Approximation of combinatorial generalization bounds for threshold classifiers

Shaura Ishkina (shaura-ishkina@yandex.ru)

Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences

Intelligent Data Processing: Theory and Applications

October 10 – 14, 2016

Outline

Combinatorial theory of overfitting

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — a finite *universe set* of objects;
$\mathbb{A} = \{a_1, \ldots, a_D\}$ — a finite set of *classifiers*;

$I(a, x) = [\text{classifier } a \text{ makes an error on object } x]$;

Loss matrix of size $L \times D$ with distinct columns:

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	0	0	1	$\cdots$	1	X — observable
$\ldots$	0	0	0	0	1	1	$\cdots$	1	(training) sample
x_ℓ	0	0	1	0	0	0	$\cdots$	0	of size ℓ
$x_{\ell+1}$	0	0	0	1	1	1	$\cdots$	0	$\bar{X}$ — hidden
$\ldots$	0	0	0	1	0	0	$\cdots$	1	(validation) sample
x_L	0	1	1	1	1	1	$\cdots$	0	of size $k = L - \ell$

Weak probability assumption:

All partitions $X \sqcup \bar{X} = \mathbb{X}$ are equiprobable.

Generalization ability estimation problem

$n(a, X) = \sum_{x \in X} I(a, x)$ — *number of errors* $a \in \mathbb{A}$ on the sample $X \subset \mathbb{X}$;

$\nu(a, X) = \frac{1}{|X|} n(a, X)$ — *error rate* of a on the sample X;

Def. *Learning algorithm* $\mu \colon 2^{\mathbb{X}} \to \mathbb{A}$ takes a training sample $X \subset \mathbb{X}$ and returns a classifier $a \in \mathbb{A}$.

Def. *Empirical risk minimization* (ERM):

$$\mu X \in A(X) = \operatorname{Arg} \min_{a \in \mathbb{A}} \nu(a, X).$$

Probability of overfitting:

$$Q_\varepsilon = \mathsf{P}(|\nu(\mu X, \bar{X}) - \nu(\mu X, X)| \geqslant \varepsilon).$$

Complete cross-validation:

$$CCV = \mathsf{E}\nu(\mu X, \bar{X}).$$

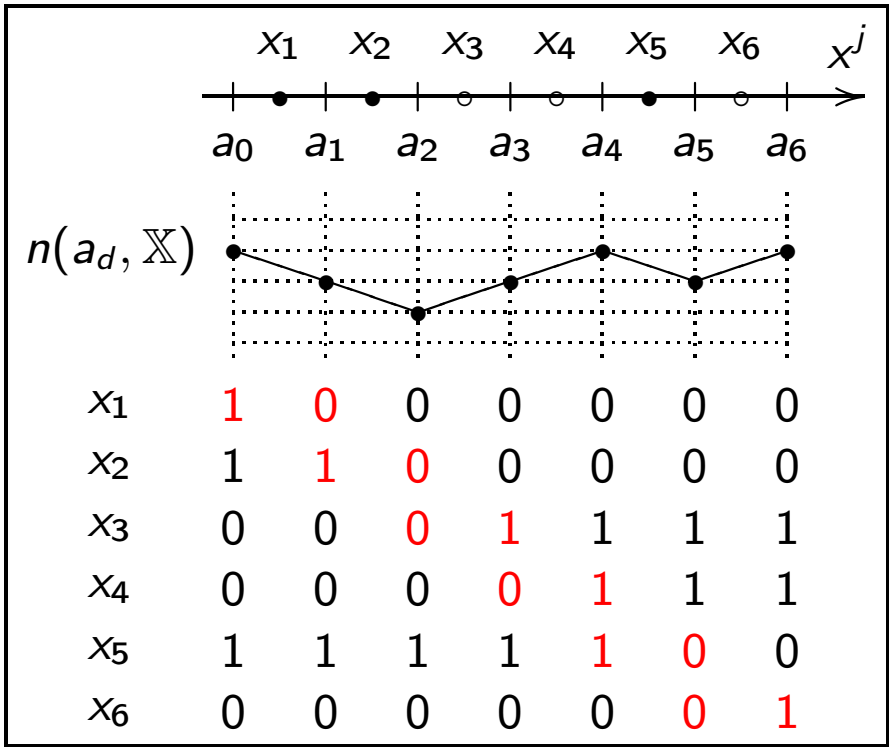
One-dimensional threshold classifiers

$x = (x^1, \ldots, x^n) \in \mathbb{X}$, $\mathbb{Y} = \{0, 1\}$.

Let $x^j \in \mathbb{R}$.

Generate set $\mathbb{A}$ by varying θ in the threshold classifier

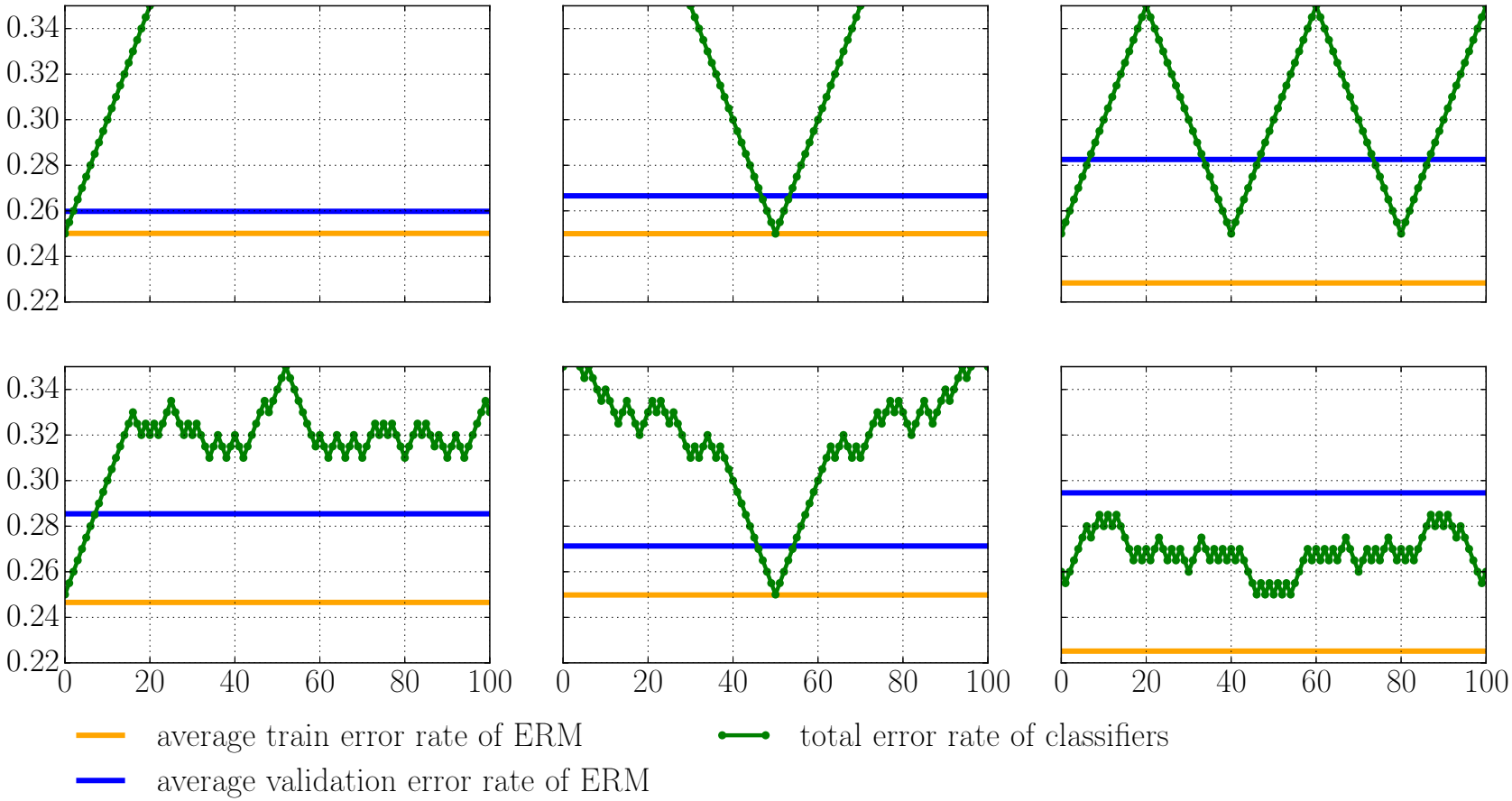
$$\mathbb{A} = \{a_\theta(x) = [x^j \geqslant \theta], \theta \in \mathbb{R}\}.$$



