Feature selection by distributions contrasting

Tsurko V.V.[1], Michalski A.I.[1,2]

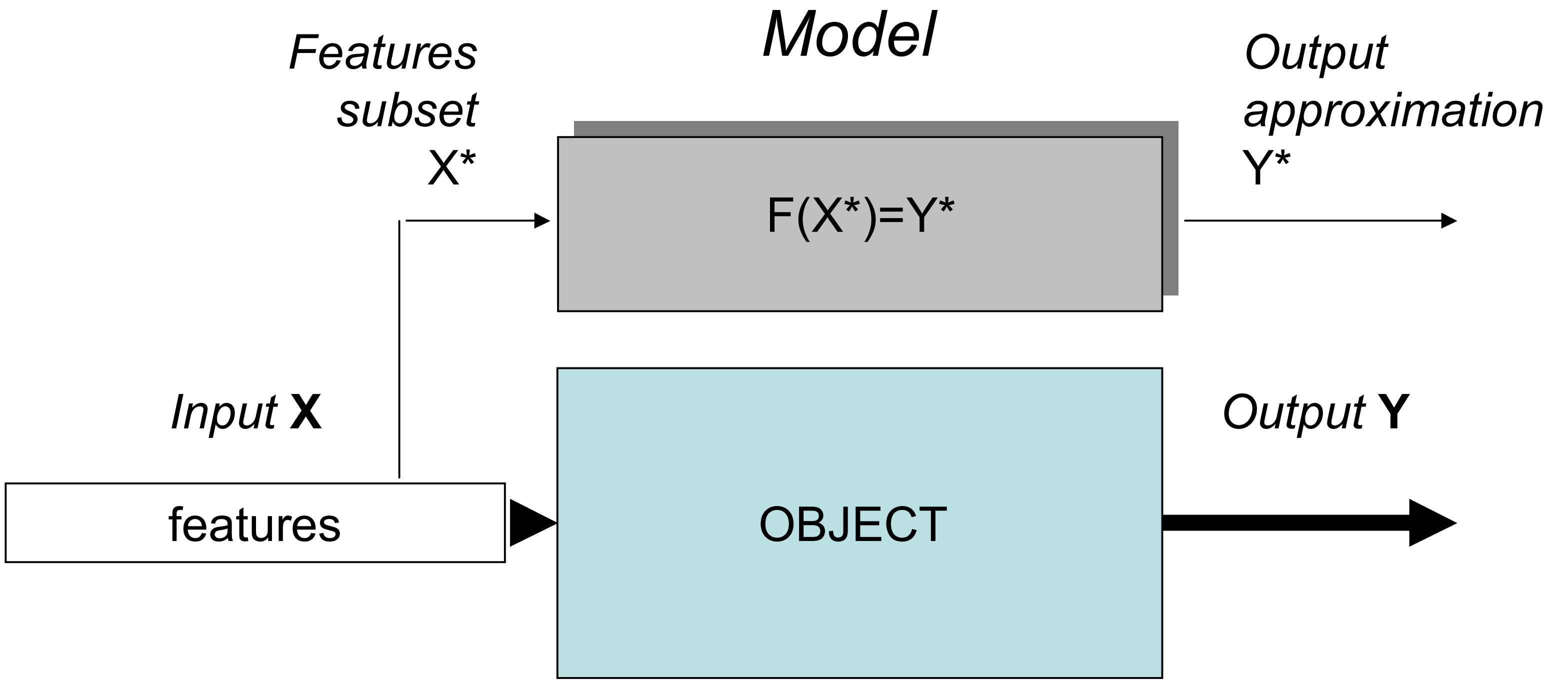
1 *Institute of Control Sciences of Russian Academy of Science, Moscow*
2 *Research University – Higher School of Economics, Moscow*

AIMSA 2014, Varna, Bulgaria

Outline

- What is feature selection
- Why do we need to select features
- How to select features
  - General setting. Loss function
  - Average risk. Empirical risk
  - Distributions contrasting
  - Practical realization
- Real life example

What is feature selection



Learning sample: pairs (X^i, Y^i) i=1,…N

What is feature selection

feature selection, also known as **variable selection**, **attribute selection** or **variable subset selection**, is the process of selecting a subset of relevant features for use in **model construction**

Example of feature selection

Let (X,Y) be a vector of two independent features.

