Differential Leave-One-Out Cross-Validation for Feature Selection in Generalized Linear Dependence Models

Alexey Morozov[1], Brian Angulo[1], Vadim Mottl[2], Alexander Tatarchuk[2], Olga Krasotkina[3]

[1] Moscow Institute of Physics and Technology, Moscow 141701, Russia

[2] Computing Center of the Russian Academy of Sciences, Moscow 119991, Russia

[3] Markov Processes International, Summit NJ 07901, USA

Abstract. Estimation of dependences from empirical data in a growing class of models is inevitably concerned with choosing the value of a structural parameter responsible for the model's complexity. The most popular cross-validation schemes, in particular, leave-one-out, suffer from the necessity to multiply repeat the model estimation on different subsamples of the training set. In this paper, we propose the method of differential leave-one-out cross-validation for generalized linear models of arbitrary dependences, which allows for estimation of the model only once with each tentative value of the structural parameter. The idea of the method is that, instead of complete deleting an object from the training set at a single step of the training process, we delete only an infinitesimally small part of each of them. The indicator of the model quality is computed as the average of partial derivatives of the errors at each of single objects by the weights of their occurrence in the training set. The computing of the model quality indicator does not increase the computational complexity of the estimation procedure.

Keywords: Dependence estimation, Leave-one-out cross-validation, Feature selection.

1 INTRODUCTION

1.1 The generalized dependence estimation problem

The classical dependence estimation problem [1] consists in the requirement to recover the unknown dependence of a hidden variable $y\in\mathbb{Y}$ associated with any real-world object on the observable vector of its numerical features $\mathbf{x}=(x_1\cdots x_n)^T\in\mathbb{R}^n$, when only a training set of real-world objects is available

$$\{(\mathbf{x}_t, y_t), t=1,...,N\}, \mathbf{x}_t=(x_{t,1}\cdots x_{t,n})^T\in\mathbb{R}^n, y\in\mathbb{Y}. \quad (1)$$

In the general case, the target variable may take values from an arbitrary scale $y\in\mathbb{Y}$, but we consider here only the cases of numerical regression, when the target variable is a real number $y\in\mathbb{Y}=\mathbb{R}$, and two-class pattern recognition, when it takes one of two values $y\in\mathbb{Y}=\{-1,1\}$.

The size of the training set is defined by the number of objects N and the number of their numerical features n

$$\mathbf{X}=(\mathbf{x}_1\cdots\mathbf{x}_N)=\begin{pmatrix}\boldsymbol{x}_1^T\\ \vdots\\ \boldsymbol{x}_n^T\end{pmatrix}(n\times N),\ \mathbf{x}_t\in\mathbb{R}^n,$$
$$\mathbf{X}^T=(\boldsymbol{x}_1\cdots\boldsymbol{x}_n)=\begin{pmatrix}\mathbf{x}_1^T\\ \vdots\\ \mathbf{x}_N^T\end{pmatrix}(N\times n),\ \boldsymbol{x}_i\in\mathbb{R}^N. \quad (2)$$

In this paper, we keep to the generalized linear approach to dependence estimation [2], which we understand as a fusion of Vapnik's principle of regularized empirical risk minimization [1] and John Nelder's idea of Generalized Linear Models [3]. The latter was originally formulated within the bounds of the classical mathematical statistics. We harness the notion of Generalized Linear Model as a means to make the sought-for mathematical model of the dependence under study irrelevant of the specific scale of the goal variable $y\in\mathbb{Y}$.

Following [3], we treat here the mathematical model of a dependence of arbitrary kind as a pair

$$\begin{cases} z(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R} \ - \text{Generalized Linear Model}, \\ q(y,z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+ \ - \ \text{link function, convex in } z \in \mathbb{R}. \end{cases} \tag{3}$$

The real-valued variable $z(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x} \in \mathbb{R}$ will be called the generalized linear feature of the real-world object represented by its feature vector $\mathbf{x} \in \mathbb{R}^n$ relative to the hyperplane in the feature space $\{\mathbf{x}' \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x}' = 0\} \subset \mathbb{R}^n$ defined by the direction vector $\mathbf{a} \in \mathbb{R}^n$. The generalized linear feature has the sense of the positive or negative distance between the point $\mathbf{x}$ and the hyperplane with respect to its positive or negative side.

The link function $q(y,z)$ (loss function in Vapnik's terminology) is to be chosen by the observer and is meant to express his/her suggestion on how the nature would penalize the estimate of the unknown $y \in \mathbb{Y}$ for an object $\mathbf{x} \in \mathbb{R}^n$ represented by its generalized numerical linear feature $z(\mathbf{x} \mid \mathbf{a})$. Since the link function is chosen, the hyperplane direction vector $\mathbf{a} \in \mathbb{R}^n$ completely defines the decision rule:

$$\hat{y}(\mathbf{x}|\mathbf{a}) = \arg\min_{y \in \mathbb{Y}} q\big(y, z(\mathbf{x}|\mathbf{a})\big) = \arg\min_{y \in \mathbb{Y}} q(y, \mathbf{a}^T\mathbf{x}) \,. \tag{4}$$

Particular dependence estimation problems differ from each other only in the choice of the link function, specifically:

- for regression $y \in \mathbb{R}$, $q(y,z) = (y-z)^2$, $\hat{y}(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x}$; (5)
- for logistic regression pattern recognition $y = \pm 1$,

$$q(y,z) = \ln\big[1 + \exp(-yz)\big], \ \hat{y}(\mathbf{x}|\mathbf{a}) = \begin{cases} 1, \mathbf{a}^T\mathbf{x} \geq 0, \\ -1, \mathbf{a}^T\mathbf{x} < 0. \end{cases} \tag{6}$$

- for SVM pattern recognition $y = \pm 1$,

$$q(y,z) = \max(0, 1-yz), \ \hat{y}(\mathbf{x}|\mathbf{a},b) = \begin{cases} 1, \mathbf{a}^T\mathbf{x} \geq 0, \\ -1, \mathbf{a}^T\mathbf{x} < 0; \end{cases} \tag{7}$$

