Recent Advances on Generalization Bounds Part I

Konstantin Vorontsov

Computing Center RAS • Moscow Institute of Physics and Technology

4th International Conference on
Pattern Recognition and Machine Intelligence (PReMI'11)
Moscow, Russian Federation • June 27 – July 1, 2011

Structure

- **Tutorial Part I. Overview of Generalization Bounds**
  June 27, 13:40–14:40 PReMI tutorial 4 (Γ-313)
  - VC, Occam Razor, Rademacher, and margin-based bounds
  - How these bounds can be used for learning algorithm design?
- **Tutorial Part II. Combinatorial Generalization Bounds**
  June 27, 15:00–16:00 PReMI tutorial 4 (continued) (Γ-313)
  - Why complexity bounds are so loose (overestimated)?
  - How to obtain tight or even exact bounds?
  - Will they be useful?
- **Part III. Tight Combinatorial Generalization Bounds for Threshold Conjunction Rules**
  June 29, 11:00-11:20 PReMI session 4 (Γ-408)
  - A practical issue from Combinatorial Generalization Bounds

Contents of the Part I

Classification problem

X — a set of *objects*, usually $\mathbb{R}^n$
Y — a set of *class* labels, usually $\{-1, +1\}$ or $\{1, \ldots, M\}$
$y: X \to Y$ — unknown *target* function
$X^\ell = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$ — *training set*, $y_i = y(x_i)$, $i = 1, \ldots, \ell$

Classification is a supervised learning problem:

find a classifier $a: X \to Y$ from a given function set A
that *generalizes well*, that is approximates well a target y
not only on the training set X^ℓ but everywhere on X.

One must specify accurately:

- what means "approximate well"?
- what means "approximate everywhere on X"?

Probabilistic model of data

Let $X \times Y$ be a probability space with unknown distribution $p(x, y)$ and observations $(x_i, y_i)_{i=1}^{\ell}$ be drawn independently from p

Define a *binary loss function* $I(a, x)$, usually

$$I(a, x) = \big[a(x) \neq y(x)\big] = \begin{cases} 1, & a(x) \neq y(x) \\ 0, & a(x) = y(x) \end{cases}$$

Empirical error (error rate, frequency of errors) of a classifier a

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

Probability of error (generalization error) of a classifier a

$$P(a) = \mathsf{P}_x\big(I(a, x) = 1\big)$$

Empirical risk minimization

Learning algorithm μ is a function that takes a training sample X^ℓ and gives a classifier a^* from A:

$$a^* = \mu(X^\ell)$$

Empirical risk minimization (ERM) is a classical example of the learning algorithm:

$$a^* = \arg\min_{a \in A} \nu(a, X^\ell)$$

Unfortunately, ERM can lead to *overfitting*, when

$$\nu(a^*, X^\ell) \ll P(a^*)$$

How generalization bounds can help to reduce overfitting

There are two things to do:

1. to give an upper bound of the probability of error that holds for any A, any p, any μ (and sometimes any X^ℓ):

$$P(a^*) \leqslant \hat{P}(a^*, X^\ell)$$

Two types of bounds exist:
$\hat{P}(a)$ — data-independent bound (usually very loose)
$\hat{P}(a, X^\ell)$ — data-dependent bound (most recent bounds)

2. to construct the *learning algorithm* μ that minimizes the generalization error bound:

$$a^* = \mu(X^\ell) \equiv \arg\min_{a \in A} \hat{P}(a, X^\ell)$$

Probably Approximately Correct (PAC) learning

Problem statement [Vapnik & Chervonenkis, 1969; Valiant, 1984]: Given only A and ℓ, find a bound $\eta(\varepsilon, \ell, A)$ on the *uniform deviation* of the error frequency from the error probability:

$$\mathsf{P}_{X^\ell}\Big(P(a^*) - \nu(a^*, X^\ell) \geqslant \varepsilon\Big) \leqslant$$

$$\mathsf{P}_{X^\ell}\Big(\sup_{a \in A}\big(P(a) - \nu(a, X^\ell)\big) \geqslant \varepsilon\Big) \leqslant \eta(\varepsilon, A, \ell) \quad - \ ?$$