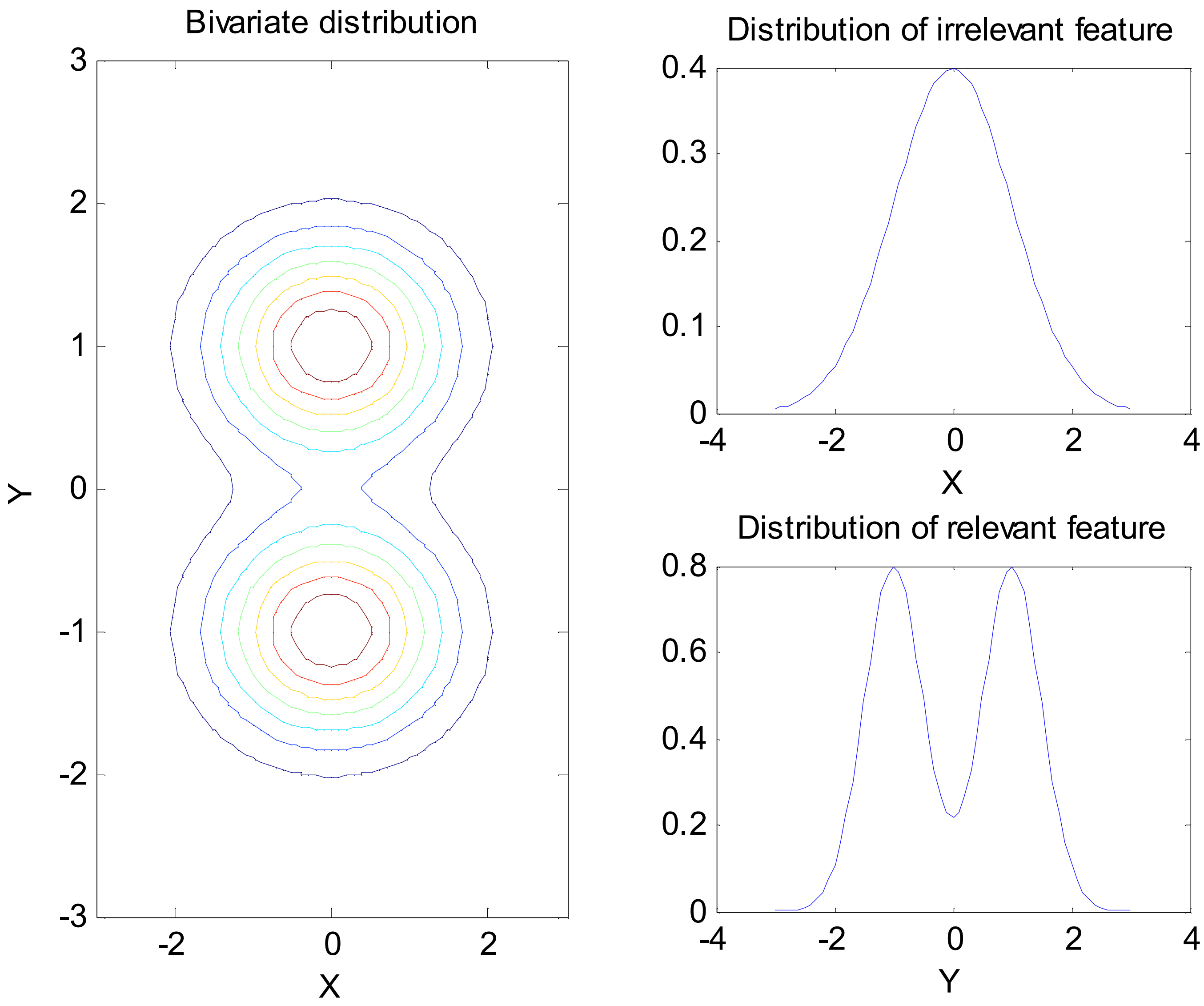
Distribution of feature X does not depend on hypothesis H_0 or H_1.

Distribution of feature Y do depend on hypothesis H_0 or H_1.

Feature X is irrelevant in

hypothesis H_0 vs H_1 testing.

