Towards tight generalization bounds (combinatorial approach)

Konstantin Vorontsov
(vokov@forecsys.ru, http://www.ccas.ru/voron)

Computing Centre of Russian Academy of Sciences,
Vavilova 40, 119991, Moscow, Russian Federation

Pattern Recognition and Image Analysis:
New Information Technologies
September, 14-20, 2008
Nizhny Novgorod, Russian Federation

Contents

1. **Generalization bounds**
   - The probability of overfitting
   - Vapnik-Chervonenkis bounds
   - Data dependent bounds
2. **VC bounds: measuring factors of overestimation**
   - Weak Probability Axiomatic (WPA)
   - Vapnik-Chervonenkis bounds under WPA
   - Causes of overestimation of the VC bound
   - Empirical results
3. **Splitting and similarity**
   - Overfitting of two-element set of classifiers
   - Overfitting of chain of classifiers
   - Conclusions

Definitions and notation

Training sample: $X^\ell = \{x_i\}_{i=1}^\ell \subset \mathbb{X}$.
Learning algorithm μ: $X^\ell \mapsto a$, where $a \in A$ is a *classifier*.
Binary loss function $I(a,x) = $ [classifier a makes an error on x].
Binary loss vector of a classifier a on a sample X^ℓ:

$$\vec{a}(X^\ell) = \bigl(I(a,x_i)\bigr)_{i=1}^\ell.$$

Frequency of errors of classifier a on a sample X^ℓ

$$\nu(a,X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} I(a,x_i).$$

Testing sample: $X^k = \{x_i\}_{i=1}^k \subset \mathbb{X}$.
Overfitting of a learning algorithm μ with respect to X^ℓ, X^k:

$$\delta(\mu, X^\ell, X^k) = \nu\bigl(\mu(X^\ell), X^k\bigr) - \nu\bigl(\mu(X^\ell), X^\ell\bigr).$$

Problem: obtain an upper bound of the *probability of overfitting*

$$\mathsf{P}_{X^\ell,X^k}\{\delta(\mu,X^\ell,X^k) > \varepsilon\} \leqslant \eta(\varepsilon), \qquad \eta(\varepsilon) - ?$$

Test set bound

Theorem (a form of the Law of Large Numbers)

For any fixed classifier a and any probability measure P *over* $X^L = X^\ell \cup X^k$ *the observable frequency* $\nu(a, X^\ell)$ *predicts the unknown frequency* $\nu(a, X^k)$*:*

$$\mathsf{P}_n\bigl\{\delta(a, X_n^\ell, X_n^k) \geqslant \varepsilon\bigr\} \leqslant H_L^\ell(\varepsilon),$$

$H_L^\ell(\varepsilon) = \max\limits_{m=0,\dots,L} \sum\limits_{t=s_0}^{s(\varepsilon)} \dfrac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}$ *is an upper bound of the left tail of hypergeometric distribution,* $s_0 = (m-k)_+$, $s(\varepsilon) = \bigl\lfloor \frac{\ell}{L}(m - \varepsilon k)\bigr\rfloor$.

- ⊕ The bound is tight (moreover, an exact variant exists).
- ⊖ But it gives no recommendations for μ construction.

Vapnik-Chervonenkis bounds [1968–1971]

For any set of classifiers A, any prob.measure P over $X^L = X^\ell \cup X^k$

$$\mathsf{P}_{X^L}\bigl\{\delta(\mu, X^\ell, X^k) > \varepsilon\bigr\} \leqslant \mathsf{P}_{X^L}\Bigl\{\sup_{a\in A} \delta(a, X^\ell, X^k) > \varepsilon\Bigr\} \leqslant$$

$$\leqslant \sum_{\vec{a}\in A(X^L)} \mathsf{P}_{X^L}\bigl\{\delta(a, X^\ell, X^k) > \varepsilon\bigr\} \leqslant \Delta^A(L)\cdot H_L^\ell(\varepsilon) \leqslant$$

$$(\text{if } \ell = k) \;\leqslant \Delta^A(L)\cdot 1.5\, e^{-\varepsilon^2\ell},$$

$A(X^L) = \bigl\{\vec{a}(X^L) \bigm| a \in A\bigr\}$ is a set of loss vectors induced by A.