Plus: it holds for any learning algorithm μ and any distribution p

Minus: it is a worst-case bound which can be very loose

The uniform convergence principle is an axiom in VC-theory, PAC-learning theory, and Rademacher Complexity theory

The inversion technique

If the bound has been obtained

$$\mathsf{P}_{X^\ell}\left(\sup_{a\in A}\big(P(a) - \nu(a, X^\ell)\big) \geqslant \varepsilon \right) \leqslant \eta(\varepsilon, A, \ell)$$

then, with probability at least $1 - \eta$ for any classifier $a \in A$

$$P(a) \leqslant \nu(a, X^\ell) + \varepsilon(\eta, A, \ell),$$

where $\varepsilon(\eta, A, \ell)$ is the *inverse function* for $\eta(\varepsilon, A, \ell)$.

A new learning algorithm: try to minimize generalization error $P(a)$

$$a^* = \arg\min_{a} \min_{A} \min_{a\in A} \Big(\nu(a, X^\ell) + \varepsilon(\eta, A, \ell) \Big)$$

(differs from ERM by penalty $\varepsilon(\eta, A, \ell)$ and extra optimization by A)

The binomial tail bound for one-classifier case

The empirical error of a fixed classifier a is distributed binomially:

$$\mathsf{P}\big(\nu(a, X^\ell) = \tfrac{s}{\ell}\big) = C_\ell^s p^s (1-p)^{\ell-s}, \quad \text{where } p = P(a)$$

Binomial tail — exact bound, tedious inversion:

$$\mathsf{P}\big(P(a) - \nu(a, X^\ell) \geqslant \varepsilon\big) = \sum_{s=0}^{\ell p - \ell\varepsilon} C_\ell^s p^s (1-p)^{\ell-s}$$

Chernoff's inequality — inflated bound, easier inversion:

$$\leqslant \exp\big(-\ell\, \mathsf{KL}(p-\varepsilon \| p)\big)$$

where $\mathsf{KL}(q\|p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ is Kullback–Leibler divergence.

Hoeffding's inequality — more inflated bound, trivial inversion:

$$\leqslant \exp\big(-2\ell\varepsilon^2\big)$$

Vapnik-Chervonenkis bound for finite set

Theorem (data-independent bound)

If A is a finite set of classifiers then for any $\varepsilon \in (0,1)$

$$\mathsf{P}\Big(\sup_{a \in A} \big(P(a) - \nu(a, X^\ell)\big) \geqslant \varepsilon \Big) \leqslant |A| \cdot \exp\big(-2\ell\varepsilon^2\big)$$

Proof sketch:

first, apply the *union bound*:

$$\mathsf{P}\Big(\sup_{a \in A} \big(P(a) - \nu(a, X^\ell)\big) \geqslant \varepsilon \Big) \leqslant \sum_{a \in A} \mathsf{P}\Big(P(a) - \nu(a, X^\ell) \geqslant \varepsilon \Big)$$

second, apply the one-classifier bound from the previous slide:

$$\leqslant |A| \cdot \exp\big(-2\ell\varepsilon^2\big)$$

Vapnik-Chervonenkis bound for infinite set

Theorem (data-independent bound)

If A is an arbitrary set of classifiers then for any $\varepsilon \in (0,1)$

$$\mathsf{P}\Big(\sup_{a \in A} \big(P(a) - \nu(a, X^\ell)\big) \geqslant \varepsilon \Big) \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2} \exp\big(-\ell\varepsilon^2\big)$$

where $\Delta^A(L)$ is the growth function of the set A.

Definition. The *growth function* $\Delta^A(L)$ of the set A is the maximal number of distinct L-dimensional binary vectors $\mathbf{a} = \big(I(a, x_1), \ldots, I(a, x_L)\big)$ induced by all classifiers $a \in A$ on a sample $X^L = (x_1, \ldots, x_L)$

Informally, $\Delta^A(L)$ is a *complexity measure* of the set A.

Vapnik-Chervonenkis dimension

Definition: The VC-dimension of the set A is the maximal sample size h such that $\Delta^A(h) = 2^h$.