x_1	1	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0
x_3	0	0	0	1	1	1	1
x_4	0	0	0	0	1	1	1
x_5	1	1	1	1	1	0	0
x_6	0	0	0	0	0	0	1

Dependence of overfitting on the classes structure

Overfitting $\delta(\mu, X) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$.



Parameters: $L = 200$, $\ell = 100$, $m = 50$, Monte-Carlo method on $N = 10^5$ partitions.

Vapnik-Chervonenkis bound

Hypergeometric distribution function

$$\mathcal{H}_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \; z = 0, \ldots, m.$$

Theorem (1971)

For all $\mathbb{X}$, μ, A and $\varepsilon \in (0,1)$

$$Q_\varepsilon \leqslant \mathsf{P} \max_{a \in \mathbb{A}} \big[\delta(a, X, \bar{X}) \geqslant \varepsilon\big] \leqslant$$

$$\leqslant |\mathbb{A}| \cdot \max_m \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}\big(m - \varepsilon k\big)\right) \leqslant \frac{3}{2}|\mathbb{A}| e^{-\varepsilon^2 L}.$$

Vapnik V. N., Chervonenkis A. Ya. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 16(2):264–280

Splitting-Connectivity bound

Hamming distance: $\rho(a,b) = \sum_{x \in \mathbb{X}} [I(a,x) \neq I(b,x)]$;

Partial order $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ for all $x \in \mathbb{X}$;

Precedence $a \prec b$: $a \leqslant b$ и $\rho(a,b) = 1$.

Theorem (2011)

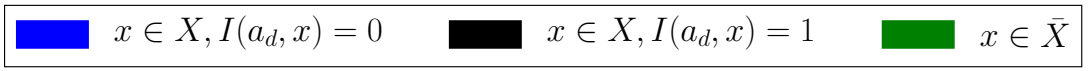
For all motononic methods μ, for all $\mathbb{X}$, $\mathbb{A}$ and $\varepsilon \in (0,1)$

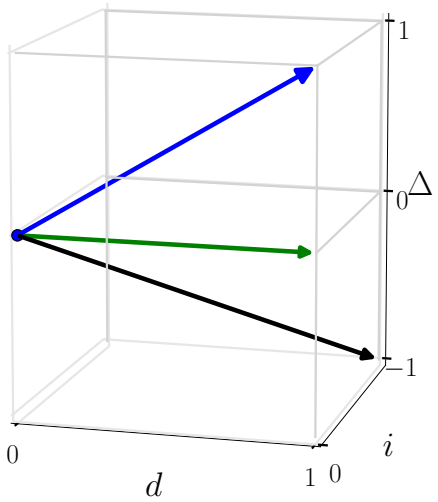
$$Q_\varepsilon \leqslant \sum_{a \in \mathbb{A}} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\, m-q} \left(\tfrac{\ell}{L} \left(m - \varepsilon k \right) \right),$$

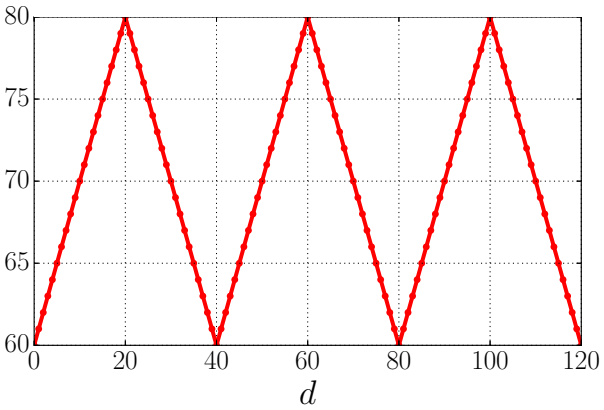
where $u \equiv u(a) = \left| \{ b \;\middle|\; a \prec b \} \right|$,
$q \equiv q(a) = \left| \left\{ x \in \mathbb{X} \;\middle|\; \exists b \in \mathbb{A} \colon\, b < a,\; I(b,x) < I(a,x) \right\} \right|$,
$m \equiv m(a) = n(a, \mathbb{X})$.

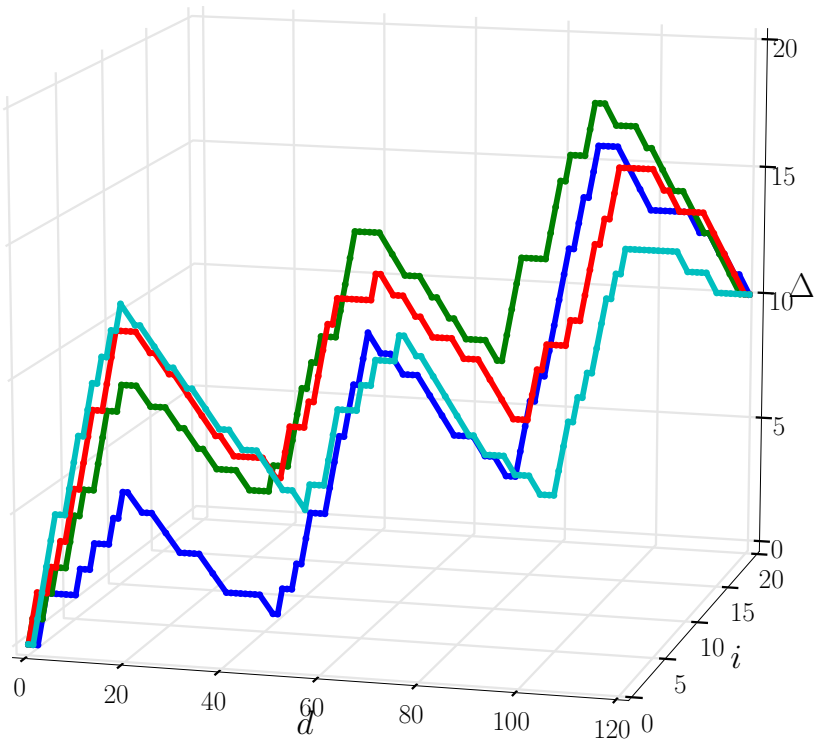
Vorontsov K. V., Ivahnenko A. A. 2011. Tight combinatorial generalization bounds for threshold conjunction rules. *PReMI'11*. LNCS. 66–73