$\Delta^A(L) = \max\limits_{X^L} |A(X^L)|$ is a *shatter coefficient* of the set A,

$\Delta^A(L) \leqslant 1.5\,\frac{L^h}{h!}$, where h is the *VC-dimension* of the set A.

- ⊕ The bound leads to the Structural Risk Minimization method.
- ⊖ But it is highly overestimated and almost useless in practice.

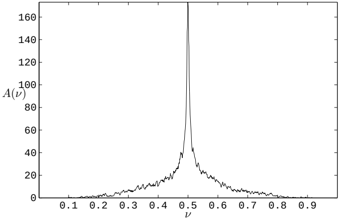
The causes of overestimation: the recent understanding

- The «worst case» bound does not take into account:
  - peculiarities of the data X^L;
  - peculiarities of the learning algorithm μ.
- The *effect of splitting* (or *localization*): the worse classifier is, the less is a chance that it would be obtained from learning. The set A is split into data-dependent subsets.
- The «union bound» $P(S_1 \cup \dots \cup S_\Delta) \leqslant P(S_1) + \dots + P(S_\Delta)$, is loose when events $S_d = \bigl\{\delta(a_d, X^\ell, X^k) > \varepsilon\bigr\}$ are similar.
- The exponent factor $e^{-\varepsilon^2\ell}$ is also an upper bound.

40 years ago: the problem remains open

- Uniform convergence [Vapnik, Chervonenkis, 1968]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Data-dependent bounds [Haussler, 1992; Bartlett, 1998]
- Connected function classes [Sill, 1995]
- Similar classifiers VC bounds [Bax, 1997]
- Self-bounding learning algorithms [Freund, 1998]
- Microchoice bounds [Langford, Blum, 2001]
- Algorithmic stability [Bousquet, Elisseeff, 2002]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Shell bounds [Langford, 2002]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]

Shell bounds: main ideas



- Most classifiers $a \in A$ are concentrated near to $\nu(a, X^L) = 0.5$
- Only classifiers from the left tail of the histogram have chances to be choused by ERM: $\nu(a, X_n^\ell) \to \min\limits_{a \in A}$
- The bound is too complicated, requires Monte-Carlo simulation, and not tight enough in practice.

John Langford. Quantitatively Tight Sample Complexity Bounds. PhD (Carnegie Mellon). 2002.

Similar classifiers VC bounds: main ideas

Theorem

Suppose the set of loss vectors $\bigl\{\vec{a}(X^L) \bigm| a \in A\bigr\}$ is clusterized with the Hamming distance on S clusters, each of the radius r. Then

$$\mathsf{P}_{X^L}\bigl\{\delta(a, X^\ell, X^k) > \varepsilon + r/\ell\bigr\} \leqslant S \cdot H_L^\ell(\varepsilon).$$

- If A is separating hyperplanes, then $S = \Delta^A(L)/(2r+1)$.
- Optimization over r (open problem: how r depends on the dimension of the object space X?)
- The bound is not tight, even after optimization over r.

Bax E. Similar Classifiers and VC Error Bounds. CalTech-CS-TR97-14, June 1997. `citeseer.ist.psu.edu/bax97similar.html`

Connected function classes

Definition

The set A is *connected* if for any $\vec{a} \in A(X^L)$ with probability 1 exists $\vec{a}_1 \in A(X^L)$ such that Hamming distance $\|\vec{a} - \vec{a}_1\| = 1$.

- SVM, two layer ANN, RBF, etc. are connected.
- **Theorem:** *if A is connected, then*

$$\mathsf{P}_{X^L}\{\delta(a, X^\ell, X^k) > \varepsilon\} \leqslant \frac{1}{\sqrt{\pi L}} \Delta^A(L) \cdot H_L^\ell(\varepsilon).$$

- The bound is not tight. It differs a little from the VC bound.

Sill J. Generalization Bounds for Connected Function Classes. 1995. `http://citeseer.ist.psu.edu/127284.html`

Sill J. Monotonicity and Connectedness in Learning Systems. PhD thesis, CalTech, 1998.