The link functions for these three typical cases are convex and continuous, for regression and logistic regression they are, in addition, smooth, and for SVM it is piecewise smooth. There exist simple expressions for derivatives and, almost everywhere (7), second derivatives

$$q'(y,z) = \big(\partial/\partial z\big) q(y,z) \,, \quad q''(y,z) = \big(\partial^2/\partial z^2\big) q(y,z) \,. \tag{8}$$

1.2 Regularized empirical risk minimization

From the viewpoint of the Generalized Linear Approach to dependence estimation, the quality of the hyperplane parameter $\mathbf{a} \in \mathbb{R}^n$ is the average value of the loss $q(y, \mathbf{a}^T\mathbf{x})$ over all the real-world objects $(\mathbf{x}, y) \in \mathbb{R}^n \times \mathbb{Y}$, which is usually called the average risk of error, let it be denoted as $AvR(\mathbf{a})$. However, average risk minimization $AvR(\mathbf{a}) \to \min$ is problematic because the hypothetical universe is unattainable to immediate observation.

Instead, it is commonly adopted to approximately estimate the average risk from the training set (1) as the arithmetic mean of the attainable loss values. This is the famous criterion of Empirical Risk minimization [1], which in our terms has the form [2]

$$EmpR(\mathbf{a}) = \frac{1}{N} \sum_{t=1}^{N} q(y_t, \mathbf{a}^T\mathbf{x}_t) \to \min\big(\mathbf{a} \in \mathbb{R}^n\big) \,. \tag{9}$$

This optimization problem is convex if the link function $q(y,z)$ is chosen as convex.

When the practical problem originates from a medical or industrial domain, the number of features often far dominates that of training objects $n \gg N$. If so, the problem of empirical risk minimization (9) becomes ill posed – there exist a continuum of models $\mathbf{a} \in \mathbb{R}^n$ that totally approximate the training data. The commonly adopted way to overcome the ambiguity is to heuristically add some additional independent convex regularization criterion

$$Reg(\mathbf{a}) \to \min\big(\mathbf{a} \in \mathbb{R}^n\big), \tag{10}$$

and to solve the two-criteria problem (9)-(10) from the principle of pareto optimality.

In this paper, we apply this principle in the form of regularized empirical risk minimization with the feature-selective ridge regularization

$$J(\mathbf{a}|\gamma,\mu)=\gamma\underbrace{\sum_{i=1}^{n}\begin{pmatrix}2\mu|a_i|,\ |a_i|\le\mu\\ \mu^2+a_i^2,|a_i|>\mu\end{pmatrix}}_{\text{feature-selective regularization}}+\sum_{t=1}^{N}q\left(y_t,\sum_{i=1}^{n}a_i x_{t,i}\right)\to\min(\mathbf{a}\in\mathbb{R}^n),\tag{11}$$

where the regularization parameters γ and μ are to be chosen from some outward considerations. Here $\gamma>0$ is the usual ridge coefficient. As to $\mu\ge 0$, this is the selectivity parameter first proposed in [4], which is to be adjusted to the training set (1) along with finding the direction vector of the generalized linear model of the dependence. It is clear that the regularization function is convex and smooth.

The crucial property of the selective regularization is that the direction vector estimated as the minimum point of (11)

$$\hat{\mathbf{a}}_{\gamma,\mu}=(\hat{a}_{\gamma,\mu,i},\ i=1,...,n)=\arg\min J(\mathbf{a}|\gamma,\mu)\tag{12}$$

will consist of elements some of which have exactly zero values. The subset of its $\hat{n}_{\gamma,\mu}\le n$ nonzero components

$$\hat{\mathbb{I}}_{\gamma,\mu}=\left\{i:|\hat{a}_{\gamma,\mu,i}|>0\right\}\subseteq\{1,...,n\},\ \hat{n}_{\hat{\mathbb{I}}_{\gamma,\mu}}=\left|\hat{\mathbb{I}}_{\gamma,\mu}\right|,\tag{13}$$

is the subset of active features. This is just the effect of feature selection.

If $\mu=0$, the regularization function in (11) coincides with the usual ridge regularization $\gamma\mathbf{a}^T\mathbf{a}+EmpR(\mathbf{a},b)\to\min$, and all the elements of the direction vector remain, generally speaking, be different from zero. But when the selectivity parameter grows $\mu>0$, the penalty $\mu|a_i|$ in (11) drives to zero those of the coefficients at redundant features which weakly contribute to diminishing of the empirical risk. Further growth of the selectivity parameter $\mu\to\infty$ results finally in complete zeroing of all the coefficients.

The essential necessity to find the Golden Mean makes unavoidable the choice of an instrument for adjusting the hyperparameters (γ,μ) of the training criterion (11), primarily, the selectivity μ.

1.3 Differential leave-one-out cross-validation versus traditional cross-validation techniques and other methods of hyperparameter optimization

The indisputable indicator of the quality of hyper-parameter values would be the average risk of error

$$AvR(\hat{\mathbf{a}}_{\gamma,\mu})=E\left\{q\left(y,\hat{\mathbf{a}}_{\gamma,\mu}^T\mathbf{x}\right)\right\}\to\min(\gamma,\mu)\tag{14}$$

measured over the entire set of all the feasible real-world objects $(y,\mathbf{x})\in\mathbb{Y}\times\mathbb{R}$. An estimate of such an indicator would require some assumptions on the family of probability distributions associated with the data source and parametrized by hyperparameters $f\left(\mathbf{y},\mathbf{X}|\gamma,\mu\right),\ \mathbf{y}=(y_1\cdots y_N),\ \mathbf{X}=\left(\mathbf{x}_1\cdots\mathbf{x}_N\right)$ (1).

If such a family of probability distributions is accepted, the most popular method is maximization of the Marginal Likelihood, or, what is the same, Model Evidence [5,6]. However, any probabilistic model is product of the observer's imagination, and the result of hyperparameter optimization will inevitably depend on these heuristics. The same relates to the class of information criteria, such as Akaike [7], Schwarz [8] and the deviance information criterion [9].

