My first scientific paper

Week 3

State the problem

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Convey your message to the audience

You walk in the street.

Which poster do you pick?

The simplest problem statement in machine learning

$$\begin{tikzcd}
(-\lambda^2 \nabla_{\mathbf{w}} S + \mathbf{w}) \arrow[r, "T"] & \hat{\mathbf{w}}, \mathbf{x} \arrow[r, "f"] & \hat{y} \arrow[r] & S(\mathbf{w} | \mathrm{y}, \hat{\mathrm{y}}) \arrow[lll, bend right] \\
t \arrow[r] & i \arrow[r] & (\mathbf{x}, y) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

f is the forecasting model,

S is the criterion,

T is an optimization algorithm,

$\hat{\mathbf{w}}$ is some solution,

$$\hat{\mathbf{w}} = \arg\min S(\mathbf{w} | y, f).$$

Problem statement for machine learning

Formal problem statement, an analyst has to set

1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

$$\{\textbf{models} \times \textbf{data sets} \times \textbf{quality critea}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

Analyst creates an optimal model for expert to put it to operation

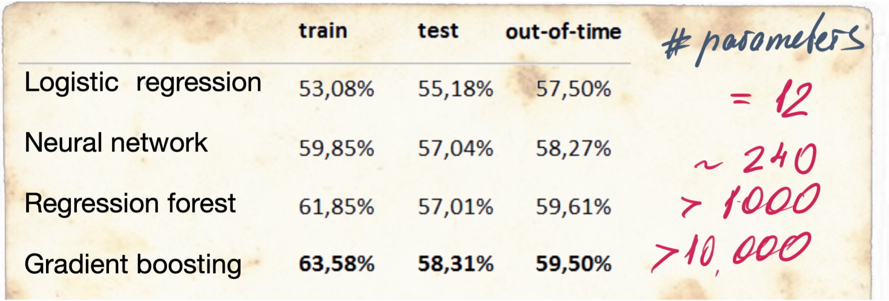
Quality criteria

- Accuracy: MAPE, AUC, F1 score
- Stability: forecasting variance, failure rate, parameter variance
- Complexity: number of parameters, Kolmogorov complexity

Origins of quality criteria

1. Theory: statistical hypotheses of data generation, algebraic structures of data, models of measurement
2. Computations: a criterion is useful to an optimisation procedure
3. Deployment: revenue, loss, failure rate

Significant increase in complexity and modest increase in accuracy



	train	test	out-of-time	# parameters
Logistic regression	53,08%	55,18%	57,50%	= 12
Neural network	59,85%	57,04%	58,27%	~ 240
Regression forest	61,85%	57,01%	59,61%	> 1000
Gradient boosting	**63,58%**	**58,31%**	**59,50%**	>10,000

Model selection is an important problem!

... it was a banking credit scoring model

Katrutsa A.M., Strijov V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 : 172-183.

Stresstest procedures for feature selection algorithms*

A. M. Katrutsa[1,2] and V. V. Strijov[1]

[1] *Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny city, 141700, Russian Federation*

[2] *Skolkovo Institute of Science and Technology, Novaya St., 100, Karakorum Building, 4th floor, Skolkovo, 143025, Russian Federation*

Abstract

This study investigates the multicollinearity problem and the performance of feature selection methods in case of datasets have multicollinear features. We propose a stresstest procedure for a set of feature selection methods. This procedure generates test data sets with various configurations of the target vector and features. A number of some multicollinear features are inserted in every configuration. A feature selection method results a set of selected features for given test data set. To compare given feature selection methods the procedure uses several quality measures. A criterion of the selected features redundancy is proposed. This criterion estimates number of multicollinear features among the selected ones. To detect multicollinearity it uses the eigensystem of the parameter covariance matrix. In computational experiments we consider the following illustrative methods: Lasso, ElasticNet, LARS, Ridge and Stepwise and determine the best one, which solve the multicollinearity problem for every considered configuration of dataset.

Keywords: regression analysis, feature selection methods, multicollinearity, test data sets, the criterion of the selected features redundancy.

1 Introduction

This study is devoted to multicollinearity problem and develops a testing procedure for feature selection methods. Assume that data sets have multicollinear features. *Multicollinearity* is a strong correlation between the features, which affect the target vector simultaneously. The multicollinearity reduces the stability of the parameter estimations. The multicollinearity problem, detection methods and methods to solve this problem are discussed in [1, 2, 3]. The

*This publication is based on work funded by Skolkovo Institute of Science and Technology (Skoltech) in the within the framework of the SkolTech/MITInitiative.

parameters are changing continuously.