Example of feature selection



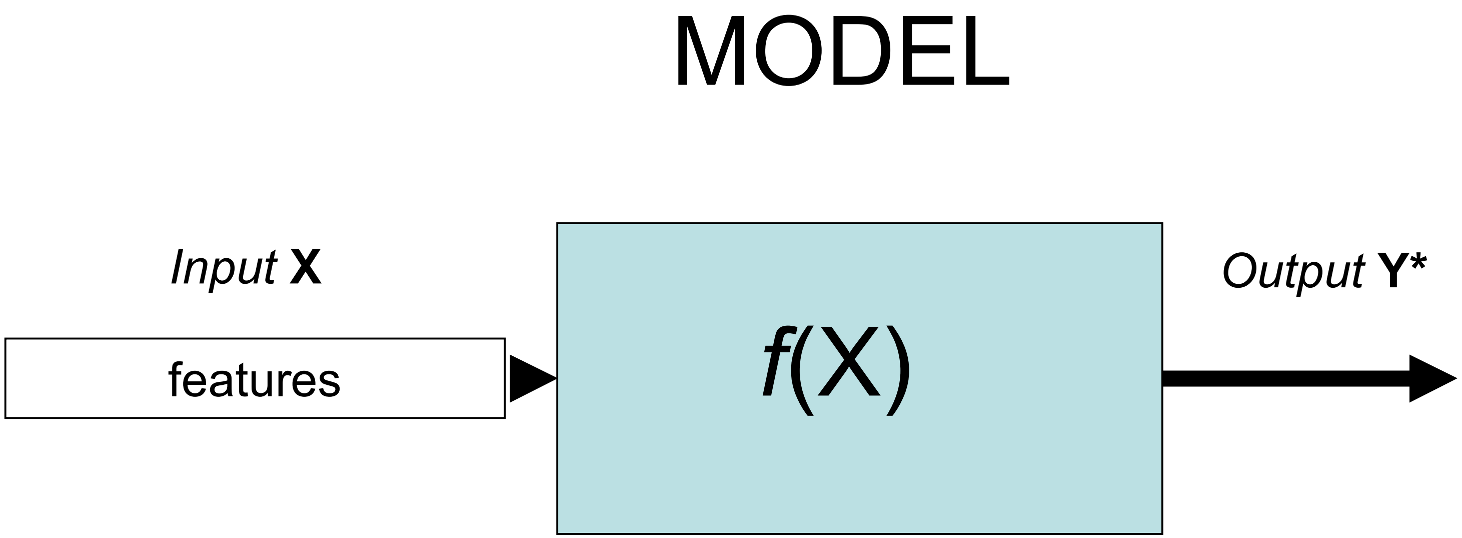
In the report we describe a method of DISTRIBUTIONS CONTRASTING

which means selection of feature subset to maximize differences between distributions under different hypothesis – inclass distributions

Why do we need to select features

- many machine learning algorithms don't operate well on the big amount of features
- as the number of features increases the algorithm run-time grows dramatically
- statistical accuracy of the algorithm decreases in the case of big number of features and the overfitting problem can arrize

General setting. Loss function.



$$L(Y, Y^*) = L_f(Y, X)$$ Loss function

General setting. Average risk.

$$M(f) = E_{X,Y}\left(L_f(Y, X)\right)$$

Examples

Classification $L_f(Y, X) = I(Y = f(X))$

Regression $L_f(Y, X) = (Y - f(X))^2$

Density estimation $L_f(X) = -\ln f(X)$

General setting. Empirical risk.

$$M_e(f) = \frac{1}{N}\sum_{i=1}^{N} L_f\left(Y^i, X^i\right)$$

Examples

Classification #error/#examples

Regression $\frac{1}{N}\sum_{i=1}^{N}\left(Y^i - f\left(X^i\right)\right)^2$

Density estimation $-\sum_{i=1}^{N}\ln f\left(X^i\right)$

Average and Empirical risks in Distributions contrasting

X – vector of features, Y={0,1} – class label (hypothesis label)

TASK: select the features subset for better classification (hypothesis testing)

Average risk in Distributions contrasting

$p(x \mid H_0)$ $p(x \mid H_1)$ conditional distributions

φ_0 φ_1 approximations for conditional distributions

Define average risk for approximations φ_0 and φ_1 as

$$M(\varphi_0, \varphi_1) = -E_{x,y}\left(y \ln \varphi_0(x) + (1-y)\ln \varphi_1(x)\right)$$

Average risk in Distributions contrasting

It is easy to see, that

$$M(\varphi_0,\varphi_1) = -E_{x,y}\left(y\ln\varphi_0(x) + (1-y)\ln\varphi_1(x)\right)$$
$$= I(\varphi_0,\varphi_1) - E_{x,y}\left(y\ln p(x\,|\,H_1) + (1-y)\ln p(x\,|\,H_0)\right)$$

where

$$I(\varphi_0,\varphi_1) = -E_{x,y}\left(y\ln\frac{\varphi_0(x)}{p(x\,|\,H_1)} + (1-y)\ln\frac{\varphi_1(x)}{p(x\,|\,H_0)}\right)$$

which shows how big is divergence between two pairs of distributions

$\varphi_0(x), p(x\,|\,H_1)$ and $\varphi_1(x), p(x\,|\,H_0)$

Average risk in Distributions contrasting

Small divergence $I(\varphi_0,\varphi_1)$ means that approximation $\varphi_0(x)$ is close to inclass distribution $p(x|H_1)$ and approximation $\varphi_1(x)$ is close to in class distribution $p(x|H_0)$. So, these approximations are not good for classification.