The principle which was designed to directly measure the reality is cross-validation. It immediately computes the estimate of the average risk as the arithmetic mean of the loss over the test set $q\left(y_t,\hat{\mathbf{a}}_{\gamma,\mu}^T\mathbf{x}_t\right),\ t\in\{Test\ set\}$, $\hat{\mathbf{a}}_{\gamma,\mu}^T\Leftarrow\{Training\ set\}$.

We prefer here the leave-one-out principle, whose essence is deleting the test objects from the training set one-by-one. Unfortunately, in the general case, it suffers from the necessity to repeatedly solve the problem when each of the training-set objects is left out.

In the rare particular case, when both link function $\left[q(y,z),\ z\in\mathbb{R}\right]$ (9) and regularization function $\left[Reg(\mathbf{a}),\ \mathbf{a}\in\mathbb{R}^n\right]$ (10) are quadratic, harnessing the well-known Sherman-Morrison-Woodbury inversion lemma [10,11] allows to compute the estimate of the direction vector for each j th left-out object $\hat{\mathbf{a}}_{\gamma,\mu}^{(j)}$ from the truncated training set $\left\{(\mathbf{x}_1,y_1),...,(\mathbf{x}_N,\ y_N)\right\}\setminus\left\{(\mathbf{x}_j,\ y_j)\right\}$ without repeatedly solving slightly different versions of the same quadratic optimization problems.

The problematic situation is that, at least, one of these functions or both of them are not quadratic in the majority of practical cases. The common idea is to consider the quadratic Taylor approximation of the training criterion and apply to it the Sherman-Morrison-Woodbury formula.

In [12], this principle is used to approximate the LOO cross-validation error in the case of ridge regression with generalized linear models. An analogous technique is applied in [13] to the problem of kernel-based logistic regression.

When the dimension of the feature space n exceeds the size of the training set N in (9)-(10), the computational complexity of solving the training problem in dual form is much lower than that of immediate solving it in primal formulation (11). Special versions of quadratic approximation to the leave-one-out estimator are presented for thus occasion in [14,15]. Probabilistic Bayesian aspects of dealing with low-rank high-dimensional data arising in this case are considered in [16]. However, any quadratic approximations of a leave-one-out estimator appear to be incorrect when a feature-selective regularizer occurs in the training criterion, like (11).

It should be kept in mind that the very idea of leave-one-out is based on the assumption that the training set is large enough, and the results of training little depends on the fact that one object is left out. But the presence of a feature-selective item in the training criterion makes it highly probable that the leaving out of an object results in changing the subset of active features, and the smooth approximation of the objective function found with the full training set will no longer be valid after its truncation. Thus, it is to be stated that the principle of approximate leave-one-out on the basis of quadratic Taylor expansion considered in above-cited papers is incorrect if the training criterion contains feature-selective items.

We studied this situation in paper [17]. The idea was to treat the subset of active features, $\hat{\mathbb{I}}_{\gamma,\mu}$ in terms of (11)-(13), as a separate hyperparameter to be verified via the leave-one-out principle.

In this paper, we consider an alternative way of reconciling the essential discontinuity of the leave-one-out principle of model verification with the smooth Taylor approximation of the training criterion. We propose the idea of a new differential leave-one-out cross-validation principle, when only a small part of an object is excluded from training, in contrast to the traditional discrete version of leave-one-out.

We exploit the same assumption as in [12-16], namely, that the link function of the sought-for dependence $q(y,z)$ (3) is convex and piecewise smooth. It is easy to see that the feature-selective regularization function in (11) is smooth function. Thus, the training criterion (11) is convex and smooth as a whole, so that it allows for exact quadratic representation as two terms of the Taylor expansion in the infinitesimally small vicinity of its minimum point $\hat{\mathbf{a}}_{\gamma,\mu}\in\mathbb{R}^n$. This fact invites for harnessing the Sherman-Morrison-Woodbury formula even for non-quadratic link functions. Moreover, the smoothness of the regularization function in (11) allows for small changes in the direction vector of the dependence model $\mathbf{a}=(a_1,...,a_n)$ without changings in the structure of the regularization penalty $|a_i|\gtrless\mu$.

2 THE SOLUTION OF THE GENERALIZED DEPENDENCE ESTIMATION PROBLEM

We shall consider an equivalent disjoint formulation of the problem (11) via the intermediate notion of generalized linear features of the training objects $z_t\in\mathbb{R}$ (3):

$$\begin{cases} \gamma\sum\limits_{i=1}^{n}\begin{pmatrix} 2\mu|a_i|, |a_i|\le\mu \\ \mu^2+a_i^2, |a_i|>\mu \end{pmatrix}+\sum\limits_{t=1}^{N}q\left(y_t,z_t\right)\to\min(a_1,...,a_n,z_1,...,z_N), \\ z_t=\mathbf{a}^T\mathbf{x}_t,\ t=1,...,N. \end{cases} \tag{15}$$

This convex programing problem with $n+N$ variables can be reformulated in dual form with respect to N Lagrange multipliers $\boldsymbol{\lambda}=(\lambda_1\cdots\lambda_N)\in\mathbb{R}^N$. It is especially important that the values of Lagrange multipliers immediately define the subset of active features (13) with respect to the symbol $\boldsymbol{x}_i\in\mathbb{R}^N$ (2):

$$\mathbb{I}(\boldsymbol{\lambda}|\mu)=\left\{i:\left|\mathbf{x}_i^T\boldsymbol{\lambda}\right|>\mu\right\}\subseteq\{1,...,n\}. \tag{16}$$

The dual form of the feature-selective dependence estimation criterion (15) is twice differentiable convex problem

$$W(\boldsymbol{\lambda}\,|\gamma,\mu)=\frac{1}{2}\boldsymbol{\lambda}^T\left(\sum_{i\in\mathbb{I}(\boldsymbol{\lambda}|\mu)}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{\lambda}+\sum_{t=1}^{N}\varphi_t(y_t,\lambda_t|\gamma)\to\min(\boldsymbol{\lambda}\in\mathbb{R}^N), \tag{17}$$

where $\varphi_t(y_t,\lambda_t|\gamma)$ are inverse link functions.

$$\varphi(y_t,\lambda_t\,|\,\gamma)=-\min_{z_t\in\mathbb{R}}\left(\frac{1}{2\gamma}q(y_t,z_t)+\lambda_t z_t\right),\ \lambda_t\in\mathbb{R}\,. \tag{18}$$

It is proved in [2] that each inverse link function is convex in $\lambda_t\in\mathbb{R}$ if the link function $q(y_t,z_t)$ (3) is convex in $z_t\in\mathbb{R}$.