2 Feature selection problem statement

Let $\mathfrak{D} = \{(\mathbf{X}, \mathbf{y})\}$ be the given data set, where the design matrix

$$\mathbf{X} = [\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_j, \dots, \boldsymbol{\chi}_n], \ \mathbf{X} \in \mathbb{R}^{m \times n} \text{ and } j \in \mathcal{J} = \{1, \dots, n\}.$$

The vector $\boldsymbol{\chi}_j$ is called the j-th feature and the vector $\mathbf{y} = [y_1, \dots, y_m]^\mathsf{T} \in \mathbb{Y} \subset \mathbb{R}^m$ is called the target vector. Assume that the target vector $\mathbf{y}$ and design matrix $\mathbf{X}$ are related through the following equation:

$$\mathbf{y} = \mathbf{f}(\mathbf{w}, \mathbf{X}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{f}$ maps the cartesian product of the feasible parameter space and the space of the $m \times n$ matrices to the target vector domain, and $\boldsymbol{\varepsilon}$ is the residual vector. The data fit problem is to estimate the parameter vector $\mathbf{w}^*$,

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^n} S(\mathbf{w} | \mathfrak{D}_\mathcal{L}, \mathcal{A}, \mathbf{f}), \tag{2}$$

where S is the error function. The set $\mathfrak{D}_\mathcal{L} \subset \mathfrak{D}$ is a training set and the set $\mathcal{A} \subseteq \mathcal{J}$ is the *active index set* used in computing the error function S. In the stresstest procedure we use the quadratic error function

$$S = \|\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\|_2^2 \tag{3}$$

and the linear regression function $\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w}$. The introduced stresstest procedure could be applied to the generalised linear model selection algorithms, where the model is $\mathbf{f} = \boldsymbol{\mu}^{-1}(\mathbf{X}\mathbf{w})$ and $\boldsymbol{\mu}$ is a link function.

Definition 2.1 Let $\mathcal{A}^*$ denote *the optimum index set*, the solution of the problem

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S_\mathfrak{m}(\mathcal{A} | \mathbf{w}^*, \mathfrak{D}_\mathcal{C}, \mathbf{f}), \tag{4}$$

where $\mathfrak{D}_\mathcal{C} \subset \mathfrak{D}$ is the test set, $\mathbf{w}^*$ is the solution of the problem (2) and $S_\mathfrak{m}$ is an error function corresponding to a feature selection method $\mathfrak{m}$ (5).

The feature selection problem (4) is to find the optimum index set $\mathcal{A}^*$. It must exclude indices of noisy and multicollinear features. It is expected that if one uses features indexed by the set $\mathcal{A}^*$ then it brings more stable solution of the problem (2), in comparison to the case of $\mathcal{A} \equiv \mathcal{J}$.

In the computational experiment we consider the feature selection methods from the set $\mathfrak{M} = \{$Lasso, LARS, Stepwise, ElasticNet, Ridge$\}$.

Definition 2.2 A feature selection method $\mathfrak{m} \in \mathfrak{M}$ is a map from the complete index set $\mathcal{J}$ to active index set $\mathcal{A} \subseteq \mathcal{J}$:

$$\mathfrak{m} : \mathcal{J} \to \mathcal{A}. \tag{5}$$

According to this definition we consider the terms feature selection problem and the model selection problem to be synonyms.

Definition 2.3 Let a model be a pair $(\mathbf{f}, \mathcal{A})$, where $\mathcal{A} \subseteq \mathcal{J}$ is an index set. The model selection problem is to find the optimum pair $(\mathbf{f}^*, \mathcal{A}^*)$ which minimizes the error function S (3).

Definition 2.4 Call *the model complexity* C the cardinality of the active index set $\mathcal{A}$, number of the selected features:

$$C = |\mathcal{A}|.$$

Definition 2.5 Define *the model stability* R be logarithm of the condition number κ of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$R = \ln \kappa = \ln \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and the minimum non-zero eigenvalue of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. The features with indices from the corresponding active set $\mathcal{A}$ are used in computing the condition number κ.

3 Multicollinearity analysis in feature selection

In this section we give definitions of multicollinear features, correlated features and features correlated with the target vector. In the following subsections we list and study the multicollinearity criteria.

Assume that the features $\boldsymbol{\chi}_j$ and the target vector $\mathbf{y}$ are normalized:

$$\|\mathbf{y}\|_2 = 1 \text{ and } \|\boldsymbol{\chi}_j\|_2 = 1, \; j \in \mathcal{J}. \tag{6}$$

Consider active index subset $\mathcal{A} \subseteq \mathcal{J}$.

Definition 3.1 The features with indices from the set $\mathcal{A}$ are called *multicollinear* if there exist the index j, the coefficients a_k, the index $k \in \mathcal{A} \setminus j$ and sufficiently small positive number $\delta > 0$ such that

$$\left\| \boldsymbol{\chi}_j - \sum_{k \in \mathcal{A} \setminus j} a_k \boldsymbol{\chi}_k \right\|_2^2 < \delta. \tag{7}$$

The smaller δ the higher *degree of multicollinearity.*

Definition 3.2 Call the features indexed i, j be *correlated* if there exists sufficiently small positive number $\delta_{ij} > 0$ such that:

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}. \tag{8}$$

From this definition it follows that $\delta_{ij} = \delta_{ji}$. In the special case $a_k = 0 \; k \neq j$ and $a_k = 1 \; k = j$ the inequalities (8) and (7) are identically.