Motivation for measuring factors of overestimation

- **Ultimate aim (OPEN PROBLEM)**
  to obtain tight and useful bounds.
- **Immediate aim (DONE — see below)**
  to understand the causes of overestimation by comparing them quantitatively in experiments on real data sets
- **Problem:**
  Standard probabilistic techniques used to obtain bounds induce a sequence of uncontrollably overestimated inequalities
- **Why so?**
  It is usual to introduce and handle unobservable probabilities that can be hardly measured
- **What is proposed:**
  A theory that handles only measurable quantities

Weak (minimalistic, combinatorial) Probability Axiomatic (WPA)

1. $X^L = \{x_i\}_{i=1}^L$ — a given finite set of objects.
2. All partitions $X^L = X_n^\ell \cup X_n^k$ are *equally probable*, where $n = 1, \dots, N$, $N = C_L^k$, $L = \ell + k$;
   X_n^ℓ — observable training subset;
   X_n^k — hidden testing subset.

Overfitting at n-th partition: $\delta_n(\mu) \equiv \delta(\mu, X_n^\ell, X_n^k)$.

Probability of overfitting is defined as the "fraction of partitions":

$$\mathsf{P}_n\{\delta_n(\mu) > \varepsilon\} = \frac{1}{N}\sum_{n=1}^{N}\bigl[\delta_n(\mu) > \varepsilon\bigr].$$

Remark. The notion of probability is introduced without theory of measure and without passage to the limit $L \to \infty$.

Advantages of Weak Probability Axiomatic

- Not redundant. Can give not asymptotical exact bounds.
- Any probability can be measured empirically:
  $$\hat{\mathsf{P}}_n\{\delta_n > \varepsilon\} = \frac{1}{|N'|}\sum_{n\in N'}[\delta_n > \varepsilon] \overset{N'\to N}{\to} \mathsf{P}_n\{\delta_n > \varepsilon\}.$$
- Easy transition to Kolmogorov's axiomatic:
  if $\mathsf{P}_n\{\delta(X_n^\ell, X_n^k) > \varepsilon\} \leqslant \eta(\varepsilon, X^L)$,
  then $\mathsf{P}_{X^L}\{\delta(X^\ell, X^k) > \varepsilon\} \leqslant \mathsf{E}_{X^L}\eta(\varepsilon, X^L)$.
- Sufficient to prove fundamental facts:
  - the Law of Large Numbers (exact bound);
  - Kolmogorov-Smirnov criterion (also exact);
  - many statistical hypothesis tests (order statistics etc.);
  - Vapnik-Chervonenkis generalization bounds (see later);

The test set bound (Law of Large Numbers) under WPA

Consider a fixed classifier a, $\nu(a, X^L) = m/L$.

Theorem (exact bound)

The observable frequency $\nu\bigl(a, X^\ell\bigr)$ predicts the hidden frequency $\nu\bigl(a, X^k\bigr)$:

$$\mathsf{P}_n\bigl\{\delta(a, X_n^\ell, X_n^k) \geqslant \varepsilon\bigr\} = H_L^{\ell,m}\left(s(\varepsilon)\right),$$

where $H_L^{\ell,m}(s) = \sum\limits_{t=s_0}^{s} \dfrac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}$ *— the left tail of hypergeometric distribution;* $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$; $s_0 = \max\{0, m-k\}$.

Vapnik-Chervonenkis bounds under WPA

For any learning algorithm μ and any set X^L:

$$Q_\varepsilon = \mathsf{P}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\} \leqslant$$
$$\leqslant \sum_{m=1}^{L} D_m \cdot H_L^{\ell,m}(s(\varepsilon)) \leqslant$$
$$\leqslant \Delta_L^\ell \cdot H_L^\ell(\varepsilon) \overset{\ell=k}{\leqslant} \Delta^A(L) \cdot 1.5\, e^{-\varepsilon^2 \ell};$$

$\Delta_L^\ell(\mu, X^L)$ — *local shatter coefficient* (LSC) — shatter coefficient of the set of classifiers $\{a_n = \mu(X_n^\ell) \mid n = 1, \dots, N\}$;

$D_m(\mu, X^L)$, $m = 0, \dots, L$ — *shatter profile* — a sequence of shatter coefficients of the sets of classifiers having m errors on X^L:
$\{a_n = \mu(X_n^\ell) \mid \nu(a_n, X^L) = \frac{m}{L},\ n = 1, \dots, N\}$.

The effective local shatter coefficient

Conclusions:

- The exponential approximation $1.5\, e^{-\varepsilon^2 \ell}$ is avoided.
- The *splitting* of A is (partially) taken into account, but:
  — it's not clear, how to estimate D_m;
  — it's not clear, whether this will give a gain.
- The *similarity* of classifiers is not taken into account.