The solution of the problem $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}=\arg\min W(\boldsymbol{\lambda}\,|\gamma,\mu)$ defines:

- each of n components of the direction vector $\hat{\mathbf{a}}_{\gamma,\mu}=\arg\min J(\mathbf{a}|\gamma,\mu)$ (11) independently of each other

$$\hat{a}_{\gamma,\mu,i}=\begin{cases}0, & \left|\boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right|\le\mu,\\ \boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}, & \left|\boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right|>\mu,\end{cases}\quad i=1,...,n\,; \tag{19}$$

- the subset of active features $\hat{\mathbb{I}}_{\gamma,\mu}=\left\{i:\left|\boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right|>\mu\right\}$ (16);
- the linear models of the objects $\hat{\mathbf{z}}_{\gamma,\mu}=(\hat{z}_{\gamma,\mu,1}\cdots\hat{z}_{\gamma,\mu,N})^T\in\mathbb{R}^N$ in the sought-for dependence (3):

$$\begin{aligned}&\hat{z}_{\gamma,\mu,t}=\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{t,i}\boldsymbol{x}_i^T\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu},\ t=1,...,N,\\ &\hat{\mathbf{z}}_{\gamma,\mu}=\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\in\mathbb{R}^N.\end{aligned} \tag{20}$$

The algorithm of numerically solving the dual problem (17) is built by the principle of Newton's iteration with variable step length. The essence of each iteration is quadratic Taylor representation of both convex summands in (17) in a small vicinity of the current vector of Lagrange multipliers $\boldsymbol{\lambda}^k$.

In the first summand, it is enough to fix the subset of active features

$$\boldsymbol{\lambda}^T\left(\sum_{i\in\mathbb{I}(\boldsymbol{\lambda}|\mu)}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{\lambda}\cong\boldsymbol{\lambda}^T\left(\sum_{i\in\hat{\mathbb{I}}^k}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{\lambda},\ \hat{\mathbb{I}}^k=\mathbb{I}(\boldsymbol{\lambda}^k|\,\mu)\,. \tag{21}$$

As to the second summand, it is proved in [2] that in the case of regression (5), when the link function is quadratic, the inverse link function is quadratic, too, and no additional Taylor expansion is required:

$$q(y_t,z_t)=(y-z)^2,\quad \varphi(y_t,\lambda_t\,|\,\gamma)=\frac{1}{2}\gamma\lambda_t^2-y_t\lambda_t\,. \tag{22}$$

In the case of logistic regression (6), it is enough to consider the Taylor expansion of the link function in a vicinity of the current point $\hat{z}_t^k=\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{t,i}\boldsymbol{x}_i\right)\hat{\boldsymbol{\lambda}}^k$ (20):

$$\begin{aligned}&q(y_t,z_t)=\ln\left[1+\exp(-y_t z_t)\right]\cong\hat{g}_t^k(\hat{y}_t^k-z_t)^2+const,\\ &\hat{g}_t^k=\frac{1}{2}\frac{\exp(-y_t\hat{z}_t^k)}{\left(1+\exp(-y_t\hat{z}_t^k)\right)^2},\quad \hat{y}_t^k=\hat{z}_t^k+y_t\left(1+\exp(-y_t\hat{z}_t^k)\right).\end{aligned} \tag{23}$$

Now, when the link function is approximated to be quadratic, the inverse link function for logistic regression is approximately quadratic like in (22) with respect to the current iteration:

$$\varphi(y_t,\lambda_t\,|\,\gamma)\cong\frac{1}{2}\gamma\frac{1}{\hat{g}_t^k}\lambda_t^2-\hat{y}_t^k\lambda_t\,. \tag{24}$$

Let the parameters of Taylor expansions (21)-(24) at the last iteration step be denoted as

$$\begin{cases}\hat{\mathbb{I}}_{\gamma,\mu}\subseteq\{1,...,n\},\\ \hat{\mathbf{G}}_{\gamma,\mu}=\begin{pmatrix}Diag(1\cdots1)-\text{regression}\\ Diag(\hat{g}_{\gamma,\mu,1}\cdots\hat{g}_{\gamma,\mu,N})-\text{logistic regression}\end{pmatrix}(N\times N),\\ \hat{\mathbf{y}}_{\gamma,\mu}=\begin{pmatrix}(y_1\cdots y_N)-\text{regression}\\ (\hat{y}_{\gamma,\mu,1}\cdots\hat{y}_{\gamma,\mu,N})-\text{logistic regression}\end{pmatrix}\in\mathbb{R}^N,\end{cases} \tag{25}$$

Then the solution of the selective dependence estimation problem (11) in the dual form (17) $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}=\arg\min W(\boldsymbol{\lambda}\,|\,\gamma,\mu)$ will be represented as the minimum point of the quadratic function

$$\hat{W}_{\gamma,\mu}(\boldsymbol{\lambda}\,|\,\gamma,\mu)=\frac{1}{2}\boldsymbol{\lambda}^T\left(\hat{\mathbf{G}}_{\gamma,\mu}\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)\boldsymbol{\lambda}-\hat{\mathbf{y}}_{\gamma,\mu}^T\hat{\mathbf{G}}_{\gamma,\mu}\boldsymbol{\lambda}\to\min(\boldsymbol{\lambda}\in\mathbb{R}^N). \tag{26}$$

The process of Newton's iterations [2] is required, actually, only for finding the "proper" parameters of Taylor expansion (25).

