ISSN 1064-5624, Doklady Mathematics, 2017, Vol. 96, No. 2, pp. 494–497. © Pleiades Publishing, Ltd., 2017.
Original Russian Text © Z.M. Shibzukhov, 2017, published in Doklady Akademii Nauk, 2017, Vol. 476, No. 5, pp. 495–499.

MATHEMATICS

On the Principle of Empirical Risk Minimization Based on Averaging Aggregation Functions

Z. M. Shibzukhov

Presented by Academician of the Russian Academy of Sciences Yu. I. Zhuravlev April 17, 2017

Received May 11, 2017

Abstract—An extended version of the principle of empirical risk minimization is proposed. It is based on the application of averaging aggregation functions, rather than arithmetic means, to compute empirical risk. This is justified if the distribution of losses has outliers or is substantially distorted, which results in that the risk estimate becomes biased from the very beginning. In this case, for optimizing parameters, a robust estimate of the mean risk should be used. Such estimates can be constructed by using averaging aggregation functions, which are the solutions of the problem of minimizing the function of penalty for deviation from the mean value. An iterative reweighting scheme for numerically solving the problem of empirical risk minimization is proposed. Illustrative examples of the construction of a robust procedure for estimating parameters in the linear regression problem and in the problem of linearly separating two classes based on the application of an averaging mean function, which replaces the α-quantile, are given.

DOI: 10.1134/S106456241705026X

INTRODUCTION

The solution of many recognition and prediction problems is based on the application of the principle of empirical risk minimization (ERM) [1], which is a kind of the general extremum principle for searching an optimal recognition algorithm [2]. This principle consists in minimizing the mean loss caused by the incorrect operation of a trained system on a given finite set of precedents. The value of empirical risk is estimated as the arithmetic mean of the loss:

$$\mathrm{ER}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N} L_k(\mathbf{w}), \tag{1}$$

where $L_k(\mathbf{w})$ is the loss function associated with the kth precedent. The sought parameter values $\mathbf{w}^*$ must minimize empirical risk:

$$\mathrm{ER}(\mathbf{w}^*) = \min_{\mathbf{w}} \mathrm{ER}(\mathbf{w}). \tag{2}$$

In the case where the empirical loss distribution has outliers, estimate (1) may become biased. The outliers may be caused by both distortions of the initial data and the inadequacy of the model. This difficulty might be overcome by using the weighted empirical risk

$$\mathrm{ER}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N} \nu_k L_k(\mathbf{w}). \tag{3}$$

However, the complexity of searching for adequate weights $\nu_1,\ldots,\nu_N$, which would compensate the contribution of outliers is comparable with that of searching for outliers themselves. For this reason, another approach is used, which is based on the application of empirical mean estimates stable with respect to outliers, such as medians [4, 5] or quantiles [6].

This paper considers a generalizing approach, in which mean loss is estimated by using arbitrary averaging aggregation functions (AFs) [7, 8]; such functions have already proved efficient in the construction of correcting operations on algorithms which preserve algorithm correctness [9, 10].

EMPIRICAL RISK AND AVERAGING AGGREGATION FUNCTIONS

Let M be any averaging AF. In the presence of outliers, M must be a stable estimate of the mean. The empirical risk is defined by

$$\mathrm{ER}(\mathbf{w}) = \mathrm{M}\{L_1(\mathbf{w}),\ldots,L_N(\mathbf{w})\}. \tag{4}$$

Consider the standard class of averaging AFs, which includes the overwhelming majority of the known averaging AFs. An important property of standard AFs is the

Moscow State Pedagogical University, Moscow, 119882 Russia

Institute of Applied Mathematics and Automation, Nalchik, 360000 Kabardino-Balkariya, Russia

e-mail: szport@gmail.com

possibility of determining $\operatorname{grad}\mathrm{M}\{z_1,\dots,z_N\}$ as an implicit function, which makes it possible to reduce searching for $\mathbf{w}^*$ which minimizes (4) to solving a system of equations or apply the method of descent along directions computed from its gradient.