$$M(\varphi_0,\varphi_1)\xrightarrow[\varphi_0,\varphi_1\in\Psi]{}\max$$

equivalent

$$I(\varphi_0,\varphi_1)\xrightarrow[\varphi_0,\varphi_1\in\Psi]{}\max$$

Ψ – class of different features sets distributions

Average risk maximization in Distributions contrasting

Distribution contrasting task: find such a features set *F*, that approximations $\varphi_0(x)$ and $\varphi_1(x)$, produced using these features, deliver maximum for average risk

$$\max_{\varphi_0,\varphi_1 \in \Psi_F} M(\varphi_0,\varphi_1) \xrightarrow[F]{} \max$$

here

Ψ_F - class of distributions approximations $\varphi_0(x)$ and $\varphi_1(x)$, produced using features from set *F*

Empirical risk maximization in Distributions contrasting

We substitute this problem with empirical risk maximization

$$\max_{\varphi_0, \varphi_1 \in \Psi_F} M_e(\varphi_0, \varphi_1) \xrightarrow[F]{} \max$$

here

Ψ_F - class of distributions approximations $\varphi_0(x)$ and $\varphi_1(x)$, produced using features from set *F*

Average risk vs Empirical risk

If we know that with a given probability

$$\sup_{\varphi_0,\varphi_1 \in \Psi_F} \left| M(\varphi_0,\varphi_1) - M_e(\varphi_0,\varphi_1) \right| < \varepsilon(\Psi_F)$$

then with the same probability for any $\varphi_0(x)$ and $\varphi_1(x)$ in Ψ_F

$$M_e(\varphi_0,\varphi_1) - \varepsilon(\Psi_F) < M(\varphi_0,\varphi_1)$$

and we can maximize the **penalized empirical risk**

$$M_e(\varphi_0,\varphi_1) - \varepsilon(\Psi_F) \xrightarrow[F]{} \max$$

What distributions approximations $\varphi_0(x)$ and $\varphi_1(x)$ use in distributions contrasting problem

and

how to calculate the penalty term $\varepsilon\left(\Psi_F\right)$?

Distributions approximation in Distributions contrasting

For inclass distributions approximation we use Bayesian histograms

$$\varphi^{b}(i) = \frac{n_i + 1}{\sum_{j=1}^{k} n_j + k}$$

n_i – number of sample elements in *i*-th bin

k – number of bins in histogram

Distributions contrasting

Loss function

$$L_{\varphi_0^b,\varphi_1^b}(x,y) = -y\ln\varphi_0^b(x) - (1-y)\varphi_1^b(x)$$

Average risk

$$M(\varphi_0^b,\varphi_1^b) = -E_{x,y}\left(y\ln\varphi_0^b(x) + (1-y)\ln\varphi_1^b(x)\right)$$

Empirical risk for Bayesian histograms

$$M_e(\varphi_0^b,\varphi_1^b) = -\frac{1}{l_0+l_1}\sum_{i=1}^{k}\left(m_i\ln\varphi_0^b(i) + n_i\ln\varphi_1^b(i)\right)$$

Rademacher penalty term

General formula

$$R = \sup_{f} \left| \frac{1}{N} \sum_{i=1}^{N} \delta_i L_f \left(Y^i, X^i \right) \right|$$

Formula in distribution contrasting problem

$$R = \sup_{\varphi_0^b, \varphi_1^b \in \Psi_F} \left| \frac{1}{l_0 + l_1} \left(\sum_{i=1}^{l_1} \delta_i^1 \ln \varphi_0^b(i) + \sum_{j=1}^{l_0} \delta_j^0 \ln \varphi_1^b(i) \right) \right|$$

$\delta_i, \delta_i^0, \delta_j^1$ Independent random variables with equally probable values

-1 and +1

Main inequalities

For the class of functions uniformly bounded by a constant U for all t>0 it holds (*Koltchinskii, 1999*)

$$P\left\{\sup_{\varphi}\left|M(\varphi)-M_e(\varphi)\right|\geq 2R+\frac{3tU}{\sqrt{l}}\right\}\leq \exp\left(-\frac{t^2}{2}\right)$$