Definition 3.3 A feature $\boldsymbol{\chi}_j$ is called *correlated with the target vector* $\mathbf{y}$ if there exists sufficiently small positive number $\delta_{yj} > 0$ such that

$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta_{yj}.$$

Further used the following notations RSS (Residual Sum of Squares) and TSS (Total Sum of Squares):

$$\text{RSS} = S(\mathfrak{D}_{\mathcal{L}}, \mathbf{w}^*) = \|\boldsymbol{\varepsilon}\|_2^2 \quad \text{and} \quad \text{TSS} = \sum_{i=1}^{m}(y_i - \overline{y})^2, \text{ where } \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i. \tag{9}$$

3.1 Variance inflation factor

The variance inflation factor VIF_j is used as a multicollinearity indicator [17]. The VIF_j is defined for j-th feature and shows a linear dependence between j-th feature and the other features.

To compute VIF_j estimate the parameter vector $\mathbf{w}^*$ according to the problem (1) assuming $\mathbf{y} = \boldsymbol{\chi}_j$ and extracting j-th feature from the index set $\mathcal{J} = \mathcal{J} \setminus j$. The functions RSS and TSS are computed similar to (9). The VIF_j is computed with the following equation:

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

where $R_j^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$ is the coefficient of determination.

According to [17] any $\text{VIF}_j \gtrsim 5$ indicates that the associated elements of the vector $\mathbf{w}^*$ are poorly estimated because of multicollinearity. Denote by VIF the maximum value of VIF_j for all $j \in \mathcal{J}$:

$$\text{VIF} = \max_{j \in \mathcal{J}} \text{VIF}_j.$$

However, VIF_j can be infinitely large for some features. In this case it is impossible to determine which features must be removed from the active set. This is major disadvantage of the variance inflation factor.

Another multicollinearity indicator is the condition number κ of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. The condition number is defined as:

$$\kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}},$$

where the λ_{max} and λ_{min} are the maximum and minimum non-zero eigenvalues of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$.

The condition number shows how much does the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ close to the singular matrix. The larger κ the more ill-conditioned matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$.

3.2 The Belsley criterion

To detect and remove indices of the multicollinear features from the active index set we state the direct optimization problem using the Belsley criterion. We propose the new criterion to compare feature selection methods: *the criterion of the selected features redundancy.* This criterion uses the maximum cardinality of the redundant index set, which can be removed within the error function does not raised above given value. The features are removed according to the Belsley criterion described below. The formal definition of the the maximum cardinality of the redundant index set is given by (16).

Assume that the parameter vector $\mathbf{w} \in \mathbb{R}^n$ has the multivariate normal distribution with the expectation $\mathbf{w}_{\mathrm{ML}}$ and the covariance matrix $\mathbf{A}^{-1}$,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\mathrm{ML}}, \mathbf{A}^{-1}).$$

The estimation $\hat{\mathbf{A}}^{-1}$ of the covariance matrix $\mathbf{A}^{-1}$ in the linear model is

$$\hat{\mathbf{A}}^{-1} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}.$$

To inverse $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ we use the singular value decomposition of the $m \times n$ matrix $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices, and $\mathbf{\Lambda}$ is the diagonal matrix with the singular values $\sqrt{\lambda_i}$ on the diagonal, such that

$$\sqrt{\lambda_1} \geq \ldots \geq \sqrt{\lambda_i} \geq \ldots \geq \sqrt{\lambda_r} > 0,$$

where $i = 1, \ldots, r$ and $r = \min(m, n)$. Thus, the inversion $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is following:

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1}.$$

The columns of the matrix $\mathbf{V}$ is the eigenvectors and the squares of the singular values λ_i are the eigenvalues of the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ since $\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\mathsf{T}}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2$.

Definition 3.4 The ratio of the maximum eigenvalue $\lambda_{\max}$ to the i-th eigenvalue λ_i is called *the condition index* η_i

$$\eta_i = \frac{\lambda_{\max}}{\lambda_i}.$$

The large value of η_i indicates the close-to-linear relation between the features. The larger value of η_i the closer relation between features to linear.

The variance of the vector $\mathbf{w}^*$ elements are estimated as diagonal entries of the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\mathsf{T}}$:

$$\mathrm{Var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2}.$$

Definition 3.5 *The coefficient variance proportion* q_{ij} is the j-th feature contribution to the variance of the i-th element of the optimal parameter vector $\mathbf{w}^*$. The formal definition of the coefficient variance proportion q_{ij} is

$$q_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum\limits_{j=1}^{n} v_{ij}^2/\lambda_j^2},$$

where $[v_{ij}] = \mathbf{V}$ and λ_j is the eigenvalue of the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$.