Following [7, 8], we define standard AFs as follows. Consider functions of the form

$$P(z_1,\dots,z_N,u)=\sum_{k=1}^{N}\rho(z_k,u), \tag{5}$$

where $\rho(z,u)$ is the function dissimilarity (see [8] for the definition).

Let us define a function M_ρ based on a dissimilarity function ρ.

Definition. We set

$$\mathrm{M}_\rho\{z_1,\dots,z_N\}=\arg\min_u P(z_1,\dots,z_N,u) \tag{6}$$

if $\mathbf{M}_{z_1\dots z_N}=\{z\colon P(z_1,\dots,z_N,z)=\min_u P(z_1,\dots,z_N,u)\}$ is a singleton and

$$\mathrm{M}_\rho\{z_1,\dots,z_N\}=\frac{a+b}{2}$$

if $\mathbf{M}_{z_1\dots z_N}$ is an interval with endpoints a and b.

If $\rho(z,u)=g(h(z),h(u))$, where g is a nonnegative convex function and h is a monotone invertible function, then M_ρ is an AF [8]. If $g(z,z)=0$, then M_ρ is an averaging AF.

According to the likelihood maximum principle, an estimate of an empirical mean $z_1,\dots,z_N$ by using an averaging AF M_ρ is adequate if the loss value is distributed according to a law $p(z)\propto e^{-\rho(z,\overline{z})}$, where $\overline{z}=\mathrm{M}_\rho\{z_1,\dots,z_N\}$. In this sense, M_ρ can be called the *M-mean function*.

The class of standard averaging AFs is large enough for computing stable estimates of the mean value. For example, it includes

(i) the family of symmetric means

$$\mathrm{M}^\gamma\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^{N}|z_k-u|^{1+\gamma}, \tag{7}$$

where $0\le\gamma\le 1$; here, M^0 is the median function and M^1 is the arithmetic mean function;

(ii) the family of asymmetric means

$$\mathrm{M}^\gamma_\alpha\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^{N}|z_k-u|^{1+\gamma}_\alpha, \tag{8}$$

where $|u|^{1+\gamma}_\alpha=(\alpha-[u>0])u|u|^\gamma$, $0\le\gamma\le 1$; here, M^0_α is the α-quantile and M^1_α is the α-expectile;

(iii) the family of Kolmogorov-type symmetric means

$$\mathrm{M}^\gamma_g\{z_1,\dots,z_N\}=\arg\min_u\sum_{k=1}^{N}|g(z_k)-g(u)|^{1+\gamma},$$

where g is a self-inverse function and $0\le\gamma\le 1$; here, M^0_g is the scalable median

$$\operatorname{med} g\{z_1,\dots,z_N\}=g^{-1}(\operatorname{med}\{g(z_1),\dots,g(z_N)\}),$$

and M^1_g is the Kolmogorov mean

$$\mathrm{M}_g\{z_1,\dots,z_N\}=g^{-1}\left(\frac{g(z_1)+\dots+g(z_N)}{N}\right).$$

If the function ρ has partial derivatives up to the second order, then the mean value $\overline{z}=\mathrm{M}_\rho\{z_1,\dots,z_N\}$ is a solution of the equation

$$\sum_{k=1}^{N}\rho'_u(z_k,\overline{z})=0,$$

which can be regarded as the definition of the averaging AF as an implicit function. Therefore,

$$\operatorname{grad}\mathrm{ER}_\rho(\mathbf{w})=\sum_{k=1}^{N}\alpha_k(\mathbf{w},\overline{z})\operatorname{grad}L_k(\mathbf{w}),$$

where

$$\alpha_k(\mathbf{w},\overline{z})=\frac{-\rho''_{uz}(L_k(\mathbf{w}),\overline{z})}{\rho''_{uu}(L_1(\mathbf{w}),\overline{z})+\dots+\rho''_{uu}(L_N(\mathbf{w}),\overline{z})}. \tag{9}$$