Theorem

If such h exists then $\Delta^A(L) \leqslant C_L^0 + \cdots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}$

Consider a two-class classification problem $Y = \{-1, +1\}$ and a set A of linear classifiers in n-dimensional object space $X = \mathbb{R}^n$:

$$a(x) = \text{sign}(w_1 x^1 + \cdots + w_n x^n), \quad x = (x^1, \ldots, x^n) \in X.$$

Theorem

$\text{VCdim}(A) = n$

The inversion technique

VC-bound $\mathsf{P}\Big(\sup\limits_{a\in A}\big(P(a)-\nu(a,X^\ell)\big)\geqslant\varepsilon\Big)\leqslant\Delta\cdot\exp\big(-\ell\varepsilon^2\big)$

gives with probability at least $1-\eta$

$$P(a)\leqslant\underbrace{\nu(a,X^\ell)}_{\substack{\text{empirical}\\\text{risk}}}+\underbrace{\sqrt{\frac{1}{\ell}\ln\Delta+\frac{1}{\ell}\ln\frac{1}{\eta}}}_{\text{complexity penalty}}$$

VC-bound $\mathsf{P}\Big(\sup\limits_{a\in A}\big(P(a)-\nu(a,X^\ell)\big)\geqslant\varepsilon\Big)\leqslant\frac{3}{2}\frac{L^h}{h!}\cdot\frac{3}{2}\exp\big(-\ell\varepsilon^2\big)$

gives with probability at least $1-\eta$

$$P(a)\leqslant\underbrace{\nu(a,X^\ell)}_{\substack{\text{empirical}\\\text{risk}}}+\underbrace{\sqrt{\frac{h}{\ell}\ln\left(\frac{2e\ell}{h}\right)+\frac{1}{\ell}\ln\frac{4}{9\eta}}}_{\text{complexity penalty}}$$

Structural Risk Minimization (SRM)

Given a system of nested subsets of increasing dimensions

$$A_0 \subset A_1 \subset \cdots \subset A_h \subset \cdots$$

Find an optimal dimension h^*:

$$P(a) \leqslant \underbrace{\min_{a \in A_h} \nu(a, X^\ell)}_{\substack{\text{empirical} \\ \text{risk minimization}}} + \underbrace{\sqrt{\frac{h}{\ell} \ln\left(\frac{2e\ell}{h}\right) + \frac{1}{\ell} \ln \frac{4}{9\eta}}}_{\text{complexity penalty}} \to \min_h$$

The main disadvantage of SRM approach:
VC-bound is very loose (overestimated)
Then, h^* may be suboptimal (oversimplified)

Practitioners prefer to use Cross-Validation instead of the bound

Two main reasons of the VC-bound looseness

- **The uniform deviation bound** is highly overestimated when most classifiers have a vanishing probability to be obtained by the learning algorithm.

  In practice, the distribution over classifiers

  $$q(a) = \mathsf{P}\big(\mu(X^\ell) = a\big), \quad a \in A$$

  is essentially nonuniform!
- **The union bound** is highly overestimated when there are a lot of similar classifiers.

  In practice, this is usually the case!

Let us start with the first problem...

Occam Razor bound

One can not know $q(a) = \mathsf{P}\big(\mu(X^\ell) = a\big)$, but one can make a shot.
Let $p(a)$ be a normalized function — "prior" distribution over A.

Theorem (Occam Razor bound)

For any "prior" $p(a)$ over A, for any $\eta \in (0,1)$, for all $a \in A$

$$P(a) \leqslant \nu(a, X^\ell) + \sqrt{\frac{1}{\ell} \ln \frac{1}{p(a)} + \frac{1}{\ell} \ln \frac{1}{\eta}}.$$

with probability at least $1 - \eta$.

Statement 1. If one guess well, $p(a) = q(a)$, then this bound is most tight.

Statement 2. Although it is still very overestimated because only first of two reasons of looseness has been treated...

Occam Razor bound: how to specify $p(a)$?

Example 1.
The *uniform prior* $p(a) = \frac{1}{|A|}$ gives the VC-bound:

$$P(a) \leqslant \nu(a, X^\ell) + \sqrt{\frac{1}{\ell} \ln |A| + \frac{1}{\ell} \ln \frac{1}{\eta}}.$$

with probability at least $1 - \eta$. Nothing new...