Idea: to estimate the causes of overestimation empirically

Definition

The *effective* local shatter coefficient (ELSC):

$$\hat{\Delta}_L^\ell(\varepsilon) = \frac{\hat{\mathsf{P}}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\}}{H_L^{\ell,m}(s(\varepsilon))} = \frac{\hat{\mathsf{P}}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\}}{\hat{\mathsf{P}}_n\{\delta(a, X_n^\ell, X_n^k) > \varepsilon\}}.$$

Causes of overestimation of the VC bound

The rate of overestimation can be factorized into 4 parts:

$$\frac{\Delta^A(L)\cdot\frac{3}{2}\,e^{-\varepsilon^2\ell}}{\hat{Q}_\varepsilon} = \underbrace{\frac{\Delta^A(L)}{\Delta_L^\ell}}_{r_1}\cdot\underbrace{\frac{\Delta_L^\ell}{\hat{\Delta}_L^\ell(\varepsilon)}}_{r_2(\varepsilon)}\cdot\underbrace{\frac{\hat{\Delta}_L^\ell(\varepsilon)\cdot H}{\hat{Q}_\varepsilon}}_{r_3(\varepsilon)}\cdot\underbrace{\frac{\frac{3}{2}\,e^{-\varepsilon^2\ell}}{H}}_{r_4(\varepsilon)}$$

where $H = \max_m H_L^{\ell,m}(s(\varepsilon))$.

Causes of overestimation:

- $r_1 \geqslant 1$: the disregard of splitting
- $r_2 \geqslant 1$: the disregard of similarity (due to union bound)
- $r_3 \geqslant 1$: the flat upper bound of the shatter profile
- $r_4 \geqslant 1$: exponent approximation of hypergeometric tail

Rule induction machine

- The *rule* is a predicate $\phi_y\colon X \to \{0,1\}$ that covers mainly objects of the class y.
- *Weighted voting* of rules:
$$a(x) = \arg\max_{y\in Y} \sum_{t=1}^{T_y} w_y^t \phi_y^t(x),$$
where $\phi_y^t(x)$ — t-the rule of the class y, w_y^t — its weight.
- *Rule learning algorithm* of class y:
$\mu_y\colon X^\ell \mapsto \bigl\{\phi_y^t(x) \bigm| t = 1,\dots,T_y\bigr\}$.
- Why the rule induction machine is convenient for the analysis of VC bounds overestimation:
    - the shatter coefficient $\Delta^A(L)$ is known;
    - the LSC $\Delta_L^\ell(\mu, X^L)$ can be easily (lower) bounded;
    - the ELSC $\hat\Delta_L^\ell(\varepsilon)$ can be easily estimated.

The experimental framework

- 7 tasks from UCI repository, two classes
- 20×2-fold cross-validation, $\ell = k$
- Learning algorithm Forecsys LogicPro®
  [Vorontsov, Kochedykov, Ivakhnenko]

			the average test set errors				
Task	L	n	C4.5	C5.0	RIPPER	SLIPPER	LogicPro
crx	690	15	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	18.1	22.7	19.0	18.9	12.0 ± 2.0

L — sample size; n — number of features.

Results

Causes of overestimation of the VC bound
(thresholds $\varepsilon_0, \varepsilon_1, \varepsilon_2$ correspond to the significance $\hat{Q}_\varepsilon = 0.05$, 0.1, 0.01).

Task	y	r_1	$r_2(\varepsilon_0)$	$r_3(\varepsilon_0)$	$r_4(\varepsilon_0)$	$\hat{\Delta}_L^\ell[\varepsilon_1, \varepsilon_2]$	$\hat{\Delta}_L^\ell(\varepsilon_0)$
crx	0	890	680	3.1	32.6	[10; 41]	24
	1	690	1700	1.6	11.6	[11; 180]	12
german	1	8 950	1500	1.7	10.9	[38; 530]	54
	2	37 000	9000	1.2	9.9	[1.0; 2.2]	1.9
hepatitis	0	23	280	13.4	9.5	[11; 148]	83
	1	55	680	2.4	22.5	[12; 27]	15
horse-colic	1	72	4500	2.1	7.2	[2; 9]	7
	2	140	3400	3.6	7.3	[3; 6]	6
hypothyroid	0	61 000	400	32.2	16.5	[3; 220]	21
	1	153 000	460	3.8	28.7	[2; 44]	30
promoters	0	94	340	5.9	9.8	[36; 230]	72
	1	150	790	3.4	6.9	[9; 22]	18