From this we write for distribution contrasting problem that with probability not less than 1-η the next inequality is true

$$M\left(\varphi_0^b,\varphi_1^b\right)>M_e\left(\varphi_0^b,\varphi_1^b\right)-2R-\frac{3\sqrt{-2\ln\eta}\,\ln\left(l_0+l_1+k\right)}{\sqrt{l_0+l_1}}$$

Feature selection by distribution contrasting algorithm

1. Order features from 1 till *d* (total number);
2. For *k* changing from 1 till *d* calculate
   - *k* -fold Bayesian histogram $\varphi^k_0(x)$ for one class sample and histogram $\varphi^k_1(x)$ for the other class sample;
   - calculate empirical risk value;
   - calculate value for Rademacher penalty term. It is done analytically;
   - by formula calculate lower bound for mean risk;
3. Take as optimal the set of features corresponding to *k* for which the lower bound for mean risk is maximal.

Classification states of real process

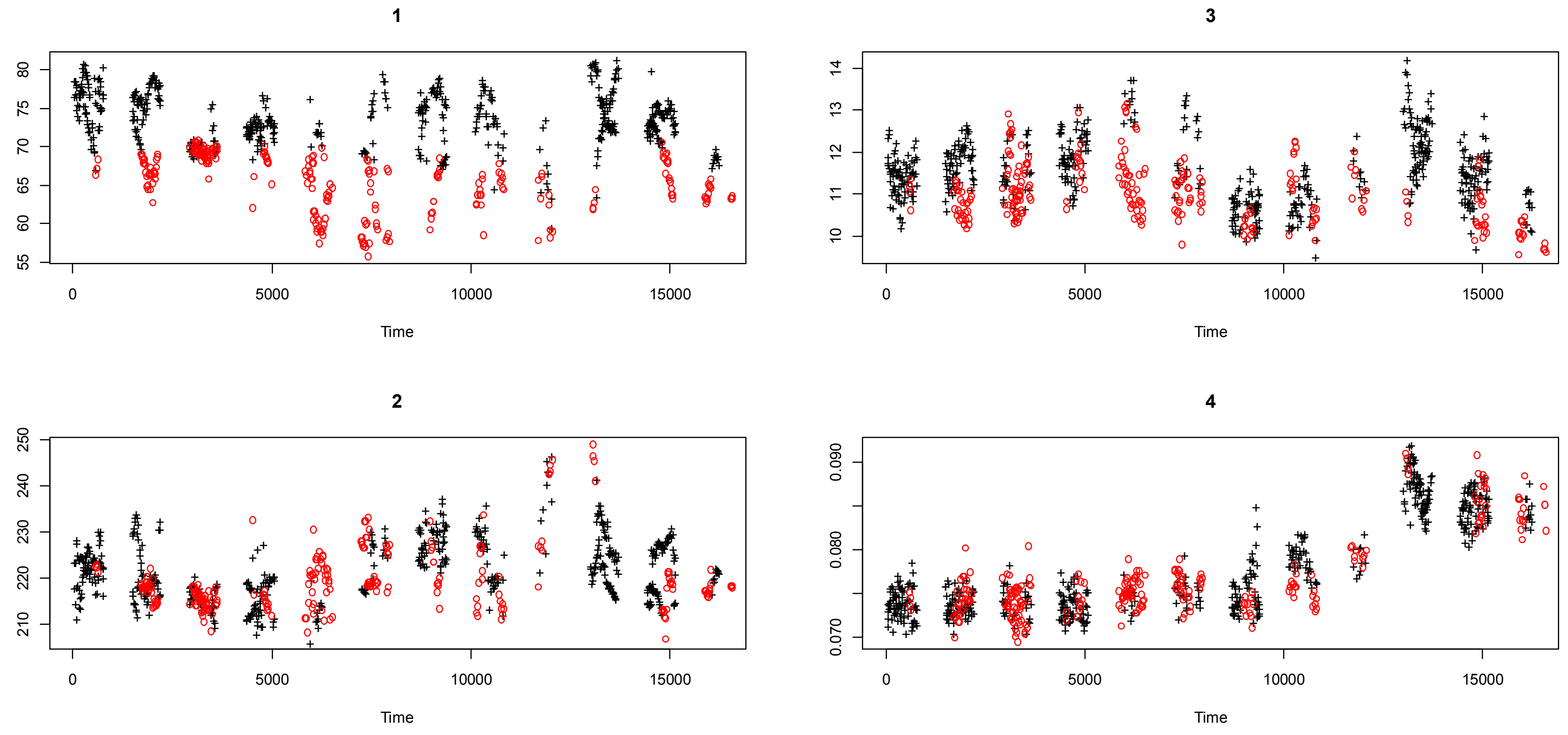
Data

- Time records of 10 parameters
- Two states labeled by experts. 562 points in the first class, 268 points in the second class

