, no. 127.
4. Beliakov G., Sola H., Calvo T. *A Practical Guide to Averaging Functions*. 2016, Springer Publ., 329 p.
5. Calvo T., Beliakov G. Aggregation functions based on penalties. *Fuzzy Sets and Systems*. 2010, vol. 161, no. 10, pp. 1420–1436.
6. Le Roux N., Schmidt M., Bach F. *A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets*. nips.org, 2012. Available at: <http://arxiv.org/abs/1202.6258> (accessed October 5, 2016).
7. Schmidt M., Le Roux N., Bach F. *Minimizing Finite Sums with the Stochastic Average Gradient*. arXiv.org, 2013. Available at: <http://arxiv.org/abs/1309.2388> (accessed October 5, 2016).
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. *Dokl. Math.* 2015, vol. 91, no. 3, pp. 391–393.
10. Rousseeuw P.J. Least Median of Squares Regression. *Jour. of the American Statistical Association*. 1984, no. 79, pp. 871–880.
11. Rousseeuw P.J., Leroy A.M. *Robust Regression and Outlier Detection*. NY, John Wiley and Sons Publ., 1987.