Conclusions

- The shatter coefficient $\hat{\Delta}_L^\ell$ should take value about 10^2 or less to bound be tight. No recent theory can provide so low estimates.
- The *effective local VC-dimension* (if we would like to define it) degenerates and becomes less that 1.
  Open problem 1:
  What new complexity characteristic to be introduced?
- There is a little sense to estimate the shatter profile D_m.
- **Open problem 2 (towards tighter bounds):**
  How to take into account both *splitting* and *similarity*?

Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, no. 2. — Pp. 243–259.

A toy example: the pair of classifiers

Consider two classifiers a_1, a_2 with m_1, m_2 errors on X^L:

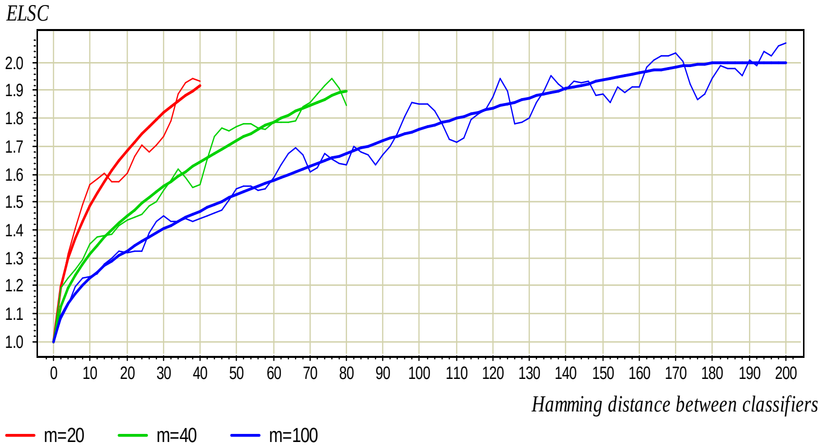
$$\vec{a}_1(X^L) = (\overbrace{1\,1\,1\;1\,1\,1\,1\,1}^{m_1}\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0);$$
$$\vec{a}_2(X^L) = (0\,0\,0\,\underbrace{\overbrace{1\,1\,1\,1\,1}^{m_0}\;1\,1\,1\,1\,1}_{m_2}\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0).$$

Theorem (exact probability of overfitting)

$$\mathsf{P}_n\{\delta(\mu, X_n^\ell, X_n^k) \geqslant \varepsilon\} = \sum_{s_0=0}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s_2=0}^{m_2} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{L-m_0-m_1-m_2}^{\ell-s_0-s_1-s_2}}{C_L^\ell} \times$$
$$\times \bigl[m_0 + m_1 + m_2 - k \leqslant s_0 + s_1 + s_2 \leqslant \ell\bigr] \times$$
$$\times \Bigl([s_1 < s_2]\, \bigl[s_0 + s_1 \leqslant \tfrac{\ell}{L}(m_0 + m_1 - \varepsilon k)\bigr] +$$
$$+ [s_1 \geqslant s_2]\, \bigl[s_0 + s_2 \leqslant \tfrac{\ell}{L}(m_0 + m_2 - \varepsilon k)\bigr] \Bigr).$$

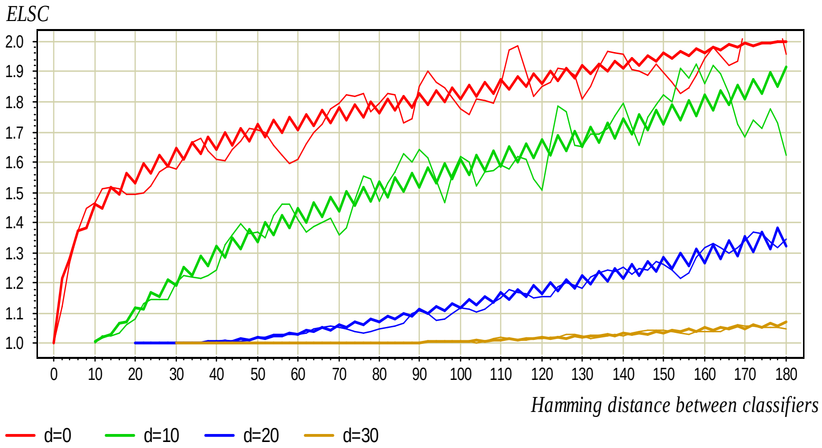
Experiment 1. Two classifiers of the equal quality