Exact combinatorial bound









Result: Algorithm of *CCV* computation with polynomial complexity $O(L^5)$.

Surrogate modelling

1. Generate representative training sample of pairs $(\mathbb{A}, CCV)$, where $\mathbb{A}$ is a set of threshold classifiers on a various $\mathbb{X}$;
2. Generate features, describing objects $\mathbb{A}$;
3. Build regression model, approximating CCV.

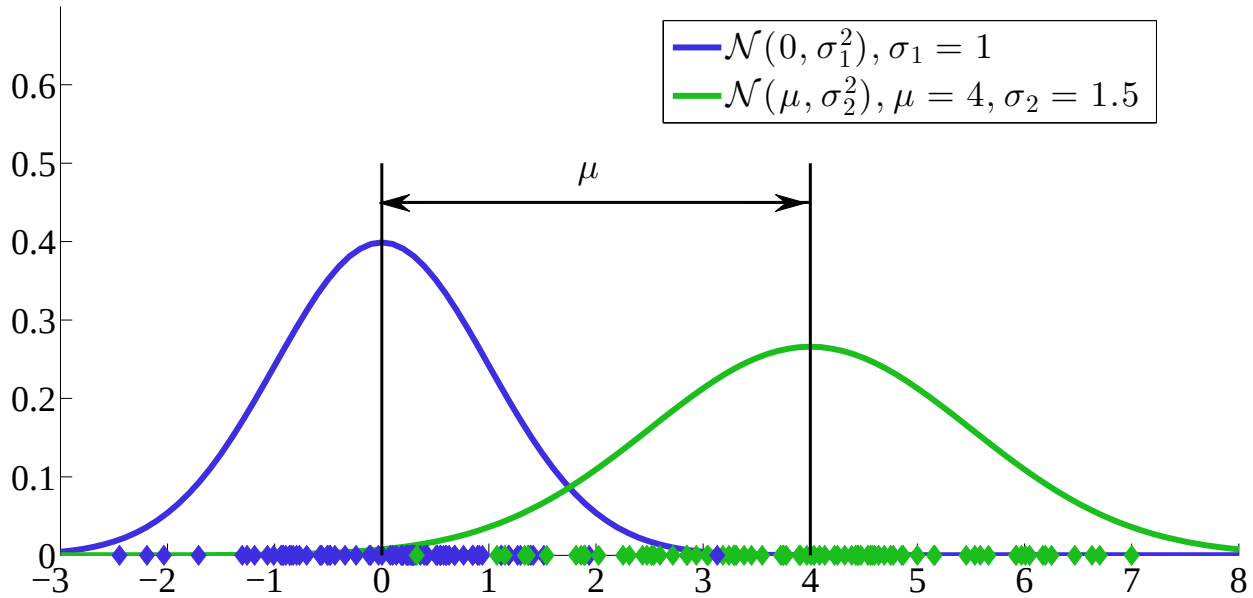
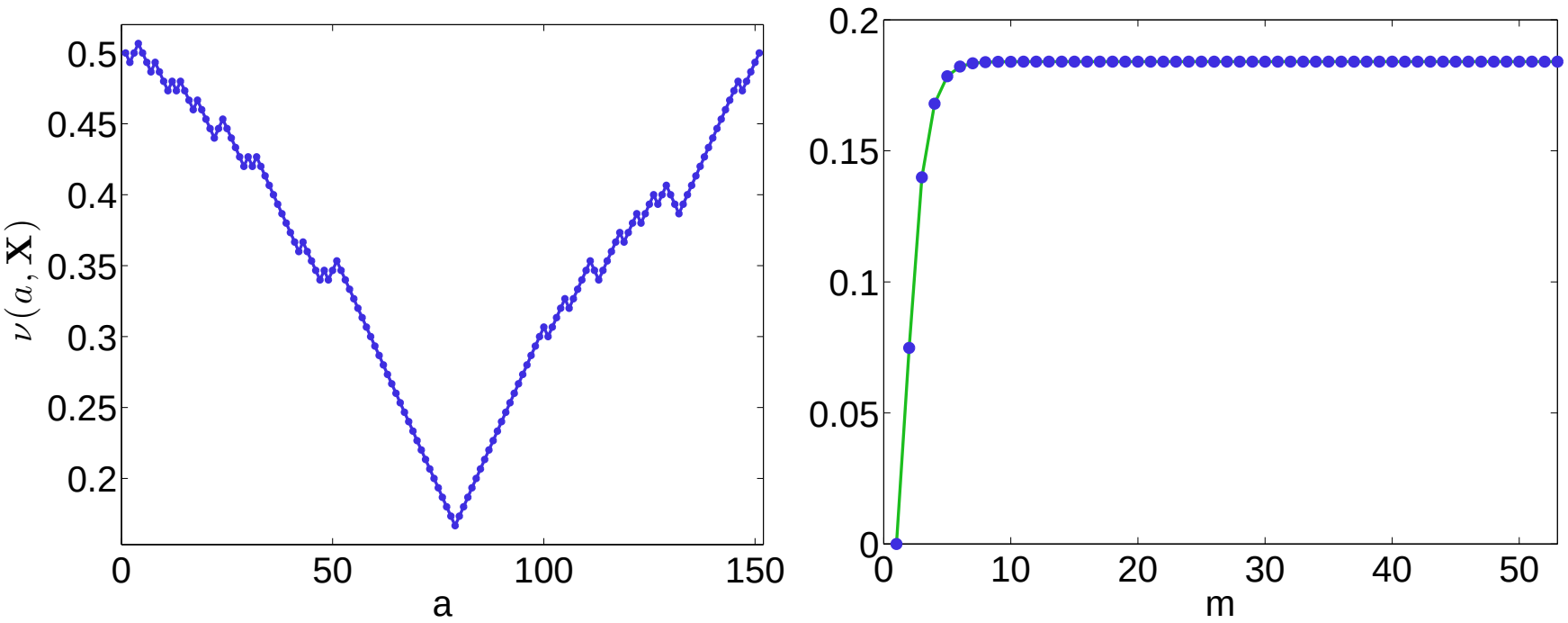


Figure: Example of $\mathbb{X}$ used for generating $\mathbb{A}$ from training sample. $\mathbb{X}$ is a one-dimensional set of two Gaussian classes. Varied parameters are μ, σ_1, σ_2, $L = |\mathbb{X}|$ and $\ell = |X|$

Generating features

Layer m $\quad \{a \in \mathbb{A} : n(a, \mathbb{X}) = m\}$.