Motrenko A., Strijov V., Weber G.-W. Bayesian sample size estimation for logistic regression // Journal of Computational and Applied Mathematics, 2014, 255 : 743-752.

Sample Size Bayesian Estimation for Logistic Regression[☆]

Anastasiya Motrenko[a], Vadim Strijov[b], Gerhard-Wilhelm Weber[c]

[a] *Moscow Institute of Physics and Technology, Moscow, Russia*
[b] *Computing Center of the Russian Academy of Sciences, Moscow, Russia*
[c] *Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey*

Abstract

The problem of sample size estimation is important in the medical applications, especially in the cases of expensive measurements of immune biomarkers. The papers describes the problem of logistic regression analysis including model feature selection and includes the sample size determination algorithms, namely methods of univariate statistics, logistics regression, cross-validation and Bayesian inference. The authors, treating the regression model parameters as a multivariate variable, propose to estimate the sample size using the distance between parameter distribution functions on cross-validated data sets.

Keywords: logistic regression, sample size, feature selection, Bayesian inference, Kullback-Leibler divergence

1. Introduction

The paper is devoted to the logistic regression analysis [1], applied to classification problems in biomedicine. A group of patients is investigated as a sample set; each patient is described with a set of features, named as biomarkers and is classified into two classes. Since the patient measurement is expensive the problem is to reduce number of measured features in order to increase sample size.

The responsive variable is assumed to follow a Bernoulli distribution. Also, parameters of the regression function are evaluated [2, 3].

With given set of features, the model is excessively complex. The problem is to select a set of features of a smaller size, that will classify patients effectively. In logistic regression, features are usually selected by stepwise regression [4, 5]. In the computational experiment, exhaustive search is implemented. This makes the experts sure that all possible combinations of the features were considered. The authors use the area under ROC curve [6] as the optimum criterion in the feature selection procedure.

The problem of classification is associated with minimum sample size determination. In the paper, the following methods are discussed:

[☆] This project was supported by the Russian Foundation for Basic Research, grant 12-07-31095.

Email address: strijov@ccas.ru (Vadim Strijov)

1. Method of confidence intervals: a method of univariate statistics.
2. Method of sample size evaluation in logistic regression [7, 8]: unlike the previous one, this method considers the distribution of the responsive variable according to the logistic regression model.
3. Cross-validation: a method which evaluates sample size by observing potential overfitting [9, 10].
4. Comparing different subsets of the same sample by computing Kullback-Leibler [11] divergence between probability density functions of model parameters, evaluated at these subsets.

The data, used while conducting computational experiment can be found here [12].

2. Classification problem

Consider the sample set $D = \{(\mathbf{x}_i, y_i) : i = 1, \dots, m\}$, of m objects (patients). Each patient is described by n features (biomarkers), $\mathbf{x}_i \in \mathbb{R}^n$ and belongs to one of two classes: $y_i \in \{0, 1\}$. The logistic regression problem assumes that the vector of responsive variables $\mathbf{y} = [y_1, \dots, y_m]^T$ is a vector of Bernoulli random variables, $y_i \sim \mathcal{B}(\theta_i)$ with the probability density function

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} \theta_i^{y_i} (1 - \theta_i)^{1-y_i}. \tag{1}$$

We use the maximim likelihood method, write the error function for (1) as

$$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln \theta_i + (1 - y_i) \ln (1 - \theta_i). \tag{2}$$

find vector of parameters $\hat{\mathbf{w}}$ of regression function, one has to solve the following optimization problem:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w}). \tag{3}$$

Let us define the probability of a case as

$$f(\mathbf{x}_i^T \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \mathbf{w})} = \theta_i. \tag{4}$$

To solve the problem (3), using

$$\frac{df(\xi)}{d\xi} = f(1 - f),$$

we compute gradient of the error function $E(\mathbf{w})$:

$$\nabla E(\mathbf{w}) = -\sum_{i=1}^{m} \big(y_i(1 - \theta_i) - (1 - y_i)\theta_i\big)\mathbf{x}_i = \sum_{i=1}^{m} (\theta_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\theta} - \mathbf{y}),$$

in which $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]^T$ and the matrix $\mathbf{X} = \left[\mathbf{x}_1^T, \dots, \mathbf{x}_m^T\right]^T$ represents features sets.

Parameters are evaluated by Newton-Raphson method. Denote by $\mathbf{\Sigma}$ a diagonal matrix with diagonal elements $\Sigma_{ii} = \theta_i(1-\theta_i)$ $(i = 1, \ldots, m)$. Set the initial value $\mathbf{w} = [w_1, \ldots, w_n]^T$ of $\hat{\mathbf{w}}$

$$w_j = \sum_{i=1}^{m} y_i(1-y_i) \quad (j = 1, \ldots, n),$$

Then the $(k+1)$-th iteration of evaluation of $\hat{\mathbf{w}}$ is

$$\begin{gathered} \mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\theta} - \mathbf{y}) = \\ (\mathbf{X}^T\mathbf{\Sigma}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}(\mathbf{X}\mathbf{w}_k - \mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{y})). \end{gathered} \tag{5}$$

The process is repeated until the Euclidean distance $\| \mathbf{w}_{k+1} - \mathbf{w}_k \|$ is sufficiently small.