Thus, the sought functions $\mathbf{w}^*$ and $\overline{z}^*=\mathrm{ER}_\rho(\mathbf{w}^*)$ are solutions of the system of equations

$$\begin{aligned}&\sum_{k=1}^{N}\rho'_u(L_k(\mathbf{w}),u)=0,\\ &\sum_{k=1}^{N}\alpha_k(\mathbf{w},u)\operatorname{grad}L_k(\mathbf{w})=0.\end{aligned} \tag{10}$$

THE CASE OF A MEDIAN AND A QUANTILE

If ρ does not have a second derivative, then the above approach cannot be applied directly. For example, in the case of an α-quantile $\rho_\alpha(z,u)=|z-u|_\alpha$ (which is the median for $\alpha=0.5$), instead of $|r|_\alpha$, we can consider the one-parameter family $\{\rho_{\alpha,\varepsilon}(r)\colon\varepsilon>0\}$ of functions satisfying the following conditions:

(i) for each $\varepsilon>0$, the derivatives $\rho'_{\alpha,\varepsilon}$ and $\rho''_{\alpha,\varepsilon}$ exist;

(ii) $\lim\limits_{\varepsilon\to 0}\rho_{\alpha,\varepsilon}(r)=|r|_\alpha$;

(iii) $\lim\limits_{\varepsilon\to 0}\rho'_{\alpha,\varepsilon}(r)=(1-\alpha)[r<0]+\alpha[r>0]$.

Examples of such functions $\rho_{\alpha,\varepsilon}$ are

(i) $\rho_{\alpha,\varepsilon}(r)=[|r|-\varepsilon\ln(\varepsilon+|r|)+\varepsilon\ln\varepsilon]_\alpha$;

Algorithm IRLAL

$t \leftarrow 0$

repeat

$u_t \leftarrow \mathrm{M}_\rho\{{}_1(\mathbf{w}_t), \ldots, {}_N(\mathbf{w}_t)\}$

$v_k = \alpha_k(\mathbf{w}_t, u_t),\ k = 1 \ldots, N$

$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k\, {}_N(\mathbf{w})$

$t \leftarrow t + 1$

until $\{u_t\}$ and $\{\mathbf{w}_t\}$ converge

Algorithm IRLS

$t \leftarrow 0$

repeat

$v_k = \varphi\{r_k(\mathbf{w}_t)\}, k = 1, \ldots, N$

$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k r_k^2(\mathbf{w})$

$t \leftarrow t + 1$

until $\{\mathrm{w}_t\}$ converge

Fig. 1. The algorithms IRLAL and IRLS.

(ii) $\rho_{\alpha,\varepsilon}(r) = [\sqrt{\varepsilon^2 + r^2} - \varepsilon]_\alpha$.

Here, $[S]_\alpha = (1-\alpha)[S < 0] + \alpha[S > 0]$.

Now, for sufficiently small ε_0, we can use the averaging AF $\mathrm{M}_{\rho_{\alpha,\varepsilon_0}}$ as a "substitute" for the α-quantile.

ALGORITHMS FOR SOLVING THE SYSTEM

The system of linear equations (10) can be solved by Seidel's method. In the framework of this method, the first and second equations can be solved by any known iteration algorithm. However, in solving the second equation, it is often necessary to calculate $\alpha_k(\mathbf{w}_t, u_t)$, which has high computational cost for large N. This can be overcome by introducing the new variables $v_k = \alpha_k(\mathbf{w}, u)$, $k = 1, \ldots, N$:

$$\sum_{k=1}^{N} \rho'_u(L_k(\mathbf{w}), u) = 0,$$

$$v_k - \alpha_k(\mathbf{w}, u) = 0, \quad k = 1, 2, \ldots, N,$$

$$\sum_{k=1}^{N} v_k \operatorname{grad} L_k(\mathbf{w}) = 0.$$

The application of Seidel's method to solve this nonlinear system of equations yields the IRLAL (Iteratively Reweighted Least Averaged Losses) algorithm, which is a version of the iterative reweighting scheme. This algorithm reduces solving the initial minimization problem (4) to solving a sequence of problems of minimizing the weighted empirical risk (3).