- Classifiers from the same layer have the same contribution;
- The contribution depends on the layer.

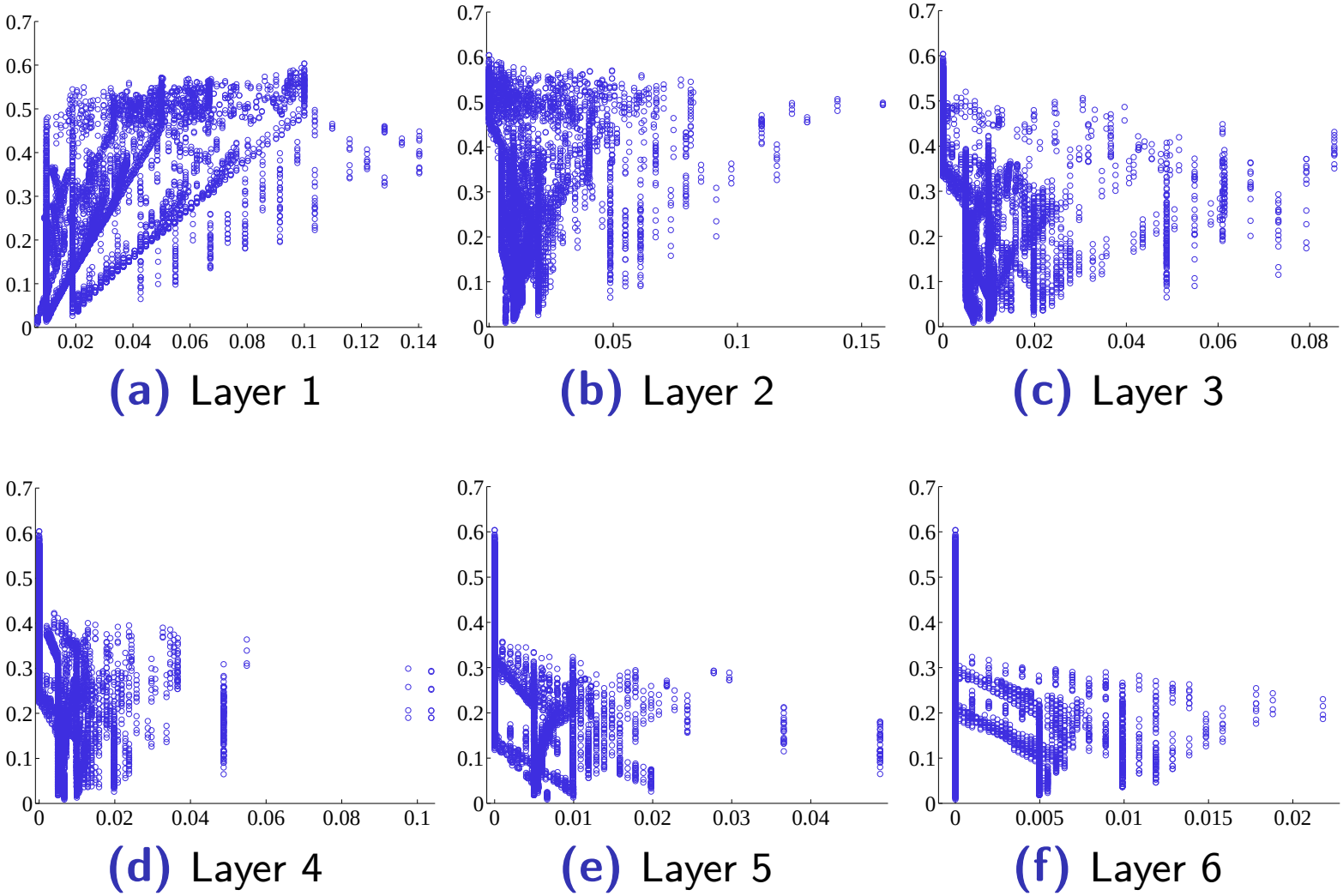


(a) Example of an object $\mathbb{A}$ **(b)** Cumulative contribution

Figure: Cumulative contribution of the first layers in $\mathbb{A}$

Training sample: 7000 objects with 53 features.