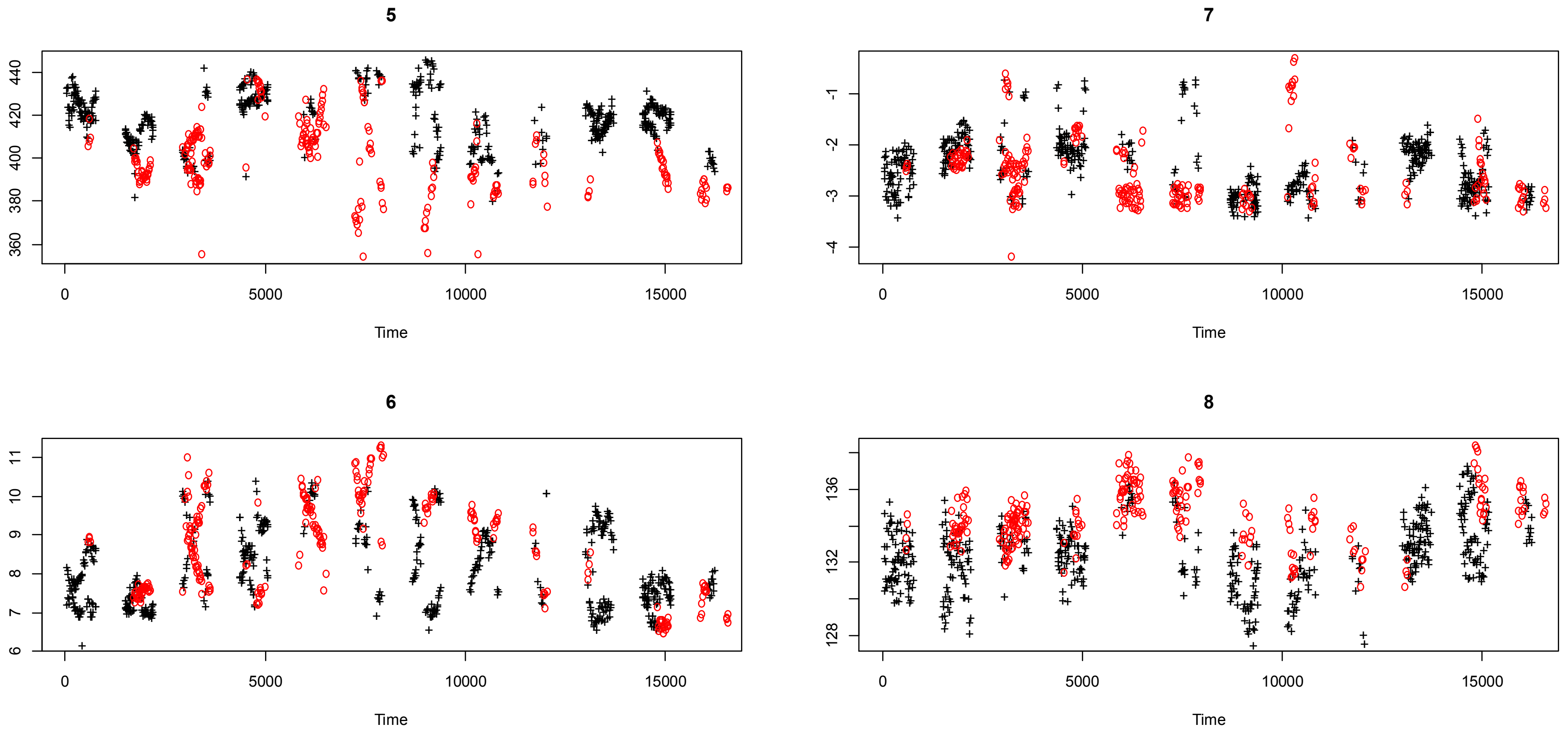
Task

Find a set of parameters for reliable classification of the process state

Data. Time records for parameters #1- #4 in two classes

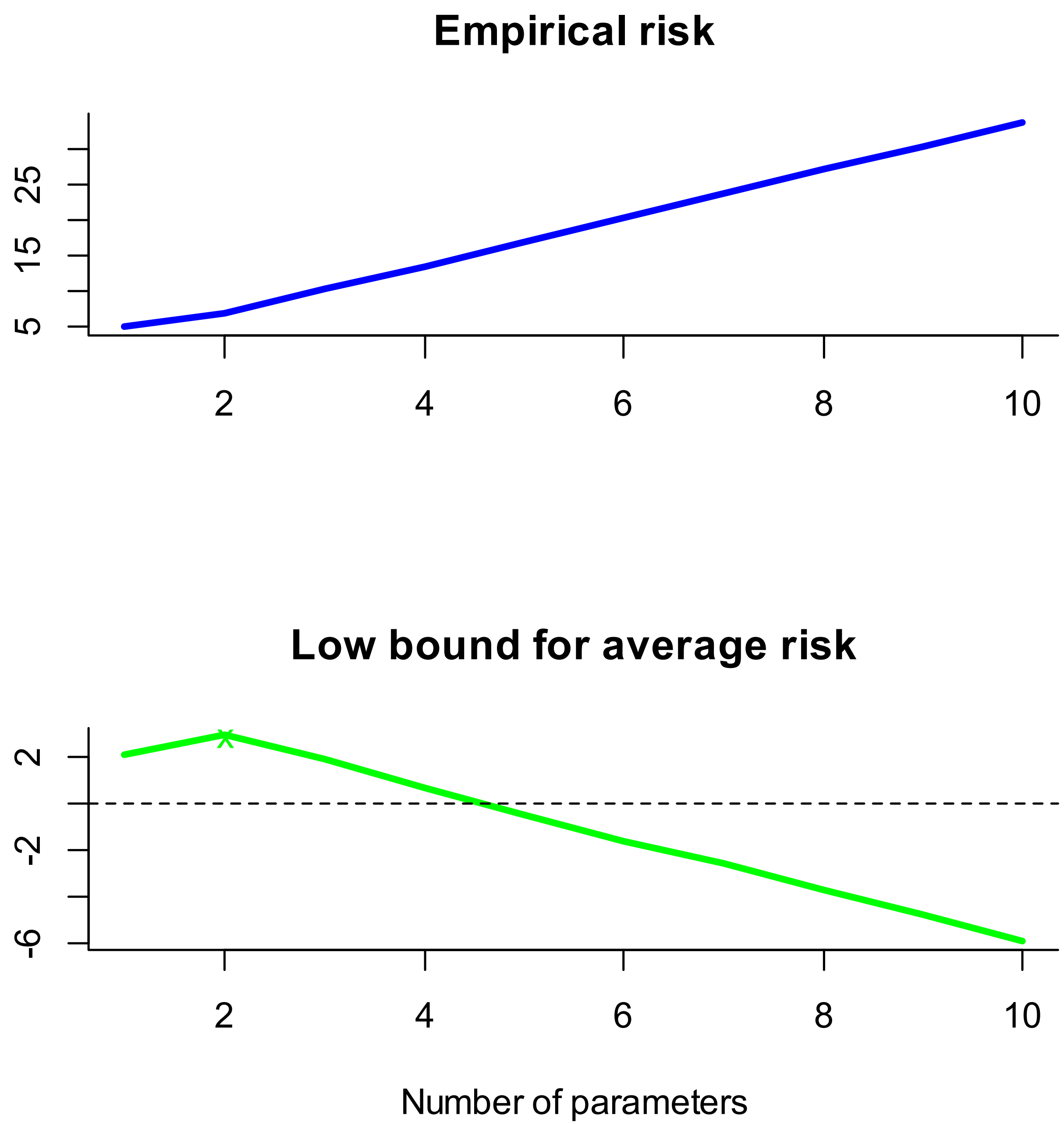


Data. Time records for parameters #5- #8 in two classes



Ordering

- Find a parameter with the maximal value of empirical risk. Fix it as #1.
- Iterate pairs, composed by #1 and one from the rest parameters. Find a pair with the maximal value of empirical risk. Fix new parameter as #2.
- Iterate triples, composed by #1, #2 and one from the rest parameters. Find a triple with the maximal value of empirical risk. Fix new parameter as #3.
- Continue till order all parameters.