COMPARISON WITH THE M-METHOD

Search for $\mathbf{w}^*$ by minimizing (4) can be regarded as a generalization of the M-method. This method searches for an optimal set of parameters $\mathbf{w}^*$ is by minimizing the function

$$Q(\mathbf{w}) = \frac{1}{N} \sum_{k=1}^{N} g(r_k(\mathbf{w})), \tag{11}$$

where $r_k(\mathbf{w}) = |f(\mathbf{x}_k, \mathbf{w}) - y_k|$ (in the regression problem), $r_k(\mathbf{w}) = -y_k f(\mathbf{x}_k, \mathbf{w})$ (in the two-classification problem), or some other function and $g(r)$ is a nonnegative quasi-convex function having a unique minimum and such that $g(0) = 0$.

The minimum of (11) can also be sought by the well-known IRLS (Iteratively Reweighted Least Squares) algorithm (with weight function $\varphi(r) = \frac{g'(r)}{r}$), which we give here in order to compare it with the IRLAL algorithm.

The IRLAL algorithm differs from the IRLS algorithm in the method for recalculating the weights $v_1, \ldots, v_N$. In the algorithm,

$$v_k = \frac{\rho''_{uz}(r_k(\mathbf{w}), \bar{r}(\mathbf{w}))}{\sum_{p=1}^{N} \rho''_{uu}(r_p(\mathbf{w}), \bar{r}(\mathbf{w}))},$$

where

$$\bar{r}(\mathbf{w}) = \mathrm{M}_\rho\{r_1(\mathbf{w}), \ldots, r_N(\mathbf{w})\},$$

while in the IRLS algorithm, $v_k = \varphi(r_k(\mathbf{w}))$. Thus, in the former case, the weight depends on the deviation of $r_k(\mathbf{w})$ from the averaged value $\bar{r}(\mathbf{w})$, and in the latter, from the value $r_k(\mathbf{w})$ itself.

Note that if g is an invertible function, then the problem of minimizing $Q(\mathbf{w})$ is equivalent to the problem of minimizing the Kolmogorov mean of the squared error, that is,

$$\mathrm{M}_{[g]}\{r_1(\mathbf{w}), \ldots, r_N(\mathbf{w})\} = g^{-1}\left(\frac{1}{N}\sum_{k=1}^{N} g(r_k(\mathbf{w}))\right), \tag{12}$$

where $\rho(z, u) = (g(z) - g(u))^2$. Thus, in this case, the M-method is equivalent to the extension of the ERM method suggested here, while the mean loss is estimated by using the Kolmogorov mean with scaling function g. Therefore, the robustness of the M-method is directly related to that of (12).

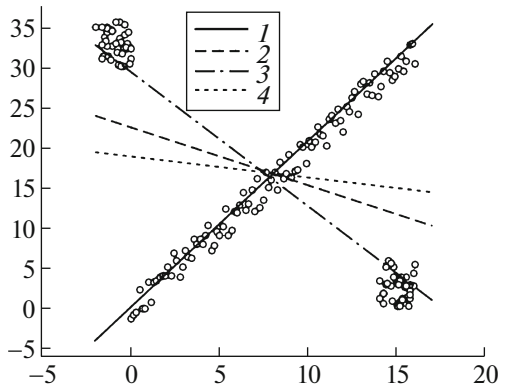

Fig. 2. Linear regression with 80% of outliers in data: the solutions are obtained (*1*) by minimizing the AF Mρ averaging the absolute error $\rho(z, u) = \sqrt{1 + (z-u)^2} - 1$ and by applying the M-method with function (*2*) r^2, (*3*) $|r|$, and (*4*) Tukey function.