In this paper, for finding a computationally efficient technique of leave-one-out verification of feature selection, we exploit the fact that, in closed form, the condition (26) boils down to inversion of an $(N\times N)$ matrix:

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu}=\left(\hat{\mathbf{G}}_{\gamma,\mu}\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)^{-1}\hat{\mathbf{G}}_{\gamma,\mu}\hat{\mathbf{y}}_{\gamma,\mu}\ . \tag{27}$$

3 DIFFERENTIAL LEAVE-ONE-OUT CROSS-VALIDATION OF HYPERPARAMETERS

3.1 Weighted quadratic representation of the dependence estimation problem in a vicinity of its solution

The principle of differential leave-one-out cross-validation is based on a weighted formulation of the dependence estimation problem (15), in which the link (loss) functions of training-set objects are supplied with some weights $r_t q(y_t,z_t)$. The weights are initially equal to one $r_t=1$, but at each step of the mental leave-one-out procedure one of them slightly decreases $r_j=1-p$:

$$\text{Differential leave } j\text{th out: } r_t=1,\ t\neq j,\ \ r_j=1-p,\ p\to 0,$$

$$\begin{cases}\gamma\sum\limits_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum\limits_{t=1}^{N}q(y_t,z_t)-pq\left(y_j,z_j\right)\to\min(a_1,...,a_n,z_1,...,z_N),\\ z_t=\mathbf{a}^T\mathbf{x}_t,\ t=1,...,N.\end{cases} \tag{28}$$

We shall consider the diagonal matrix

$$\mathbf{R}_j(p)=Diag\left(r_1\cdots r_N\right),\ r_t=1,\ t\neq j,\ r_j=1-p,\ p\to 0\ . \tag{29}$$

Since $\mathbf{R}_j(p)$ differs little from the identity matrix $\mathbf{I}_N$, the product of any two matrices of respective dimensions $(?\times N)$ and $(N\times ?)$ remains practically the same if we insert $\mathbf{R}_j(p)$ between them $\mathbf{AB}\cong\mathbf{AR}_j(p)\mathbf{B}$.

It is easy to show that the weighted problem (28) in dual form is equivalent to insertion of $\mathbf{R}_j(p)$ after both matrices $\hat{\mathbf{G}}_{\gamma,\mu}$ in (26). Respectively, the solution of the weighted problem will be expressed by the transformed equality (27):

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}=\left(\hat{\mathbf{G}}_{\gamma,\mu}\mathbf{R}_j(p)\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)^{-1}\hat{\mathbf{G}}_{\gamma,\mu}\mathbf{R}_j(p)\hat{\mathbf{y}}_{\gamma,\mu}. \tag{30}$$

In these terms, the direction vector of the undisturbed model with $p=0$ will be $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,0)}=\hat{\boldsymbol{\lambda}}_{\gamma,\mu}$ (27). The effect of damping the j th training-set object is expected to expose itself by slight increase in the respective loss in accordance with (20):

$$q(y_j,\hat{z}_{\gamma,\mu,j}^{(j,p)})=q\left(y_j,\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}\right)\underset{\approx}{\geq}q\left(y_j,\left(\sum\nolimits_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,0)}\right).$$

(31)

Our desire is to compute the growth rates $(\partial/\partial p)q(y_j,\hat{z}_{\gamma,\mu,j}^{(j,p)})\geq 0$ for all the objects $j=1,...,N$ at the point of zero damping rate $p=0$:

$$\frac{\partial}{\partial p}q(y_j,\hat{z}_{\gamma,\mu,j}^{(j,p)})\bigg|_{p=0}=q'(y_j,\hat{z}_{\gamma,\mu,j})\left(\frac{\partial}{\partial p}\hat{z}_{\gamma,\mu,j}^{(j,p)}\right)_{p=0}. \tag{32}$$

Here $\hat{z}_{\gamma,\mu,j}^{(j,p)} = \left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}} x_{j,i}\boldsymbol{x}_i\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$ (20). To find these growth rates, we have first to express the vector of Lagrange multipliers as N different functions $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$ of the damping rate p for each of the training objects $j=1,...,N$ in accordance with the rule. The inverse matrix in this equality is continuous function of the variable p, therefore, it is required to find an analytical expression for it.

3.2 Inverting perturbed matrices via Sherman-Morrison-Woodbury formula

In accordance with notation (29), the matrices in (30), in their turn, are sums of matrices:

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)} = \left(\hat{\mathbf{G}}_{\gamma,\mu}\mathbf{R}_j(p)\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)^{-1}\hat{\mathbf{G}}_{\gamma,\mu}\mathbf{R}_j(p)\hat{\mathbf{y}}_{\gamma,\mu} = \\ \left[\left(\hat{\mathbf{G}}_{\gamma,\mu}\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)-p\mathbf{1}_j\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right]^{-1}\left(\tilde{\mathbf{G}}_{\gamma,\mu}\hat{\mathbf{y}}_{\gamma,\mu}-p\mathbf{1}_j\hat{y}_j\right). \tag{33}$$

Here column $\mathbf{1}_j\in\mathbb{R}^N$ is vector of all zeros except the j th element that equals 1. As to $\hat{y}_{\gamma,\mu,j}\in\mathbb{R}$, this is, in the case of logistic regression, the j th element of vector $\hat{\mathbf{y}}_{\gamma,\mu}\in\mathbb{R}^N$, and, in the case of regression, simply the value of the hidden variable at the j th object in the training set $y_t\in\mathbb{R}$ (1), (25).

Generally speaking, finding analytical expressions for $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$ as functions of the damping rate p for all the training objects $j=1,...,N$ requires inversion of N perturbed matrices (33). All of them contain the same matrix $\left(\hat{\mathbf{G}}_{\gamma,\mu}\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)$ already inverted as result dependence estimation (27), namely, at the last step of Newton's iteration procedure. As to the additional matrices $\mathbf{1}_j\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T$, each of them is 1-rank matrix $(N\times N)$.

This situation is a slight generalization of the condition of famous Sherman-Morrison-Woodbury formula [10,11]. The following theorem formulates the result of its application to (33) as an analytical closed formula for the dependence of $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}\in\mathbb{R}^N$ on p.

Theorem 1.