Thus, the classification algorithm is defined as:

$$a(\mathbf{x}, c_0) = \operatorname{sign}\big(f(\mathbf{x}, \mathbf{w}) - c_0\big), \tag{6}$$

where c_0 is a cut-off value of regression function (4), defined by (7).

Quality of classification. Let us use an additional to (1) quality functional AUC, or the area under the ROC-curve. Introduce TPR(ξ), which stands for true positive rate

$$\mathrm{TPR}(\xi) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i, \xi) = 1][y_i = 1]$$

and FPR(ξ)means the false positive rate

$$\mathrm{FPR}(\xi) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i, \xi) = 1][y_i = 0].$$

Here, the following denotation is used:

$$[y = 1] = \begin{cases} 1, & y = 1; \\ 0, & y \neq 1. \end{cases}$$

Thus, the bigger AUC value is, the better is the classifier.

Defining c_0 *value.* Every point $[\mathrm{FPR}(c_0), \mathrm{TPR}(c_0)]$ of the ROC-curve corresponds to some $c_0 \in [0, 1]$ value. As shown in figure 1, the most distant from segment [(0,0);(1,1)] point of the ROC-curve corresponds to the c_0 value used in (6):

$$\hat{c}_0 = \arg\max_{\xi\in[0,1]} \| \big(\mathrm{TPR}(\xi), \mathrm{FPR}(\xi)\big) - (\xi, \xi) \| = \arg\max_{\xi\in[0,1]} \sqrt{(TPR(\xi) - \xi)^2 - (FPR(\xi) - \xi)^2}. \tag{7}$$

Defining $\hat{c}_0$ includes computing AUC value and, therefore, computation of (6) and iterative estimation of parameters $\mathbf{w}$ according to (5).

Generation of simple structured IR functions by genetic algorithm without stagnation

Kulunchakov A. S.[a], Strijov V. V.[b]

[a]*Moscow Institute of Physics and Technology*
[b]*Computing Centre of the Russian Academy of Sciences*

Abstract

This paper investigates an approach to construct new ranking models for Information Retrieval. The IR ranking model depends on the document description. It includes the term frequency and document frequency. The model ranks documents upon a user request. The quality of the model is defined by the difference between the documents, which experts assess as relative to the request, and the ranked ones. To boost the model quality a modified genetic algorithm was developed. It generates models as superpositions of primitive functions and selects the best according to the quality criterion. The main impact of the research if the new technique to avoid stagnation and to control structural complexity of the consequently generated models. To solve problems of stagnation and complexity, a new criterion of model selection was introduced. It uses structural metric and penalty functions, which are defined in space of generated superpositions. To show that the newly discovered models outperform the other state-of-the-art IR scoring models the authors perform a computational experiment on TREC datasets. It shows that the resulted algorithm is significantly faster than the exhaustive one. It constructs better ranking models according to the MAP criterion. The obtained models are much simpler than the models, which were constructed with alternative approaches. The proposed technique is significant for developing the information retrieval systems based on expert assessments of the query-document relevance.

Keywords: information retrieval, evolutionary stagnation, ranking function, genetic programming, overfitting

Email addresses: `kulu-andrej@yandex.com` (Kulunchakov A. S.), `strijov@gmail.com` (Strijov V. V.)

to have as high quality as the stored superpositions. This superposition highly probably will be eliminated. Therefore the population will pass to the next iteration without changes. The genetic algorithm stops actual generation.

To outperform the ranking functions found in [2], one needs to extend the set of superpositions considered there. To perform it, a modified genetic algorithm is proposed. First, it detects evolutionary stagnation and replaces the worst stored superpositions with random ones. This detection is implemented with a structural metric on superpositions. Regularizers solve the problem of overfitting. They penalize the excessive structural complexity of superpositions. The paper analyzes various pairs regularizer-metric and chooses the pair providing a selection of better ranking superpositions. All strengths and weakness of compared approaches are summarized in Table 1.

The paper [2] uses TREC collections to test ranking functions. To make the comparison of approaches consistent, the present paper also use these collections. The collection TREC-7 (trec.nist.gov) is used as the train dataset to evaluate quality of generated superpositions. The collections TREC-5, TREC-6, TREC-8 are used as test datasets to test selected superpositions.

2. Problem statement

There given a collection C consisting of documents $\{d_i\}_{i=1}^{|C|}$ and queries $Q = \{q_j\}_{j=1}^{|Q|}$. For each query $q \in Q$ some documents C_q from C are ranked by experts. These ranks g are binary

$$g : Q \times C_q \to \mathbb{Y} = \{0, 1\},$$

where 1 corresponds to relevant documents and 0 to irrelevant.

