Big Data: Dodatek statystyczny

Jacek Koronacki i Jerzy Stefanowski

Częstochowa i Poznań, 14 i 22 kwietnia 2016

MC approaches: Model selection for linear regression - Random Subspace Method (RSM)

Mielniczuk and Teisseyre (2011) and (2013): Let $T_{i,m}$ be a t-statistic for i-th predictor in a linear regression model m with $|m|$ predictors. We have:

$$\frac{T_{i,m}^2}{n-|m|} = \frac{\mathrm{RSS}_{m-\{i\}} - \mathrm{RSS}_m}{\mathrm{RSS}_m}$$

It follows that the value of $T_{i,m}^2$ can serve as a measure of, simulatneously, the importance of the i-th predictor in model m and the quality of this very model.

MC approaches: Model selection for linear regression - Random Subspace Method (RSM)

In the RSM, a random subset m of features (predictors), of size $|m|$ smaller than the number of all features d and a number of observations n, is chosen. The model is fitted in the reduced feature space by OLS. Each of the selected features is assigned a weight describing its relevance in the considered submodel.

Random selection of features is repeated many times, corresponding submodels are fitted and the final weights (scores) of all d features are computed on the basis of all submodels.

The final model can then be constructed based on predetermined number of the most significant predictors or using a selection method applied to the nested list of models given by the ordering of predictors.

MC approaches: MCFS-ID Algorithm of Draminski et al.: the Monte Carlo Feature Selection (or MCFS) part

In what follows we begin with a brief description of an effective method for ranking features according to their importance for classification regardless of a classifier to be later used. Our procedure is conceptually very simple, albeit computer-intensive.

We consider a particular feature to be important, or informative, if it is likely to take part in the process of classifying samples into classes "more often than not".

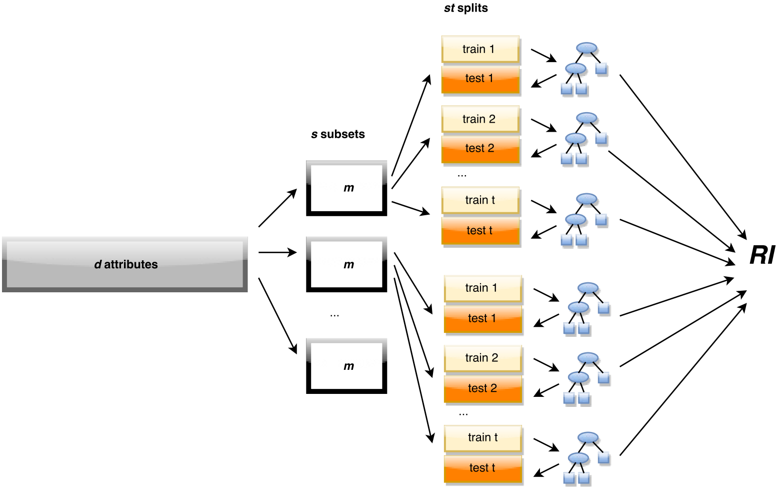
This "readiness" of a feature to take part in the classification process, termed relative importance of a feature, is measured via intensive use of classification trees. When assessing relative importance of a feature, the aforementioned "readiness" of the feature to appear in a given tree is suitably moderated by the (weighted) accuracy this tree.

MC approaches: MCFS-ID Algorithm: the MCFS part

In the main step of the procedure, we estimate relative importance of features by constructing thousands of trees for randomly selected subsets of features.

More precisely, out of all d features, s subsets of m features are selected, m being fixed and $m << d$, and for each subset of features, t trees are constructed and their performance is assessed. Each of the t trees in the inner loop is trained and evaluated on a different, randomly selected training and test sets which come from a split of the full set of training data into two subsets: each time, out of all n samples, about 66% of samples are drawn at random for training (in such a way as to preserve proportions of classes from the full set of training data) and the remaining samples are used for testing.