CCV dependence on the layer sizes

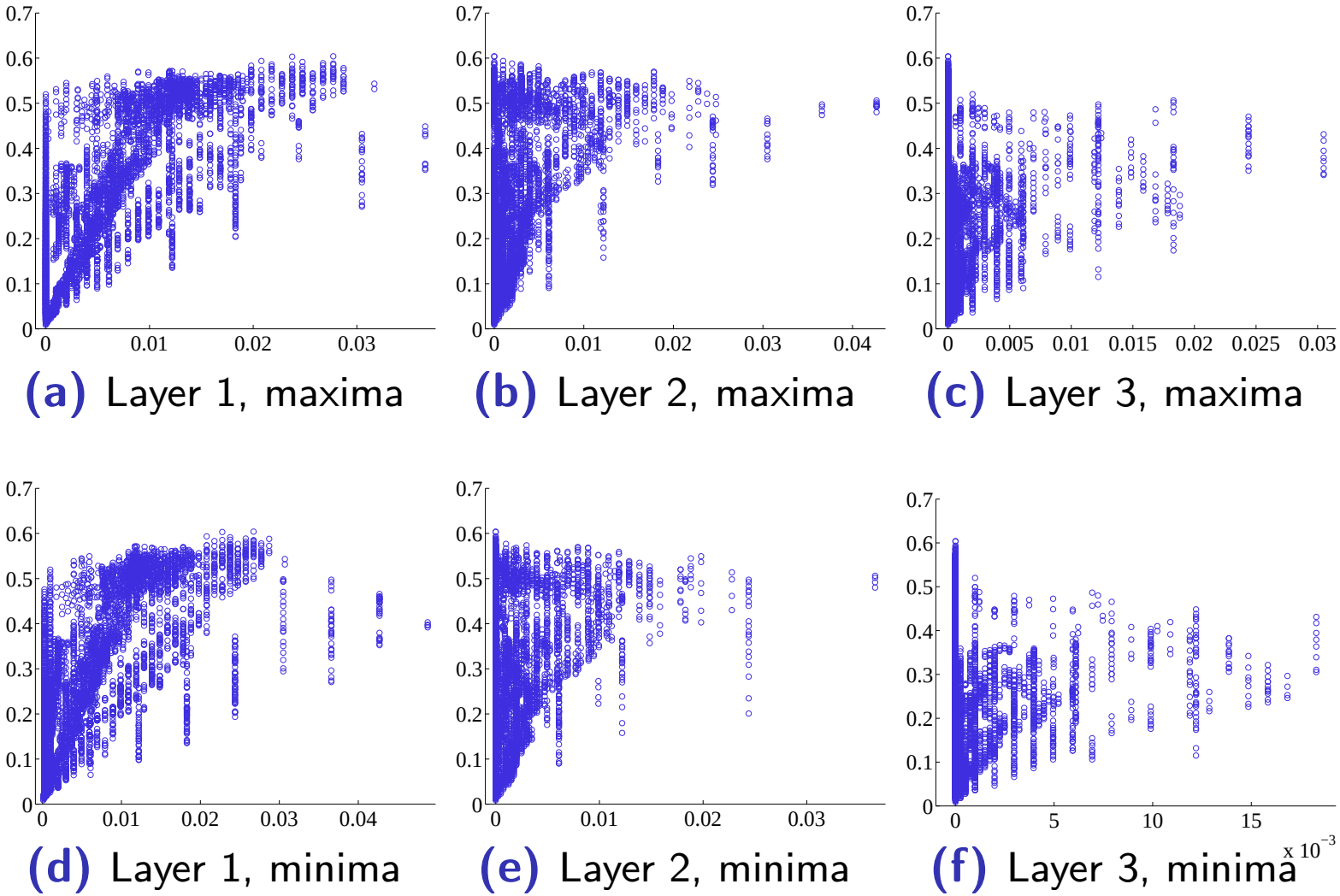


(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3

(d) Layer 4 **(e)** Layer 5 **(f)** Layer 6

Figure: *CCV* dependence on the classifiers count in each layer

CCV dependence on maxima and minima count in each layer

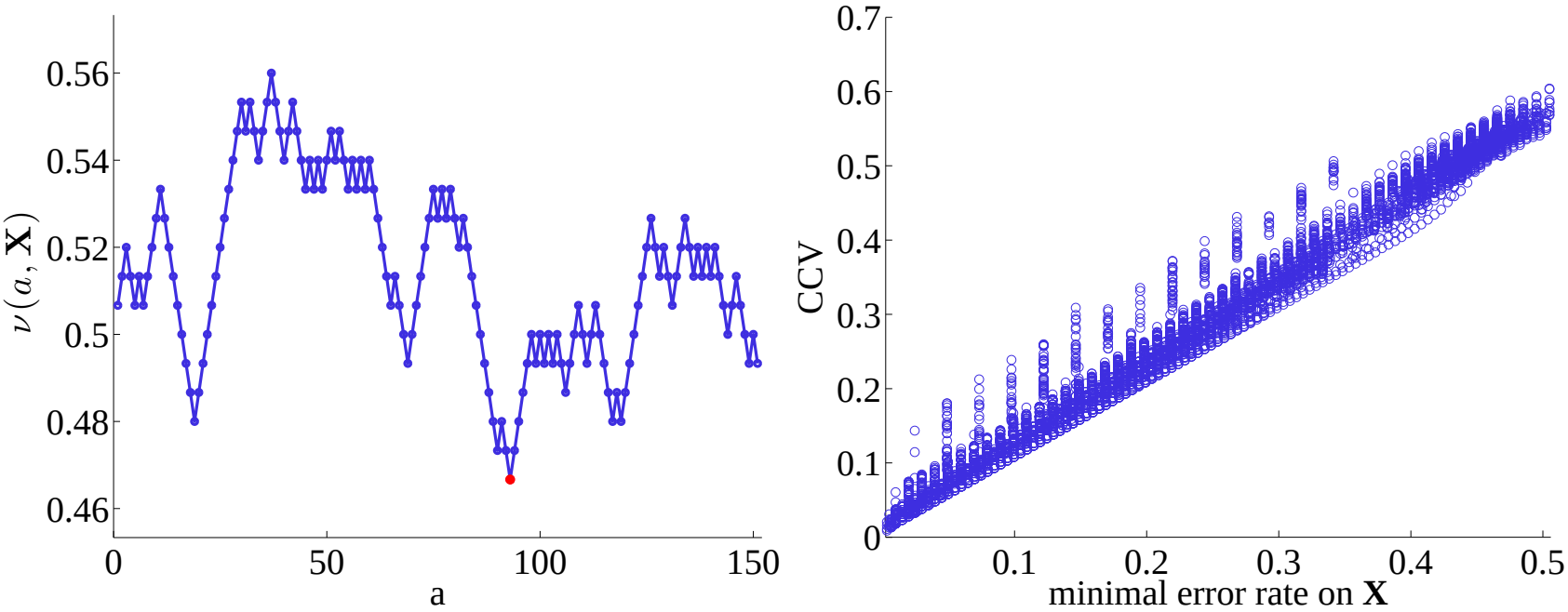


(a) Layer 1, maxima **(b)** Layer 2, maxima **(c)** Layer 3, maxima

(d) Layer 1, minima **(e)** Layer 2, minima **(f)** Layer 3, minima

Figure: *CCV* dependence on the maxima and minima count in each layer

CCV dependence on the minimal error rate on the universe set



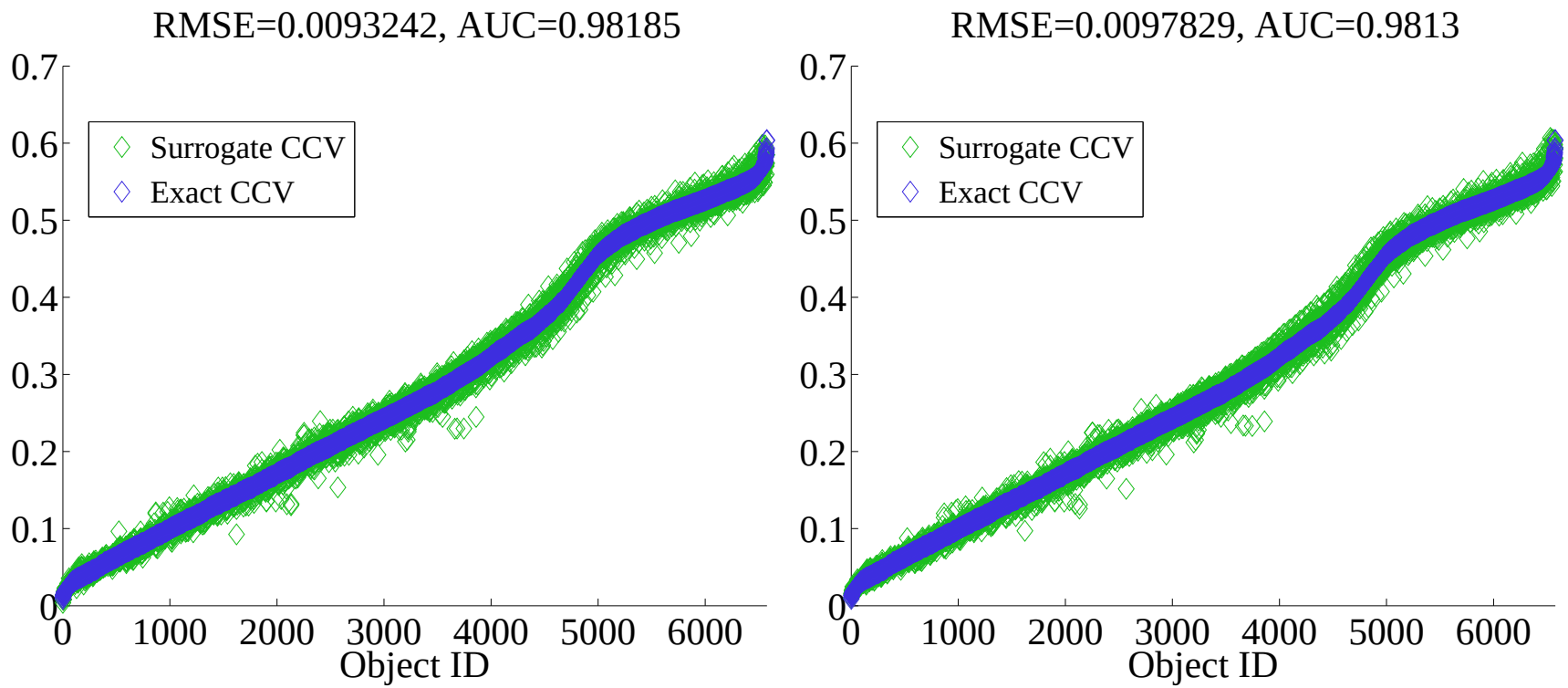
(a) Example of the object $\mathbb{A}$ **(b)** *CCV* dependence on the feature