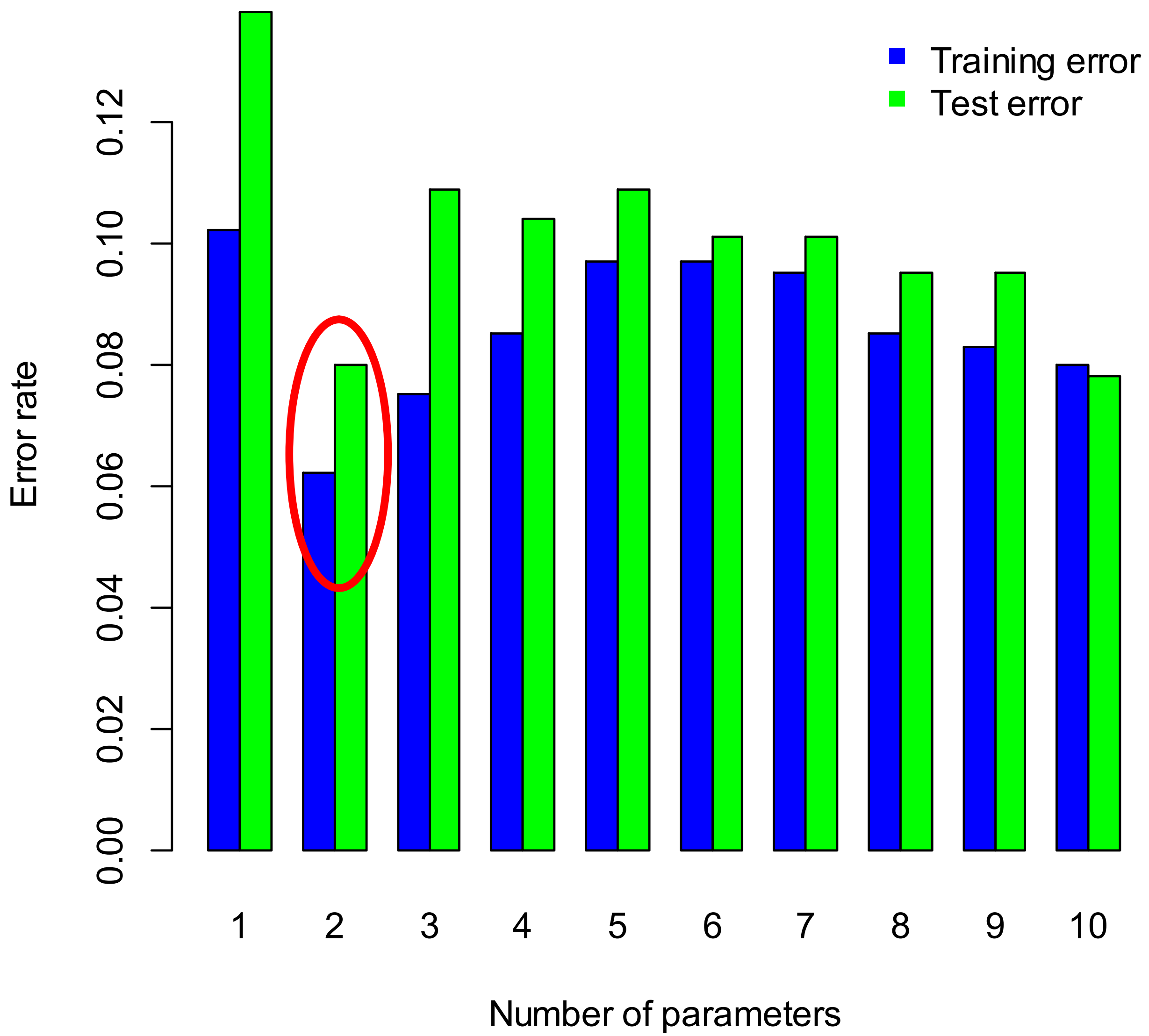


Empirical risk
5
15
25
2
4
6
8
10
Low bound for average risk
2
-2
-6
2
4
6
8
10
Number of parameters

Verification procedure

- Randomly divide data into training sample and into test sample.
- Select optimal set of parameters using only training sample data.
- Use the optimal set of parameters to classify test sample data.
- Calculate the error rate. Compare this error rate with results of test sample data classification using the other sets of parameters.

Result of verification using Naïve Bayes Classifier



Conclusion

- Distribution contrasting technique is suitable for feature selection.
- The method combines information theory approach, average risk estimation and uniform estimates of empirical risk deviation from average risk.
- This method allows to extract features mostly significant for two given hypothesis testing.
- The method has applications in analysis of links between processes of different nature. For example, between cancer mortality and non cancer morbidity

  (*V.V. Tsurko, A.I. Michalski, Advances in Gerontology, 2014, 10.1134/S2079057014030084*).

Thank you!

Any questions?