To approximate g, superpositions of grammar elements are generated. The grammar $\mathfrak{G}$ is a set $\{g_1, \dots, g_m, x_w^d, y_w\}$, where each g_i stands for an mathematical function and x_w^d, y_w stand for variables. These variables are tf-idf features of *document-query* pair (d, q). Feature x_w^d is a frequency of the word $w \in q$ in d, feature y_w is a frequency of w in C:

$$x_w^d = t_d^w \log\left(1 + \frac{l_a}{l_d}\right), \quad y_w = \frac{N_w}{|C|}, \tag{1}$$

where N_w is the number of documents from C containing w, t_d^w is the frequency of w in d, l_d is the number of words in d (the size of a document d), l_a is an average size of documents in C. Each superposition f of grammar elements is stored as a directed labeled tree T_f with vertices labeled by elements from $\mathfrak{G}$. The set of these superpositions is defined as $\mathfrak{F}$.

The value of f on a pair (d, q) is defined as a sum of its values on (d, w), where w is a word from q:

$$f(d, q) = \sum_{w \in q} f(x_w^d, y_w).$$

The superposition f ranks the documents for each q. The quality of f is the mean average precision [1]

$$\mathrm{MAP}(f, C, Q) = \frac{1}{|Q|} \sum_{q=1}^{Q} \mathrm{AveP}(f, q),$$

where

$$\mathrm{AveP}(f, q) = \frac{\sum_{k=1}^{|C_q|} \big(\mathrm{Prec}(k) \times g(k)\big)}{\sum_{k=1}^{|C_q|} \mathrm{Rel}(k)}, \quad \mathrm{Prec}(k) = \frac{\sum_{s=1}^{k} g(s)}{k},$$

where $g(k) \in \{0, 1\}$ is a relevance of the k-th document from C.

This paper aims at finding the superposition f, which maximizes the following quality function

$$f^* = \underset{f \in \mathfrak{F}}{\operatorname{argmax}} \mathcal{S}(f, C, Q), \quad \mathcal{S}(f, C, Q) = \mathrm{MAP}(f, C, Q) - \mathrm{R}(f), \tag{2}$$

where R is a regularizer controlling the structural complexity of f.

The exhaustive algorithm in [2] generates random ranking superpositions consisting at most of 8 elements of the grammar $\mathfrak{G}$. Let $\mathfrak{F}_0$ be the set of the best superpositions selected in [2]. The solution f^* is compared with the superpositions from $\mathfrak{F}_0$ with respect to to MAP.

3. Generation of superpositions

IR ranking functions are superpositions of expert-given primitive functions. These superpositions are generated by the genetic algorithm. It uses an expertly given grammar $\mathfrak{G}$ and constructs superpositions of its elements. On each iteration it keeps stores a population of the best selected superpositions. To update them and pass to the next iteration, it generates new superpositions with use of the stored ones. Since the superpositions are represented as trees, the algorithm applies crossover $c(f, h)$ and mutation $m(f)$ operations to the stored trees

$$c(f, h) : \mathfrak{F} \times \mathfrak{F} \to \mathfrak{F}, \quad m(f) : \mathfrak{F} \to \mathfrak{F},$$

Definition 1. *Crossover operation $c(f, h) : \mathfrak{F} \times \mathfrak{F} \to \mathfrak{F}$ produces a new superpositions from given f and h. This operation represents f and h as trees, uniformly selected a subtree for each of them and swaps these subtrees.*

Feature generation for classification and forecasting problems

N. P. Ivkin
Moscow Institute of Physics and Technology
ivkinnikita@gmail.com

Abstract

We propose a problem statement for analysis of complex objects such as video sequences with contents, e-mail letters with attached files, source codes of programs. The proposed problem statement helps to organize work on a project, to simplify code development and to reduce labor costs.

Feature generation problem statement

Let $\mathfrak{S}$ be a set of measurements such that

$$\mathfrak{S} = \{\mathfrak{s}_1, ..., \mathfrak{s}_m\}.$$

The element $\mathfrak{s}_i$ of the set $\mathfrak{S}$ can be a time series a video sequence or a scoring application. Let $\mathbf{y} = \{y_1, ..., y_m\}$ be a set of class labels, or target variables.

Together with the set $\mathfrak{S}$ a set $V = V(\mathfrak{S})$ is given. The set $V = V(\mathfrak{S})$ is called a vocabulary and contains knowledge about the set of measurements. The vocabulary can be obtained as the result of measurement structure analysis and used for model generation.