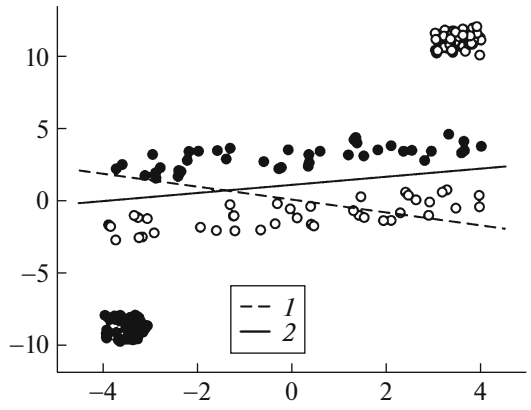

Fig. 3. Linear two-classification with 100% of outliers in data: the solution is obtained by the SVC method and by minimizing the AF Mρ averaging the hinge function, $\alpha = 0.45$.

MODEL EXAMPLES

Let us give an example of a linear regression problem and an example of a two-classification problem in which the data are selected so that the M-method surely fails. Computational results are presented in Fig. 2.

To the linear regression problem the M-method with the functions g: $|r|$, r^2, and the Tukey function $g(r) = \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{r}{c}\right)^2\right]\right)$ was also applied. The method with the averaging AF M_ρ, where $\rho(z, u) = \sqrt{1 + (z-u)^2} - 1$, and the error absolute value as the loss function was also considered. The data were composed for the model of linear regression $y = 2x$ with absolute noise ± 2 and outliers amounting to 80% of the amount of data without outliers.

To the problem of linear two-classification, the classical SVC (support vector classification) method with the best parameter values and the ERM method with the averaging AF M_ρ for the asymmetric dissimilarity function $\rho(z,u) = (\sqrt{(1 + (z-u)^2} - 1)_\alpha$, $\alpha = 0.45$, were applied. The data contained 100% of outliers (50% for each class).

The presented examples demonstrate that replacing the arithmetic mean by a more robust averaging function in estimating the mean loss makes it possible to overcome the problem of outliers.

CONCLUSION

The generalization of the ERM principle based of the application of standard averaging aggregation functions for averaging loss values, which is proposed in this paper, makes it possible to solve an ever-widening class of learning problems, especially in the presence of outliers. This is achieved by applying standard differentiable averaging AFs. For searching optimal parameters of the algorithms, the IRLAL algorithm of iterative reweighting is suggested.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 15-01-03381.

REFERENCES

1. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, Berlin, 2000).
2. Yu. I. Zhuravlev, in *Problems of Applied Mathematics and Mechanics* (Nauka, Moscow, 1971), pp. 67–74 [in Russian].
3. P. J. Huber, *Robust Statistics* (Wiley, New York, 1981; Mir, Moscow, 1984).
4. P. J. Rousseeuw, J. Am. Stat. Ass., No. 79, 871–880 (1984).
5. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection* (Wiley, New York, 1987).
6. Y. Ma, L. Li, X. Huang, and S. Wang, IFAC Proc. Vols, **44** (1), 11208–11213 (2011).
7. G. Beliakov, H. Sola, and T. Calvo, *A Practical Guide to Averaging Functions* (Springer-Verlag, Berlin, 2016).
8. T. Calvo and G. Beliakov, Fuzzy Sets Systems **161** (10), 1420–1436 (2010).
9. Z. M. Shibzukhov, Pat. Rec. Image Anal. **24** (3), 377–382 (2014).
10. Z. M. Shibzukhov, Dokl. Math. **91** (3), 391–393 (2015).

Translated by O. Sipacheva