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)} = \hat{\boldsymbol{\lambda}}_{\gamma,\mu} - p\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j\frac{\hat{y}_{\gamma,\mu,j}-\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\hat{\boldsymbol{\lambda}}_{\gamma,\mu}}{1-p\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j}, \tag{34}$$

where matrix $\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}} = \left(\hat{\mathbf{G}}_{\gamma,\mu}\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)+\gamma\mathbf{I}_N\right)^{-1}(N\times N)$ is to be computed only once as result of solving the dependence estimation problem (27) ■.

Formula (34) immediately applies to the logistic regression data model (6), (23). In the case of regression model, $\hat{y}_{\gamma,\mu,j}$ is simply the value of the goal variable at the j th object $\hat{y}_{\gamma,\mu,j}=y_j$ (5).

3.3 The criterion of differential leave-one-out cross-validation of hyperparameters

To convert the idea of evaluating the loss growth rates at the training-set objects (32) into a practically applicable criterion of differential cross-validation, it remains only to find a formula for the derivatives $(\partial/\partial p)\hat{z}_{\gamma,\mu,j}^{(j,p)}\in\mathbb{R}$, $j=1,...,N$.

Theorem 2.

$$\left(\frac{\partial}{\partial p}\hat{z}_{\gamma,\mu,j}^{(j,p)}\right)_{p=0} = -\left(\hat{y}_j-\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right)\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j. \tag{35}$$

Now we are ready to finally formulate the general criterion of differential cross-validation with respect to (32):

$$DiffLOO(\gamma,\mu) = \frac{1}{N}\sum_{j=1}^{N}\left\{-q'(y_j,\hat{z}_{\gamma,\mu,j})\left(\hat{y}_j-\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right)\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j\right\}\to\min(\gamma,\mu), \tag{36}$$

In particular, with respect to (23) and (25), we have:

$$DiffLOO(\gamma,\mu) = \frac{1}{N}\sum_{j=1}^{N}\left\{(y_j-\hat{z}_{\gamma,\mu,j})^2\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j\right\}\to\min(\gamma,\mu)\ (\text{reg.}), \tag{37}$$

$$DiffLOO(\gamma,\mu)=\frac{1}{N}\sum_{j=1}^{N}\left\{y_j\frac{\exp(-y_t\hat{z}_{\gamma,\mu,j})}{1+\exp(-y_t\hat{z}_{\gamma,\mu,j})}(\tilde{y}_j-\hat{z}_{\gamma,\mu,j})\left(\sum_{i\in\hat{\mathbb{I}}_{\gamma,\mu}}x_{j,i}\boldsymbol{x}_i^T\right)\mathbf{D}_{\gamma,\hat{\mathbb{I}}_{\gamma,\mu}}\mathbf{1}_j\right\}\to\min(\gamma,\mu)\ (\text{log. reg.}). \tag{38}$$

4 EXPERIMENTAL STUDY OF THE HYPERPARAMETER OPTIMIZATION TECHNIQUE

4.1 The aim of the experimental study

The aim of the experimental study is to demonstrate the ability of the feature selective methodology coupled with the selectivity optimization via differential LOO cross validation to estimate extremely sparse dependence models. The necessity of such models ensues from practical problems naturally complied with the assumption that coefficients at features in the generalized linear model differ from zero only within a really existing small subset of a large universe of available features, and the search for this subset (Factor Search) is the main aim of data processing.

As two glowing examples of such practical problems, we consider here the problem of recovering the hidden composition of an investment portfolio represented by a time series of its periodic returns (values of relative profitability) and the problem of gene expression pattern recognition.

4.2 Factor search in returns-based analysis of investment portfolios

As the basis for creation of an artificial portfolio we used an actual data set of monthly returns of $n=650$ stock market indexes during more than 20 years $N=251$:

$$\mathbf{X}=\underbrace{(\mathbf{x}_1\cdots\mathbf{x}_N)}_{N\text{ time series}},\ \mathbf{x}_t=\underbrace{(x_{t,1}\cdots x_{t,n})}_{\text{returns of }n\text{ securities}}. \tag{39}$$

The eigenvalues of the inner product matrix $\mathbf{X}\mathbf{X}^T$, having been placed in decreasing ordered, quickly fall $\zeta_1=16369.8,\ \zeta_2=1435.2,\ \zeta_3=1076.2,\ \zeta_{20}=134<0.01\zeta_1,\ \zeta_{50}=32.2<0.002\zeta_1$. This fact is the evidence of drastic correlation between the indexes (features).

We chose a mental portfolio composition $(a_i^*, i\in\mathbb{I}^*, |\mathbb{I}^*|=13)$ which consists of 13 active securities with equal capital distribution

$$\mathbb{I}^*\subseteq\mathbb{I}=\{i=1,...,n\},\ a_i^*=\begin{cases}0, & i\notin\mathbb{I}^*,\\ 1/13=const, & i\in\mathbb{I}^*.\end{cases} \tag{40}$$

The returns $\mathbf{y}=(y_1\cdots y_N)^T\in\mathbb{R}^N$ of the respective hypothetical portfolio (regressands in the training set) were generated as independent zero-mean Gaussian random values $y_t=\sum_{i\in\mathbb{I}^*}a_i^*x_{t,i}+\xi_t,\ t=1,...,N$ with 5% noise variance.

We set the portions of capital invested in the assumed active assets (regression coefficients in the hidden model of the dependence) to be equal to each other (40), but this fact is considered as unknown at the stage of feature-selective regression estimation, because the criterion (11) does not contain the condition $(a_i^*=const,\ i\in\mathbb{I}^*)$.

To the semi-real data set $\{(\mathbf{x}_t,y_t),t=1,...,N\}$, we multiply applied the feature-selective regression estimation procedure (17) according to the regularization path algorithm with fixed value of the ridge coefficient $\gamma=0.1$ and logarithmically growing values of the selectivity coefficient.

Fig. 1 shows the experimental plot of the differential LOO criterion $DiffLOO(\gamma,\mu)$ as function of selectivity μ (37). The solution of the selective regression estimation problem $\hat{\boldsymbol{\lambda}}_{\gamma,\hat{\mu}}$ with the experimentally optimal selectivity (37) $\hat{\mu}=\arg\min_{\mu}DiffLOO(\gamma,\mu)$ indicates the estimated portfolio of $\hat{n}=16$ securities instead of $n^*=13$ in the simulated portfolio. The compositions of both actual (simulated) portfolio and estimated one as numbers of securities in the initial data set are shown in Table 1.