Occam Razor bound: how to specify $p(a)$?

Example 2.
Consider a two-class classification problem $Y = \{-1, +1\}$ and a set A of linear classifiers in n-dimensional object space $X = \mathbb{R}^n$:

$$a(x) = \text{sign}(w_1 x^1 + \cdots + w_n x^n), \quad x = (x^1, \ldots, x^n) \in X.$$

The *Gaussian prior*: weights $w \in \mathbb{R}^n$ are independent, with zero expectation, and equal variance σ^2:

$$p(a) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\|w\|^2\right)$$

Substituting this prior into Occam Razor bound gives

$$P(a) \leqslant \nu(a, X^\ell) + \sqrt{\frac{n}{\ell} \ln \sigma\sqrt{2\pi} + \frac{\|w\|^2}{2\ell\sigma^2} + \frac{1}{\ell} \ln \frac{1}{\eta}}.$$

Regularization

The minimization of the obtained bound

$$P(a) \leqslant \nu(a, X^\ell) + \sqrt{\frac{n}{\ell} \ln \sigma\sqrt{2\pi} + \frac{\|w\|^2}{2\ell\sigma^2} + \frac{1}{\ell} \ln \frac{1}{\eta}} \;\to\; \min_w$$

can be considered as a nontrivial mixture of L_0- and L_2-regularization:

$$\nu(a, X^\ell) \;\to\; \min_w \qquad \text{ERM}$$

$$\nu(a, X^\ell) + C_0 n \;\to\; \min_w \qquad L_0\text{-regularization}$$

$$\nu(a, X^\ell) + C_1 \sum_{j=1}^{n} |w_j| \;\to\; \min_w \qquad L_1\text{-regularization}$$

$$\nu(a, X^\ell) + C_2 \sum_{j=1}^{n} w_j^2 \;\to\; \min_w \qquad L_2\text{-regularization}$$

Bound minimization leads to shrinkage and features selection.

Two main reasons of the VC-bound looseness (revisited)

- **The uniform deviation bound** is loose
  when $q(a)$ is essentially nonuniform distribution.
- **The union bound** is loose
  when there are a lot of similar classifiers.

Occam Razor bound treats only the first problem.
Its main difficulty is to guess the prior $p(a)$ well.

Below we consider approaches which treat the second problem...

Further readings on Occam Razor bounds:

[1] *Langford J.* Quantitatively Tight Sample Complexity Bounds: Ph.D. thesis. Carnegie Mellon Thesis. 2002.

[2] *Langford J.* Tutorial on practical prediction theory for classification. Journal of Machine Learning Research. 2005. Vol. 6. Pp. 273–306.

The notion of Rademacher Complexity

$\mathscr{L}: A \times X \to [-1, +1]$ — the real-valued bounded loss function
$\mathscr{L}(a, x)$ — the loss of a classifier a at the object x
$\mathbf{a} = \big(\mathscr{L}(a, x_1), \ldots, \mathscr{L}(a, x_\ell)\big)$ — *loss vector* of a classifier a.

Definition 1. *Local Rademacher complexity* of the set A on X^ℓ

$$\mathcal{R}(A, X^\ell) = \mathsf{E}_\sigma \sup_{a \in A} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} \sigma_i \mathscr{L}(a, x_i) \right|,$$

where $\sigma_1, \ldots, \sigma_\ell$ are independent Rademacher random variables, i. e. $\mathsf{P}(\sigma_i = -1) = \mathsf{P}(\sigma_i = +1) = \frac{1}{2}$.

Interpretation: If for any noise vector $(\sigma_1, \ldots, \sigma_\ell)$ one can find in A a highly covariated loss vector, then the set A is complex.