By $G = \{g_1, ..., g_n\}$ denote an expert-given set of primitive functions such that each function g_j maps an object $\mathfrak{s}_i$ to an element (i, j) of the design matrix $\mathbf{X}$:

$$g_j : \quad (\mathbf{b}_j, \mathfrak{s}_i, V) \mapsto x_{ij} \in \mathbb{R}^1,$$

where $\mathbf{b}_j$ is the set of parameters of the primitive function g_j. By f denote the regression model f together with the set of parameters $\mathbf{w}$. To find the optimal parameters $\hat{\mathbf{w}}$ we minimize a loss function $S(\mathbf{w}|f, \mathbf{X}, \mathbf{y})$ such that

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(\mathbf{w}|f, \mathbf{X}, \mathbf{y}).$$

Examples

In this section we investigate classification and forecasting problem statements as the examples of feature generation problem.

Linear regression. According to the regression problem statement the target variable y belongs to the set of real numbers, $y \in \mathbb{R}$. The model f maps each row of the matrix $\mathbf{X}$ to the set $\mathbb{R}$ such that

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w},$$

where $\mathbf{f} = [f(\mathbf{w}, \mathbf{x}_1), ..., f(\mathbf{w}, \mathbf{x}_m)]^T$. As an example of the loss function S, the sum-squared error can be considered:

$$S(\mathbf{w}|f, \mathbf{X}, \mathbf{y}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|_2^2.$$

Classification. According to the two-class classification problem the target variable y belongs to the set of class labels, $y \in \{0, 1\}$. Consider a logistic regression problem as an example of classification problem. The model f maps each row of the matrix $\mathbf{X}$ to the segment $[0, 1]$ such that

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \frac{\mathbf{1}}{\mathbf{1} + \exp(-\mathbf{X}\mathbf{w})},$$

where optimal parameters $\hat{\mathbf{w}}$ minimize a loss function

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} S(\mathbf{w}|f, \mathbf{X}, \mathbf{y}),$$

where

$$S(\mathbf{w}|f, \mathbf{X}, \mathbf{y}) = -\ln\left(\sum_{i=1}^{m} y_i \log f(\mathbf{x}_i, \mathbf{w}) + (1 - y_i) \log\bigl(1 - f(\mathbf{x}_i, \mathbf{w})\bigr)\right).$$

Некторые задачи машинного обучения

- Задача оценки параметров модели,
- задача выбора признаков или объектов выборки,
- задача выбора модели оптимальной сложности,
- задача построения и выбора структуры модели,
- задача проверки гипотезы порождения данных.

Предполагается, что функция ошибки $S(\mathrm{w}|D, f)$ задана исходя из

- гипотезы порождения данных,
- либо из практических соображений.

Задача нахождения наиболее правдоподобных параметров

Задана выборка $D = \{(\mathrm{x}_i, y_i)\}$, $i \in \mathcal{I}$, функция ошибки модели S и модель — параметрическое семейство функций $f(\mathrm{w}, \mathrm{x})$. Требуется найти такие параметры w модели, которые бы доставляли минимум функции ошибки

$$\mathrm{w}^* = \arg \min_{\mathrm{w} \in \mathbb{W}} S(\mathrm{w}|D, f). \tag{1}$$

Функция ошибки, определенная посредством логарифмической функции правдоподобия

$$S(\mathrm{w}) = -\ln\bigl(p(D|\mathrm{w}, f)\bigr),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, будут называться наиболее правдоподобными.

Примеры функции ошибки в регрессии и классификации

Регрессия

Гипотеза порождения данных: $\mathrm{y} \sim \mathcal{N}(\mathrm{f}, \mathrm{I})$.
Функция ошибки:

$$S(\mathrm{w}) = \|\mathrm{y} - \mathrm{f}\|_2^2.$$

Классификация

Гипотеза порождения данных: $\mathrm{y} \sim \mathcal{B}(f, 1 - f)$.
Функция ошибки:

$$S(\mathrm{w}) = \sum_{i \in \mathcal{I}} y_i \ln f(\mathrm{w}^\mathsf{T}\mathrm{x})_i + (1 - y_i) \ln\bigl(1 - f(\mathrm{w}^\mathsf{T}\mathrm{x})_i\bigr).$$

Задача выбора оптимального набора признаков

- Задана выборка $D = \{(\mathrm{x}_i, y_i)\}$, $i \in \mathcal{I}$.
- Задано случайное разбиение множество индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$.
- Множество независимых переменных $\mathrm{x} = [x_1, \ldots, x_j, \ldots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \ldots, n\}$.
- Задано множество моделей-претендентов $\mathfrak{F} = \{f(\mathrm{w}, \mathrm{x})\}$.
- Модель — параметрическое семейство функций $f(\mathrm{w}, \mathrm{x}) = \mu(\mathrm{w}^\mathsf{T}\mathrm{x})$, где μ — функция связи (в случае регрессии $\mu = \mathrm{id}$, в случае классификации $\mu = \frac{1}{1+\exp(-\mathrm{w}^\mathsf{T}\mathrm{x})}$.
- Структура модели $f_{\mathcal{A}}$ задана множеством индексов $\mathcal{A} \subseteq \mathcal{J}$ и означает включение переменных $\mathrm{x}_{\mathcal{A}}$. Иначе, используются только признаки-столбцы матрицы X с индексами из множества $\mathcal{A}$.
- Задана функция ошибки S.