Figure: *CCV* dependence on the minimal error rate in $\mathbb{A}$

Linear regression vs Non-negative least squares regression

NNLS:

$$AX = Y \quad \Rightarrow \quad X = \arg\min_{X \geqslant 0} \tfrac{1}{2} X^T (A^T A) X - Y^T A^T X.$$

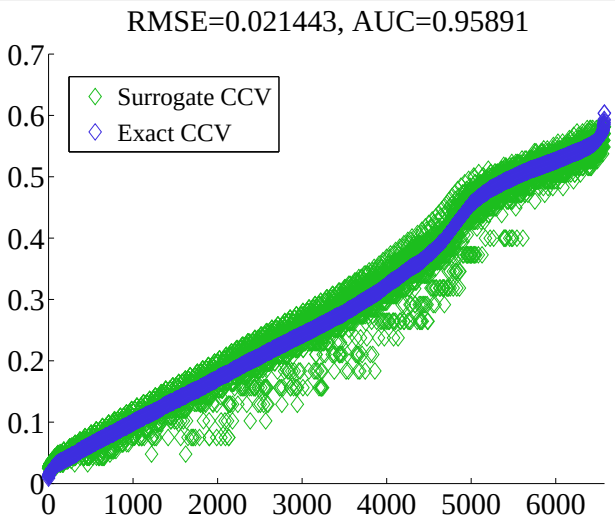


(a) Linear Regression, 43 features

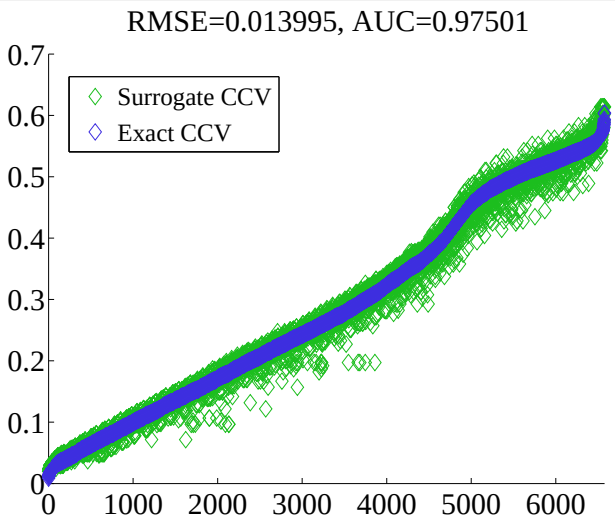
(b) NNLS, 24 features

Figure: Linear Models, 10×10-fold cross-validation. Objects are sorted by exact *CCV*

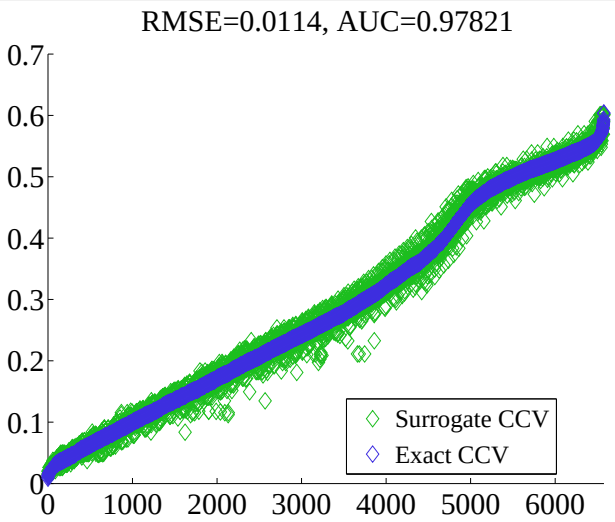
Features importance (NNLS)