MC approaches: MCFS-ID Algorithm: the MCFS part



MC approaches: Interdependency Discovery, i.e., the ID part of the MCFS-ID Algorithm

This approach to interdependency discovery is significantly different from known approaches which consist in finding correlations between features or finding groups of features that behave similarly in some sense across samples (e.g., as in finding co-regulated features).

The focus is on identifying features that "cooperate" in determining that a sample belongs to a particular class. A directed graph of such "cooperating" features is constructed.

For an exposition of the MCFS-ID algorithm, see Draminski et al. (2008), (2010) and (2016).

Regularization approaches: Model selection for linear regression - ℓ_1 regularization

The Lasso (Least Absolute Selection Operator) for linear models:

As usual, we are given n observations, each with d explanatory variables (predictors), $(x_{i1}, x_{i2}, \ldots, x_{id})$, and one response variable, y_i,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{i,d} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are i.i.d. random errors with mean 0 and unknown variance σ^2, and $\beta_0, \ldots, \beta_d$ are unknown parameters.

Minimize

$$\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

subject to

$$\sum^{p} |\beta_j| \leqslant t.$$

Regularization approaches: ℓ_1 regularization

The Lasso, in contrast to ridge regression (i.e., ℓ_2 regularization), eliminates for small t some variables from the model. It can thus be used as a feature selection method, although one should be aware that the method is likely to include too many variables.

For exhaustive account of the Lasso and related approaches see Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015). For an important extension of the idea see Pokarowski and Mielniczuk (2015), where a three-stage algorithm for selecting a regression model is proposed, with LASSO used in the 1st stage for screening of predictors (features). See also Bogdan et al. (2015), where the regularizer is a sorted ℓ_1 norm.

Regularization approaches: Support Vector Machines - ℓ_2 regularization. And more

We skip an exposition of SVMs. Regarding their use for Big Data Analytics, we refer to Tan et al. (2014) and to Priyadarshini and Agarwal (2015).

There are more statistical approaches to dealing with high-dimensional data than those already hinted to and the Bayesian ones. See Bühlmann and van de Geer (2011) for an approach which stems from undirected graphical modeling and is based on inferring zero partial correlations for variable selection (the so-called PC-simple algorithm).

A still another and promising approach, which builds on ranking the marginal correlations and is referred to as sure independence screening, has been introduced by Fan and Lv (2008); see also Fan and Song (2010).

Model selection for linear regression - Bayesian approaches

Broman and Speed (2002): Let

$$y_i = \mu + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i,$$

where $x_{ij} = 1$ or $x_{ij} = 0$ and the ε_i are i.i.d. and normally distributed, $N(0, \sigma^2)$ (in fact, x_{ij} represents genotype at marker j for individual i). The task is to select a model for which Schwarz's Bayesian Information Criterion (BIC) assumes the minimal value;

$$BIC = n \cdot \log RSS(\beta) + \frac{1}{2} k \log n,$$

where k is the number of parameters β_j in the model. It was observed by Broman i Speed that the BIC tends to overestimate the number of parameters in the model. Accordingly, they proposed the 1st modification of the BIC.

Model selection for linear regression - Bayesian approaches

The Bayesian model selection advocates choosing the model M that maximizes posterior probability of the model given the data, this probability being proportional to

$$L(y|M)\pi(M),$$

where $\pi(M)$ is a prior probability for model M (Schwartz assumed noninformative uniform prior π), and

$$L(y|M) = \int L(y|M,\beta)f(\beta|M)d\beta,$$

$f(\beta|M)$ being some prior distribution on the vector of model parameters; for a wide class of these distributions one gets

$$\log L(y|M) = \log L(y|\beta) - \frac{1}{2}(k+2)\log n.$$

For the family of normal linear regression models, maximization of this last expression is equivalent to minimization of the BIC.

Model selection for linear regression - Bayesian approaches

Bogdan et al. (2004) introduced another modification of BIC (mBIC), assuming binomial prior distribution, $\mathrm{Bin}(d, c/d)$, with some fixed c, for the model size. See Bogdan et al. (2011) for later developments and Frommlet et al. (2012) for application of their approach to Genome-Wide Association Studies.