Задача выбора оптимального набора признаков

Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathrm{w}^*, D_{\mathcal{C}})$$

на разбиении выборки D, определенном множеством индексов $\mathcal{C}$.

При этом параметры w^* модели должны доставлять минимум функции

$$\mathrm{w}^* = \arg\min_{\mathrm{w} \in \mathbb{W}} S(\mathrm{w} | D_{\mathcal{L}}, f_{\mathcal{A}})$$

на разбиении выборки, определенном множеством $\mathcal{L}$.

Свойства оценок параметров и условия Гаусса-Маркова

Для линейной модели $\mathbf{f} = X\mathbf{w}$ при гипотезе порождения данных $y \sim \mathcal{N}(\mathbf{f}, I)$ функция ошибки имеет вид (с точностью до сомножителя)

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \|\varepsilon\|^2 = \sum_{i \in \mathcal{I}} (y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2.$$

МНК предполагает выполнение следующих условий:

1) независимые переменные $\mathbf{x}$ не являются случайными величинами,
2) математическое ожидание $\mathbb{E}(\varepsilon) = \mathbf{0}$,
3) дисперсия $\mathbb{D}(\varepsilon) = \sigma_\varepsilon^2 I$ (условие гомоскедаксичности),
4) при $i \neq k$ ковариация $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0$,
5) $\mathrm{rank}(X) = n \leqslant m$.

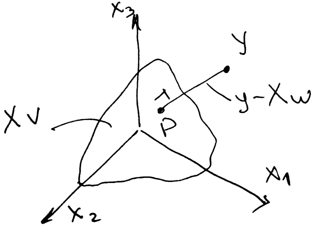
При этом оценки $\mathbf{w}$ состоятельны и несмещенны. При выполнении условия гомоскедаксичности $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I)$ они эффективны.

Проекция вектора y на пространство столбцов матрицы X

Рассмотрим вектор $\mathbf{v}$ ортогональный вектору регрессионных остатков $X\mathbf{w} - \mathbf{y}$:

$$(X\mathbf{v})^\mathsf{T}(X\mathbf{w} - \mathbf{y}) = \mathbf{v}^\mathsf{T}(X^\mathsf{T}X\mathbf{w} - X^\mathsf{T}\mathbf{y}) = 0.$$

Это равенство должно быть справедливо для произвольного $\mathbf{v}$; следовательно $X^\mathsf{T}X\mathbf{w} - X^\mathsf{T}\mathbf{y} = 0$.



При обратимости $X^\mathsf{T}X$ решение $\mathbf{w}_{\mathrm{ML}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}$ единственно.

Задачи нахождения оптимальных и наиболее правдоподобных параметров

Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$, функция ошибки модели S и модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x})$. Требуется найти такие параметры $\mathbf{w}$ модели, которые бы доставляли минимум функции ошибки

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} S(\mathbf{w}|D, f). \qquad (1)$$

Вариант 1. Функция ошибки, определенная посредством логарифмической функции правдоподобия

$$S(\mathbf{w}) = E_D = -\ln(p(D|\mathbf{w}, B, f)),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, будут называться наиболее правдоподобными, а задача (1) имеет вид

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w} \in \mathcal{W}} S(\mathbf{w}|D, B, f).$$

Задача нахождения наиболее вероятных параметров

Вариант 2. Функция ошибки

$$S(\mathbf{w}) = -\ln\big(p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)\big),$$

определенная посредством апостериорного распределения параметров модели

$$p(\mathbf{w}|D, A, B, f) = \frac{p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)}{\int\limits_{\mathbf{w}'\in\mathcal{W}} p(D|\mathbf{w}', B, f)p(\mathbf{w}'|A, f)d\mathbf{w}'},$$

обеспечивает максимизацию вероятности параметров. Параметры, найденные минимизацией такой функции ошибок, называются наиболее вероятными, а задача (1) имеет вид

$$\mathbf{w}_{\text{MP}} = \arg\min_{\mathbf{w}\in\mathcal{W}} S(\mathbf{w}|D, A, B, f).$$

Предполагается, что ковариационные матрицы A^{-1}, B^{-1} заданы.

Resume: model generation and selection cheat-sheet

A non-exhaustive list

What are hypothesis on data set?

- Set prior and posterior distribution hypothesis and construct internal criterion
- Assume one model or a mixture
- Assume outliers, class imbalances

How we generate models?

- Set a universal model
- Use primitive functions and rules of generation
- Forecast a model

How we select an optimal model?

- Use feature selection algorithms
- Use hyper-parameter analysis
- Run exhaustive search or genetic algorithm

How we check the model has the optimal structure?

- Use external criterions: AUC, BIC, Cp, Complexity, Stability