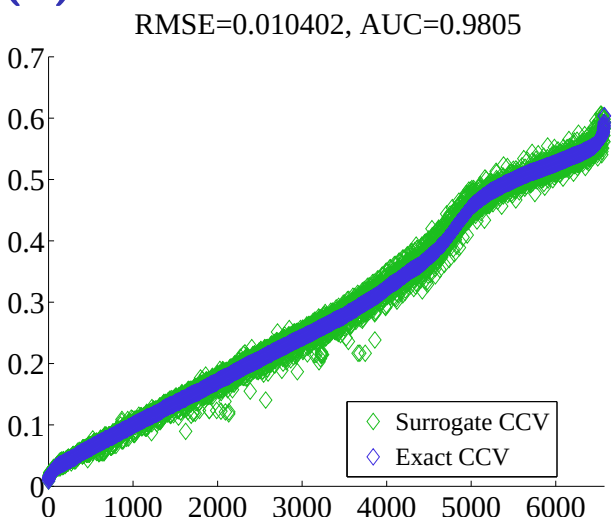
(a) Minimal error rate on $\mathbb{X}$



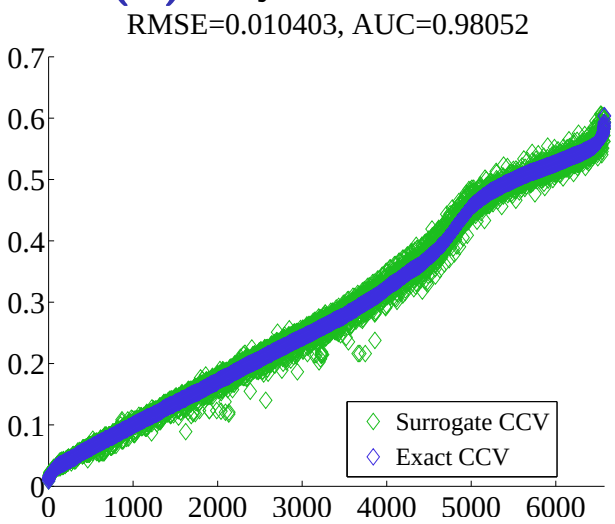
(b) Layer size, L1



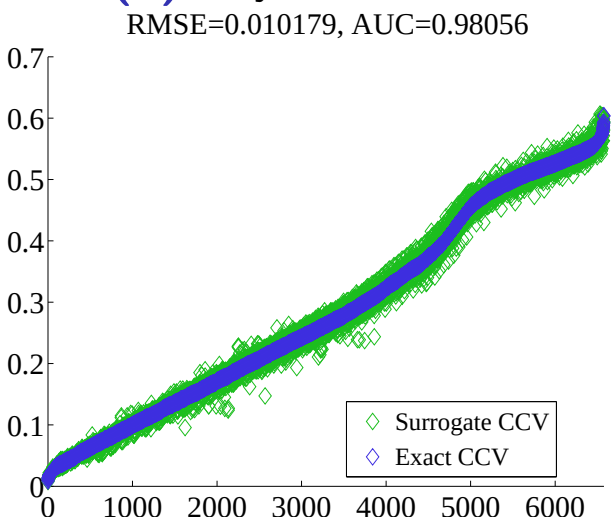
(c) Layer size, L2



(d) The ratio $\frac{\ell}{L-\ell}$



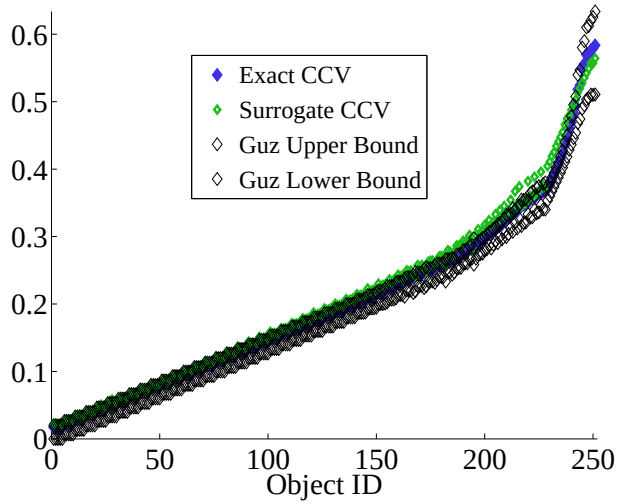
(e) Minima count, L1



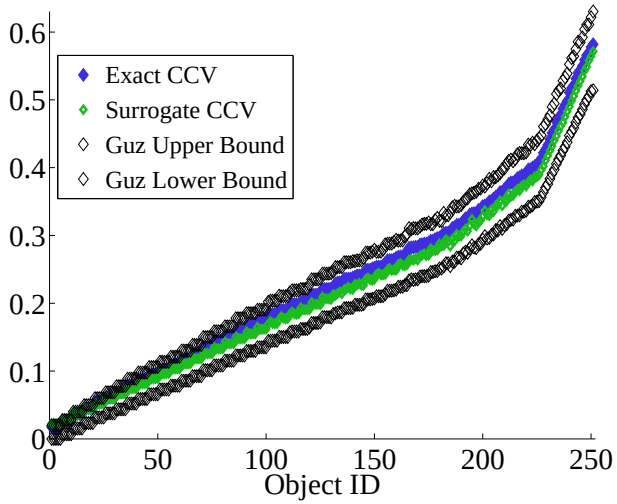
(f) Layer size, L4

Comparison with Guz Upper and Lower bounds (Guz, 2011)

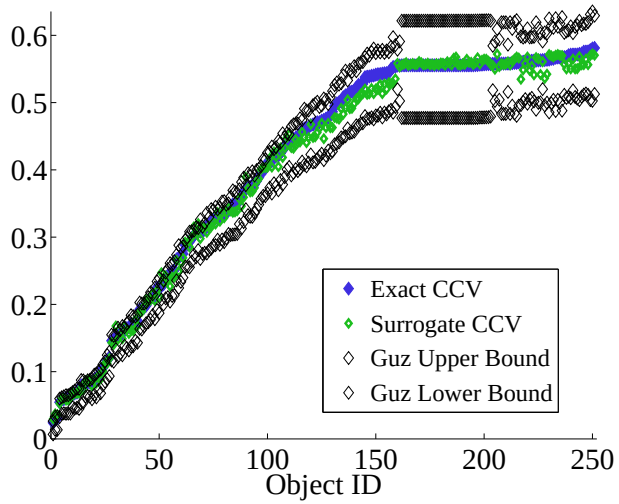
Upper and Lower bounds of CCV are computable in time $O(L^3)$:



(a) Noise far from the classes boundary



(b) Noise near the classes boundary



(c) Two Gaussian classes

Guz I.S. 2011. Constructive evaluation of the complete cross-validation for threshold classification. *Mathematical Biology and Bioinformatics* 6(2):173-189

Uspenskiy's Informational Analysis

Input: A detailed raw ECG signal.

1. Discretization:
   $S = (s_n)_0^{N-1}$ – text string on $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$.
2. Vectorization: *triplet frequency* $w = (w_0, w_1, w_2)$

$$p_w(S) = \frac{1}{N-2} \sum_{i=0}^{N-3} \big[(s_i, s_{i+1}, s_{i+2}) = (w_0, w_1, w_2)\big].$$

Output: $\{p_w(S)\}_w$. Total $|\mathcal{A}|^3 = 216$ features.

Naïve Bayes classifier

Patients are described by features $\{p_w(S)\}$, w – words.
Class labels: 1 – has disease, 0 – is healthy.

Naïve Bayes classifier

$$a(S) = \left[\ln \frac{\pi_1(S)}{\pi_0(S)} \geqslant \beta\right] = \left[\sum_w \gamma_w p_w(S) \geqslant \beta\right],$$

where $\pi_i(S)$ is a probability density function of class i.

Baseline: Select top–K by information criterion F_w, where

$$F_w = \frac{1}{|X_1|} \sum_{S \in X_1} p_w(S), \quad X_1 - \text{class 1}.$$

Feature selection methods based on the generalization bounds

Ind:

1. Make set of candidates: select top–K^* by criterion F_w;
2. Select top–K by criterion CCV from candidates.

Add: Greedy add from top–K^* by criterion F_w :

for all $j = 1, \ldots, K^*$

Given features $p_1, \ldots, p_{j-1}$ and weights $\gamma_1, \ldots, \gamma_{j-1}$

For each p from $K^* - j + 1$ candidates vary β or γ in

$$[\gamma_1 p_1(S) + \ldots \gamma_{j-1} p_{j-1}(S) + \gamma p(S) \geqslant \beta].$$

Select p by minimizing CCV.