Definition 2. *Rademacher complexity* of the set A:

$$\mathcal{R}(A) = \mathsf{E}_X \mathcal{R}(A, X^\ell)$$

Generalization bound via Rademacher Complexity

$\tilde{P}(a) = \mathsf{E}\mathscr{L}(a, x)$ — expected loss

$\tilde{\nu}(a, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathscr{L}(a, x_i)$ — empirical loss

Theorem

With probability at least $1 - \eta$ *for all* $a \in A$

$$\tilde{P}(a) \leqslant \tilde{\nu}(a, X) + 2\mathcal{R}(A) + \sqrt{\frac{1}{2\ell}\ln\frac{2}{\eta}}$$

$$\leqslant \tilde{\nu}(a, X) + 2\mathcal{R}(A, X) + 3\sqrt{\frac{1}{2\ell}\ln\frac{2}{\eta}}$$

Most important properties of Rademacher Complexity

1. Relationship with the growth function:
$$\mathcal{R}(A) \leqslant \sqrt{\tfrac{2}{\ell} \ln \Delta^A(\ell)}$$
2. For any sets of classifiers A, B and any constant $c \in \mathbb{R}$

$$\mathcal{R}(A \cup B) \leqslant \mathcal{R}(A) + \mathcal{R}(B);$$
$$\mathcal{R}(c \cdot A) = |c| \cdot \mathcal{R}(A), \qquad c \cdot A = \{c\mathbf{a} \colon a \in A\};$$
$$\mathcal{R}(A \oplus B) \leqslant \mathcal{R}(A) + \mathcal{R}(B), \quad A \oplus B = \{\mathbf{a} + \mathbf{b} \colon a \in A,\ b \in B\};$$

3. The convex hull of the set of loss vectors A has the same Rademacher complexity as A:
$$\mathcal{R}\Big\{\sum_{a \in A} c_a \mathbf{a} \colon \sum_{a \in A} |c_a| \leqslant 1\Big\} = \mathcal{R}(A).$$

Properties of Rademacher Complexity

Rademacher Complexity being defined via covariance has a lot of convenient algebraical properties.

Due to this fact Rademacher Complexity can be estimated for nontrivial and practically useful sets of classifiers.

Below we consider two of them: kernel machines and boosting.

Further readings on Rademacher Complexity

[1] *Bartlett P. L., Mendelson S.* Rademacher and Gaussian complexities: risk bounds and structural results. JMLR. 2002 No 3. Pp. 463–482.

[2] *Bartlett P., Bousquet O., Mendelson S.* Local rademacher complexities. Vol. 33. Institute of Mathematical Statistics, 2005. Pp. 1497–1537.

[3] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances. ESAIM: Probability and Statistics. 2005. No. 9. Pp. 323–375.

Continuous approximations of threshold loss function

Consider a two-class classification problem $Y = \{-1, +1\}$ and a set A of classifiers $a(x, w) = \operatorname{sign} f(x, w)$.

For linear classifier in n-dimensional object space $X = \mathbb{R}^n$:

$$f(x, w) = w_1 x^1 + \cdots + w_n x^n = \langle x_i, w \rangle.$$

Definition. *Margin* of the object x_i with a class label $y_i \in \{-1, 1\}$

$$M_i(w) = y_i f(x_i, w)$$

$M_i(w) < 0 \iff$ classifier $a(x, w)$ makes an error on x_i.