As it is seen, the estimated capital allocation is quite close to the assumed one. It consists of 16 securities, 13 of which coincide with the actual ones. The 3 redundant securities have obtained relatively small weights, they are marked in Table 1 as gray.

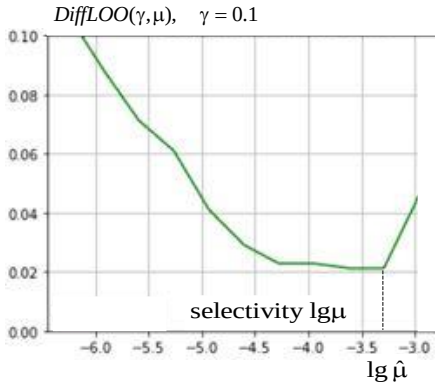

Figure 1. Experimentally optimal value of selectivity (regression).

Asset No in data set	Actual weight		Estimated weight	Asset No in data set
92	0.076		0.0723	92
			0.0109	155
164	0.076		0.0507	164
184	0.076		0.0924	184
243			0.0724	243
263	0.076		0.0753	263
332	0.076		0.0615	332
347	0.076		0.0615	347
			0.0225	399
412	0.076		0.0459	412
421	0.076		0.0773	421
			-0.0117	430
507	0.076		0.0720	507
522	0.076		0.0765	522
526	0.076		0.0753	526

Table 1. The actual and the estimated portfolio. The redundant securities are marked as gray.

Regularization path through the selectivity axis. The path consists in consecutive application of the iterative algorithm of selective regression estimation $\hat{\boldsymbol{\lambda}}_{\gamma,\mu_l}$ for each μ_l (17), starting with the previous solution $\hat{\boldsymbol{\lambda}}_{\gamma,\mu_{l-1}}$. At each step, we registered the number of active features $\hat{n}_{\mu_l}$ and the number of iterations. The result is shown in Figure 2.

The decrease of the number of active features is not surprising, it is just this what was to be expected, however, this process is not completely uniform. But the fact that the number of iterations at the slowly diminishing tentative values of the selectivity parameter remains minimally small is very notable. At most points of the selectivity axis the stopping condition was achieved after one iteration, and only at a few points 2-3 iterations were required.

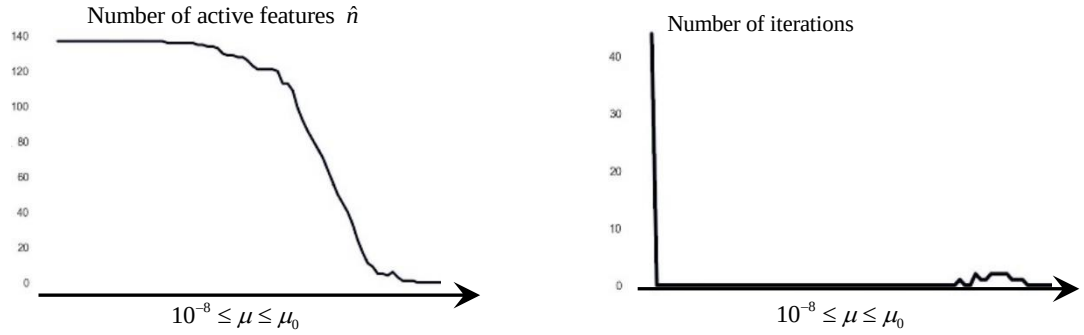

Figure 2. Regularization path through the selectivity axis.

4.3 Pattern recognition and factor search in gene expression data

For pattern recognition experiment, we use the leukemia gene expression data set [18]. This is a standard benchmark for binary classification publicly available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#leukemia .

The data set contains gene expressions samples from bone marrow and peripheral blood $\mathbf{x}_t \in \mathbb{R}^n$ measured over $n = 7129$ genes and corresponding to acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML). The dataset consisted of $N = 72$ samples marked by target variable $y_t \in \{-1,1\}$ – 49 samples of ALL $y_t \in 1$ and 23 samples of AML $y_t \in -1$. So, we have the training set $\mathbf{X}=(\mathbf{x}_1 \cdots \mathbf{x}_N)$ $(n \times N)$, $\mathbf{y} \in \{-1,1\}^N$ (1)-(2), in which the number of features $n = 7129$ far exceeds the number of samples $N = 72$.

To this data set, we multiply applied the feature-selective procedure of logistic regression pattern recognition learning (6) according to the regularization path algorithm with fixed value of the ridge coefficient $\gamma = 0.1$ and logarithmically growing values of the selectivity coefficient μ.

The experimental plot of the differential LOO criterion $DiffLOO(\gamma,\mu)$ as function of selectivity μ (37) is shown in Fig. 3.The solution of the selective logistic regression pattern recognition problem $\hat{\lambda}_{\gamma,\hat{\mu}}$ with the experimentally optimal selectivity (37) $\hat{\mu} = \arg\min_{\mu} DiffLOO(\gamma,\mu)$ indicates that there exist $n^* = 3$ features among the initial number $n = 7129$ that are sufficient for separation of the training set with only one wrongly classified sample. As we see, the process of factor search is not completely monotonic.

Regularization path through the selectivity axis. Consecutive application of the iterative algorithm of selective logistic regression pattern recognition learning $\hat{\lambda}_{\gamma,\mu_l}$ for each μ_l (17), starting with the previous solution $\hat{\lambda}_{\gamma,\mu_{l-1}}$, has led to the result shown in Fig. 4. This figure confirms once again that the decrease of the number of active features may violate the general monotonous tendency.