Comparison of feature selection methods

Run 10x10-fold cross-validation. X is training sample, $\bar{X}$ is validation sample.
Quality metric: Area under ROC-curve on hold-out test sample T.

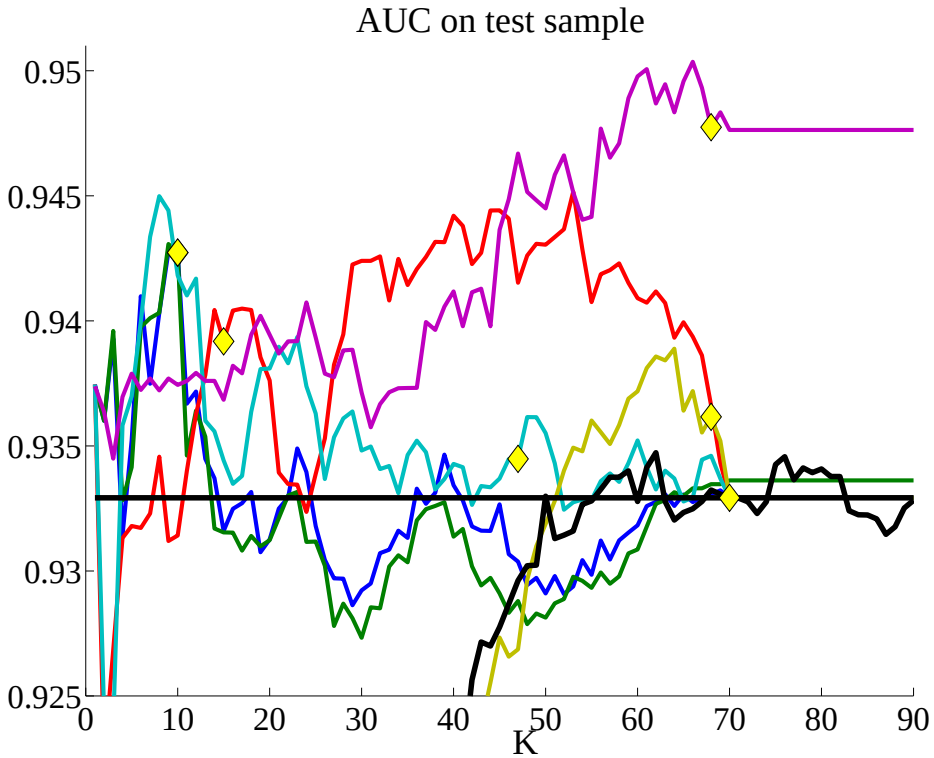


Figure: Test quality AUC_T by K

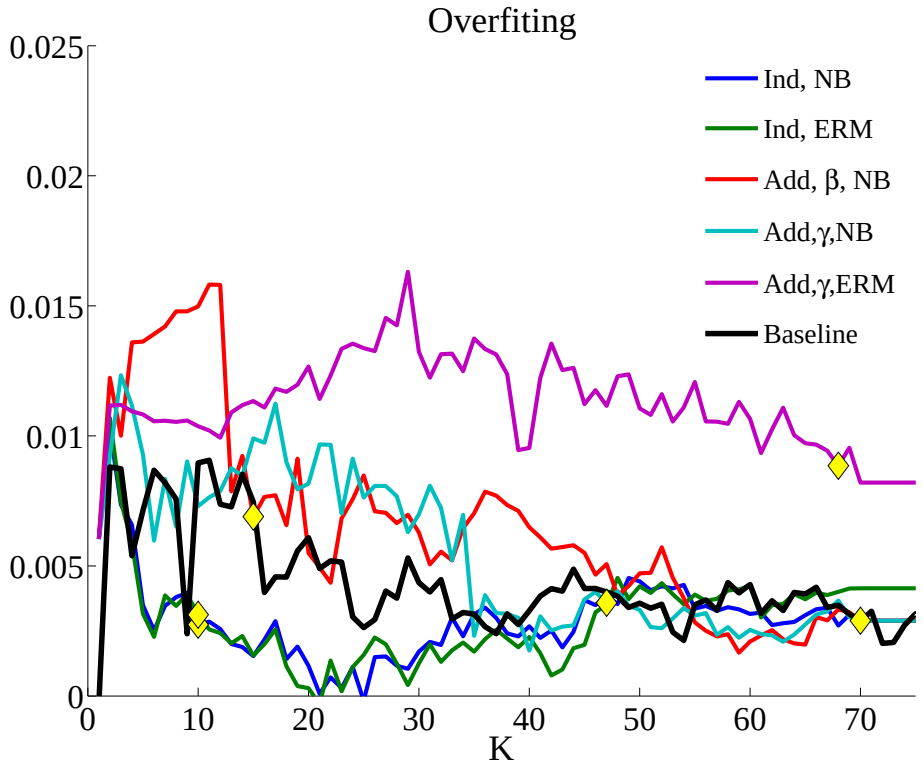


Figure: Overfitting $AUC_X - AUC_{\bar{X}}$ by K

The disease is chronic cholecystitis. Parameter K^* equals 70.

Results of using the surrogate modelling CCV in feature selection

Select optimal K that maximizes AUC on validation sample:

Disease	Baseline	Ind,NB	Ind,erm	Add,β,NB	Add,γ,NB	Add,γ,erm
benign prostatic hyperplasia	**95.20**	94.95	94.56	94.30	94.75	94.64
hypertension	**94.52**	93.22	94.40	93.81	93.39	93.93
femoral head necrosis	96.64	97.23	97.23	96.55	96.23	**97.59**
gastritis hyperacid	92.61	**95.87**	95.69	93.27	92.34	95.16
coronary heart disease	96.18	95.69	95.57	96.43	96.01	**96.72**
gastritis hypoacid	93.07	92.69	91.93	**93.86**	92.63	93.64
chronic cholecystitis	93.09	93.79	94.33	94.34	93.01	**94.62**
urolithiasis	92.37	93.55	**93.90**	93.41	92.27	93.04
cancer	**94.08**	93.00	93.03	94.05	92.50	93.29
diabetes	94.78	95.46	95.41	**96.45**	95.99	95.98
nodular goiter thyroid	92.88	93.30	93.08	**94.20**	92.20	93.96
cholelithiasis	96.32	97.00	96.84	97.61	97.34	**97.69**

Table: AUC on test sample for optimal K.

Concluding remarks

Concluding remarks

Surrogate modelling of the generalization bounds helps to reduce the number of features and to improve the classification quality of the linear Naïve Bayes classifier.

Open problems

- Use non-linear surrogate modelling (MVR Composer);
- Create new features for surrogate modelling;
- Compare approximate CCV to the existing bounds as the feature selection criterion in the greedy method.