It is easy to extend the outlined approach to include regression models with interactions. It is also possible to extend it to include generalized linear models (possibly with constraints on the model's parameters).

The outlined approach is by far not the only one possible among this strand of Bayesian approaches; e.g., a similar approach is that based on the extended BIC, and a completely different approach, which bears some relationship with support vector machines, is that of relevance vector machines. (See, e.g., Chen and Chen (2008) and (2011), and Tipping (2001), Fletcher (2010) and Saarela et al. (2010).)

Nonparametric Bayesian approaches

Let Y be a response and $X = (X^{(1)}, \dots, X^{(p)} \in R^p$ be explanatory variables. Assume

$$Y = f(X) + \varepsilon,$$

with ε normally distributed, $N(0, \sigma^2)$.

Usually, a Gaussian Process (GP) prior for f is assumed to have zero mean and square exponential covariance function (kernel function) $\exp(-\|x - x'\|^2/c)$. Such processes are smooth in a well-known sense. Other kernels can be used, and another smoothness conditions on f can be imposed.

It should be emphasized that the above mentioned use of a kernel function casts the whole approach into the area of ML with kernels (kernel machines). Indeed, some far reaching similarities (and differences) with ridge regression, SVMs, as well as with spline models are obvious and deserve separate analysis.

Nonparametric Bayesian approaches, contd.

An excellent exposition of Gaussian processes for ML is given in Rasmussen and Williams (2006); another excellent, albeit short, introduction to GPs in ML can be found in Bishop (2006). In neither of these expositions problems pertaining to dealing with Big Data are addressed, although Rasmussen and Williams (2006) has a chapter titled Approximation Methods for Large Datasets.

Nonparametric Bayesian approaches, contd.

However interesting GPs for ML are, within the context of Big Data Analytics, special emphasis has to be placed on variable selection and/or variable projections. Loosely speaking, such mechanisms can be included into the nonparametric Bayesian approach by adding more randomness into the process, i.e., introducing suitable hyperparameters. See Tokdar (2011) for variable selection and linear projection proposals which have been shown to give consistent (in probability, and at a known rate) estimators of an unknown f; e.g., for f depending on $d < p$ variables, the rate of convergence is

$$n^{-\frac{\alpha}{2\alpha+d}}(\log n)^k$$

for any $k > p + 1$.

Yang (2014) has noticed that Tokdar's proposal can be considered effective only if $d << p$.

Nonparametric Bayesian approaches, contd.

Yang (2014) has provided a general framework to assess the minimax risks for regression problems under ℓ_2 loss (see there for an excellent account of earlier, sometimes pioneering, results in the area). He has introduced a general class of Bayesian sieve estimators which, under certain (more or less restrictive) conditions, achieve the optimal minimax risk when f depends on $d << \min\{n, p\}$ variables or is a sum of finitely many, k, functions, each of which depends on $d_s << \min\{n, p\}$ variables.

He has shown also that a GP regression approach can lead to the minimax optimal adaptive rate in estimating f under some conditions when the function's domain lies on a Riemannian manifold.

See also Yang and Dunson (2014) and Yang and Tokdar (2015).

Selective bibliography:

- Bishop C.: Pattern Recognition and Machine Learning. 2006; Springer.
- Bogdan M., Ghosh J.K., Doerge R.W.: Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting Quantitatve Trait Loci. Genetics. 2004; 167, 989-999.
- Bogdan, M., Chakrabarti, A., Frommlet, F., Ghosh, J.K.: Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. Annals of Statistics. 2011; 39(3), 1551-1579.
- Bogdan M., van den Berg E., Sabatti C., Su W., Candes E.J.: SLOPE—ADAPTIVE VARIABLE SELECTION VIA CONVEX OPTIMIZATION. The Annals of Applied Statistics. 2015; 9(3), 1103–1140.
- Broman K.W, Speed T.P, A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Royal Statist. Soc. 2002; B 64, 641-656.
- Bühlmann P., van de Geer S., *Statistics for High-Dimensional Data*, Springer, 2011.
- Chen J., Chen Z., Extended Bayesian information criterion for model selection with large model space. 2008; Biometrika, 94, 759-771.
- Chen J., Chen Z., Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. 2011; arXiv:1107.2502
- Draminski M, Rada Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J.: Monte Carlo feature selection for supervised classification. Bioinformatics. 2008; 24, 110-117.
- Draminski M., Kierczak M., Koronacki J., Komorowski J.: Monte Carlo Feature Selection and Interdependency Discovery in Supervised Classification. In: Koronacki J., et al. (eds): Advances in Machine Learning II. 2010; Springer series: Studies in Computational Intelligence, Vol. 263, 371-385.
- Draminski M., Dąbrowski M.J., Diamanti K., Koronacki J., Komorowski J.: Discovering networks of interdependent features in high-dimensional problems. In: : N.Japkowicz and J.Stefanowski (eds.), Big Data Analysis: New Algorithms for a New Society. 2016; Springer, 285-304.
- Fan J., Lv J., Sure independence screening for ultra-high dimensional feature space. J. Royal Statist. Soc. 2008; B 70 (5), 849-911.
- Fan. J., Song R., Sure independence screening for generalized linear models with np-dimensionality. Ann. Statist. 2010; 38(6), 3567–3604.
- Fletcher T., Relevance vector machines explained. 2010; Tech. report, www.cs.ucl.ac.uk/staff/T.Fletcher

Selective bibliography, contd.:

- Frommlet F., Ruhaltinger F., Twaróg P., Bogdan M.: Modified versions of the Bayesian Information Criterion for genome-wide association studies Computational Statist. and Data Anal. 2012; 56(5), 1038-1051.
- Mielniczuk J., Teisseyre P.: Using Random Subset Method for prediction and variable importance assessment in linear regression. Computational Statist. and Data Anal. 2012; in press.
- Mielniczuk J., Teisseyre P.: Selection and Prediction for Linear Models using Random Subspace Methods. Proceedings of the Conference Information Technologies: Research and their Interdisciplinary Applications, Institute of Computer Science. 2013; 103-121.
- Pokarowski P., Mielniczuk J.: Combined ℓ_1 and Greedy ℓ_0 Penalized Least Squares for Linear Model Selection. J. Machine Learning Reserach. 2015; 16, 961-992.
- Priyadarshini A., Agarwal S.: A Map Reduce based Support Vector Machine for Big Data Classification. International Journal of Database Theory and Application. 2015; 8(5), 77-98.
- Rasmussen C.E., Williams C.K.I.: Gaussian Processes for Machine Learning. 2006; MIT Press.
- Saarela M., Elomaa T., Ruohonen K., An Analysis of Relevance Vector Machine Regression. In: Advances in Machine Learning, vol. 2; Springer, 2010.
- Tan M., Tsnag I.W., Wang L.: Towards Ultrahigh Dimensional Feature Selection for Big Data. J. Machine Learning Reserach. 2014; 15, 1371-1429.
- Tipping M.E.: Sparse Bayesian learning and the relevance vector machine. 2001; Journal of Machine Learning Research 1, 211-244.
- Tokdar S.T: Dimension adaptability of Gaussian process models with variable selection and projection. 2011; Tech. Rep., Duke Iniversity, arXiv:1112.0716v1.
- Yang Y.: Nonparametric Bayes for Big Data. PhD Thesis. 2014; Duke Unoversity.
- Yang Y., Dunson D.B.: Bayesian Manifold Regression. 2014; Annals of Statistics, to appear.
- Yang Y, Tokdar S.T.: MINIMAX-OPTIMAL NONPARAMETRIC REGRESSION IN HIGH DIMENSIONS. 2015; Annals of Statistics 43(2), 652–674.