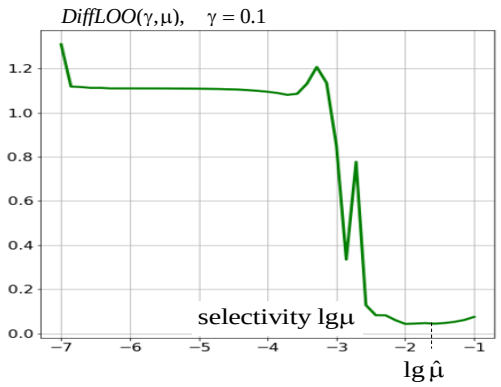

Figure 3. Experimentally optimal value of selectivity (logistic regression pattern recognition).

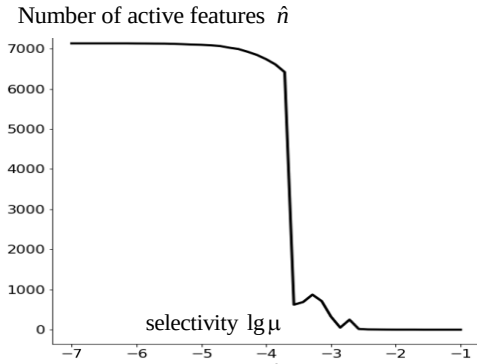

Figure 4. Regularization path through the selectivity axis in the pattern recognition experiment.

5 CONCLUSIONS

Cross-validation is the only absolutely indisputable method of measuring the generalization performance of training – the dependence model is estimated from the given data and tested on another one. This principle does not depend on any probabilistic assumptions on the data source.

We prefer the leave-one-out cross-validation. The disadvantage is extremely high computational complexity, since training must be carried out as many times as there are objects in the training set. An exception is the quadratic learning criterion, when the Woodbury formula can be applied without repeating the training.

However, this is possible only in the quadratic regression problem, in other kinds of dependence models the learning criterion is not quadratic. In addition, the selective regularization is assumed, even in the quadratic regression problem, removing one entity may result in a change in the subset of active features, then the leave-one-out method will become incorrect in principle. Our idea of differential leave-one-out cross-validation allows to overcome this obstacle, since the removal of an infinitesimally small part of an object does not change the active feature subset.

For the principle of differential leave-one-out would be applicable, it is enough that the learning criterion is quadratic in a small vicinity of the solution point. In its turn, the smoothness of the criterion is sufficient, i.e. its twice differentiability almost everywhere. All generalized linear dependence models considered in the paper correspond to this requirement.

ACKNOWLEDGEMENTS

This work is supported by the Russian Foundation for Basic Research, Grants No. 20-07-00382 (Mottl), 19-37-90159 (Morozov) and 17-07-00993 (Krasotkina).

REFERENCES

1. V. Vapnik. Estimation of Dependences Based on Empirical Data. Springer Verlag New York, 1982.
2. V. Mottl, V. Sulimova, O. Krasotkina, A. Morozov, A. Tatarchuk, I. Pugach. Computational Complexity of Dependence Estimation via Generalized Linear Models in Multidimensional Feature Spaces. 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russia, 2019, pp. 0719-0724. https://doi.org/10.1109/SIBIRCON48586.2019.8958417
3. J. Nelder, R. Wedderburn, Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General). Vol. 135, Issue 3, 1972, pp. 370-384.
4. A. Tatarchuk,, V. Mottl, A. Eliseyev, D. Windridge. Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective Support Vector Machines. Proceedings of the 19th International Conference on Pattern Recognition ICPR-2008. Vol 1-6, pp. 2336-2339.
5. D.J.C. MacKay. Hyperparameters: Optimize, or integrate out? In: G.B. Heidbreder (ed.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, 1996, pp. 43-59.
6. C. Bishop. Pattern Recognition and Machine Learning. Springer, 2006, 738 p.
7. H. Akaike. Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory. (Eds B. N. Petrov & F. Csaki.), Akademiai Ki-ado, Budapest, 1973, pp. 267-281.
8. G. Schwarz. Estimating the Dimension of a Model. Annals of Statistics, 1978,6, pp. 461-464.
9. D.J. Spiegelhalter. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, 2002, B 64, Part 4, pp. 583–639.
10. M.A. Goldberg, H.A. Cho. Introduction to Regression Analysis. WIT Press, 2004.
11. J. Thorsten. Training linear SVMs in linear time. Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA, August, 2006. ACM, New York, pp. 217–226.
12. R.J. Meijer, J.J. Goeman. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. Biometrical Journal, March 2013, Vol 55, No. 2, pp. 141-155.
13. G.C. Cawley, N.L.C. Talbot. Efficient Approximate Leave-One-Out Cross-Validation for Kernel Logistic Regression. Machine Learning, 2008, 71, pp. 243–264.
14. S. Wang, W. Zhou, H, Lu, A. Maleki, V. Mirrokni. Approximate Leave-One-Out for Fast Parameter Tuning in High Dimensions. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018, Vol. 80, pp. 5228-5237.
15. W.T. Stephenson, T. Broderick. Sparse approximate cross-validation for high-dimensional GLMs. Preprint on arXiv:1905.13657 [stat.ML]. 2019 (modified: 01 Jun 2020).
16. B. Trippe, J.H. Huggins, R. Agrawal, and T. Broderick. LR-GLM: High-dimensional Bayesian inference using low-rank data approximations. Proceedings of the 36th International Conference on Machine Learning, 2019, Vol. 97, pp. 6315-6324.
17. E. Chernousova, N. Razin, O. Krasotkina, V. Mottl, D. Windridge. Linear regression via elastic net: non-enumerative leave-one-out verification of feature selection. Springer optimization and its applications, Vol. 92, 2014, pp. 377-390.
18. T. Golub, D. Slonim K, P. Tamayo, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531-537.

Authors' background

Your Name	Title*	Research Field	Personal website
Alexey Morozov	**Phd candidate**	**Machine learning, optimization, model selection**	**https://amrzv.github.io/**
Brian Angulo	**master student**	**Model verification, robotics, optimization**	
Vadim Mottl	**full professor**	**Generalized linear models, model selection, markov models**	**http://www.machinelearning.ru/wiki/index.php?title=%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA:Vmottl**
Alexander Tatarchuk	**Phd**	**Spam detection, optimization, feature selection**	
Olga Krasotkina	**Phd**	**Financial data analysis, multivariable time-series, optimization**	