$\ell = k = 100;\ \ \varepsilon = 0.05;\ \ \underline{m_1 = m_2};\ \ m = 20,\ 40,\ 100$



Experiment 2. Two classifiers of the different quality

$$\ell = k = 100; \ \ \varepsilon = 0.05; \ \ \underline{m_0 = 20}; \ \ \underline{d \equiv m_2 - m_1} = 0,\ 10,\ 20,\ 30$$



Experiment 3. The chain of 1000 classifiers

$D = 1000$ classifiers, given by their binary loss vectors;
$\ell = k = 100$ — the size of training and testing sets ($L = 200$);
$m/L = 0.05, 0.25$ — the quality of the best classifier;
$\varepsilon = 0.05$ — the threshold of overfitting;
$N' = 1000$ random Monte-Carlo generated partitions.

A binary $L \times D$-matrix of column vectors of losses:

Example:

$$
\begin{array}{ccccccccccccccccc}
1 & 1 & \to & 0 & 0 & 0 & \to & 1 & 1 & 1 & 1 & 1 & 1 & & & & \dots \\
0 & 0 & & 0 & 0 & \to & 1 & & 1 & 1 & 1 & 1 & 1 & \to & 0 & & \dots
\end{array}
$$

$$
\begin{array}{ccccccccccccl}
1 & 1 \to 0 & 0 & 0 \to 1 & 1 & 1 & 1 & 1 & 1 & & \dots \\
0 & 0 & 0 & 0 \to 1 & 1 & 1 & 1 & 1 & 1 \to 0 & & \dots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
0 & 0 & 0 \to 1 & 1 & 1 & 1 & 1 \to 0 & 0 & 0 & 0 & \dots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \to 1 & 1 & 1 & 1 & \dots \\
0 \to 1 & 1 & 1 & 1 & 1 \to 0 & 0 & 0 \to 1 & 1 & & & \dots
\end{array}
$$

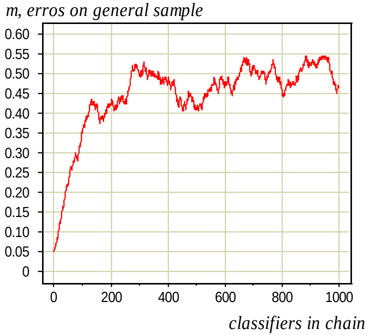
Chain is a sequence of binary loss vectors such that each subsequent vector differ from the previous one in one bit.

Chains with and without splitting

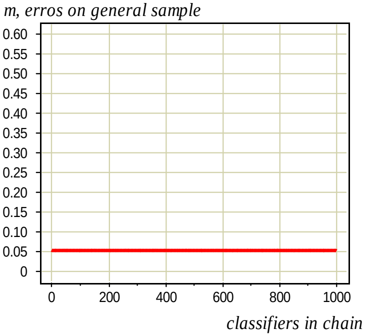
Chain is a sequence of binary loss vectors such that each subsequent vector differ from the previous one in one bit.

Two extreme types of chains:

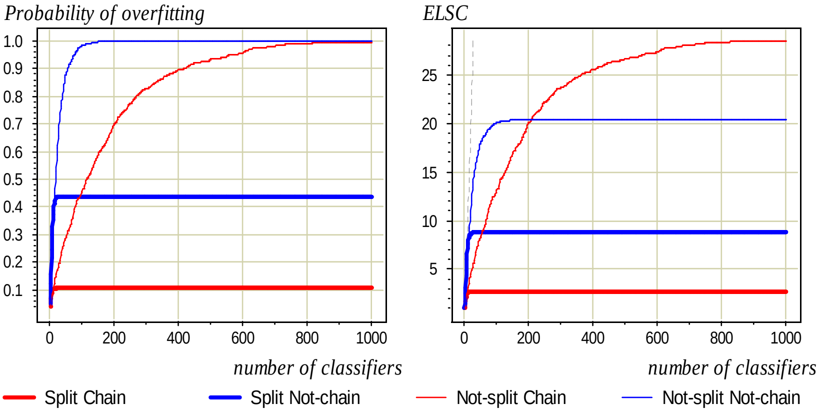
(1) split chain



(2) not-split chain

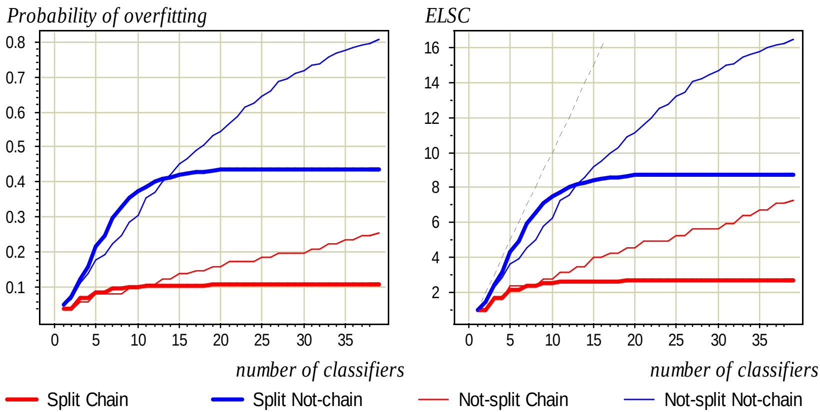


Chain and not-chain, with and without splitting ($m/L = 0.05$)



In the case of splitting and low errors ($m/L = 0.05$), the probability of overfitting newer reaches 1 with $D \to \infty$.

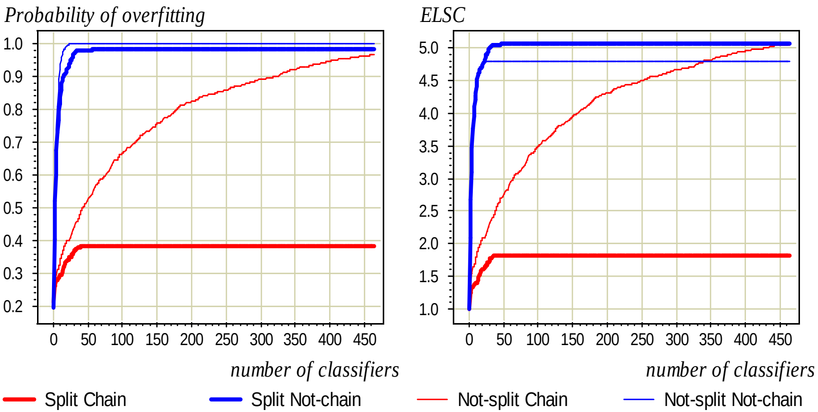
Chain and not chain, with and without splitting ($m/L = 0.05$, zoom)



According to VC theory $\hat{\Delta}(D) = D$.

This happens for not-chains and for low D only.

Chain or not chain, with or without splitting ($m/L = 0.25$)



In high error situation ($m/L = 0.25$)
only a *chain with splitting* provides a low overfitting.

Conclusions

- ELSC $\hat{\Delta}(D)$ has a horizontal asymptote whereas $\hat{\Delta} = D$ according to VC theory.
- *Chain* provides a slower growth of $\Delta(D)$.
- *Splitting* provides a lower ($\ll 1$) horizontal asymptote.
- **About the nature of overfitting:** overfitting arises as a result of choice of the best classifier (even from 2 classifiers!) on a finite training set.
- **Optimism:** *Chains with splitting* are very often encountered in practice; just in this situation overfitting is low.
- **Motivation for further research:** No theory exists that can take into account both similarity (chain) and splitting.

Thanks!
Questions?

Two invitations

1. Continually:
   virtual seminars on wiki www.MachineLearning.ru (in Russian)
   - Слабая вероятностная аксиоматика
     — Weak Probability Axiomatic
   - Расслоение и сходство алгоритмов (виртуальный семинар)
     — Splitting and similarity of classifiers (virtual seminar)
   - Участник:Vokov
     — K.Vorontsov's participant page
2. Today! 15:00, Assembly Hall, building 2
   The wiki resource www.MachineLearning.ru for research and education collaboration
   (plenary talk and discussion)