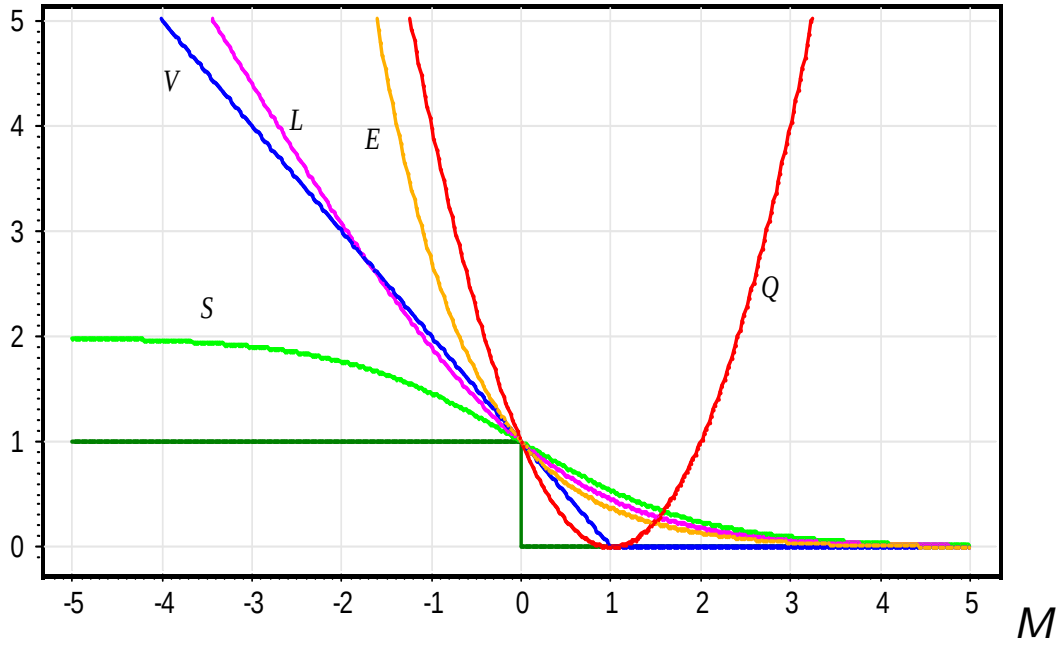
From Empirical Risk Minimization to *Approximated ERM*:

$$\nu(w, X^\ell) = \sum_{i=1}^{\ell} \big[M_i(w) < 0\big] \quad \leqslant \quad \tilde{\nu}(w, X^\ell) = \sum_{i=1}^{\ell} \mathscr{L}\big(M_i(w)\big) \to \min_w;$$

loss function $\mathscr{L}(M)$ is continuous, nonincreasing, nonnegative.

Continuous approximations of threshold loss function

Frequently used loss functions $\mathscr{L}(M)$:



$Q(M) = (1 - M)^2$ — Fisher's Discriminant;
$V(M) = (1 - M)_+$ — SVM;
$S(M) = 2(1 + e^M)^{-1}$ — sigmoidal ANN;
$L(M) = \log_2(1 + e^{-M})$ — Logistic Regression;
$E(M) = e^{-M}$ — AdaBoost.

Approximation and Regularization of the Empirical Risk

Many practical learning algorithms are based on both Approximation and Regularization of the Empirical Risk, e. g.

$$\tilde{\nu}(w, X^\ell) + C\|w\|^2 \to \min_w;$$

This can be justified from generalization bounds.

Theorem

Let A be a set of linear classifiers, loss function is bounded $[M<0] \leqslant \mathscr{L}(M) \leqslant \mathscr{L}_{\max}$ and has a Lipschitz constant λ. Then with probability at least $1-\eta$ for all $a \in A$

$$P(a) \leqslant \tilde{\nu}(w, X^\ell) + 2\lambda\mathcal{R}(A, X^\ell) + \mathscr{L}_{\max}\sqrt{\frac{2}{\ell}\ln\frac{1}{\eta}}\,.$$

Rademacher Complexity bound for Kernel Machines

Consider a kernel based linear classifier

$$a(x, w) = \operatorname{sign}\left(\sum_{i=1}^{\ell} w_i K(x_i, x) - w_0\right),$$

Theorem

If w is bounded in a sense of the norm
$\|w\|_K^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} w_i w_j K(x_i, x_j) \leqslant B^2$, *then*

$$\mathcal{R}(A, X^\ell) \leqslant \frac{2B}{\ell}\sqrt{\sum_{i=1}^{\ell} K(x_i, x_i)}\ .$$

Interpretation: $\|w\|_K \leqslant B$ is a data-dependent regularization.
Learning algorithm: minimize $\|w\|_K$ until $\tilde{\nu}(a, X^\ell)$ grow.

Rademacher Complexity bound for Kernel Machines

Learning algorithm: minimize $\|w\|_K$ until $\tilde{\nu}(a, X^\ell)$ grow.

Learning algorithm as an optimization problem:
data-dependent and kernel-dependent regularization:

$$\underbrace{\sum_{i=1}^{\ell} \mathscr{L}\big(M_i(w)\big)}_{\text{approximated ERM}} + \underbrace{\frac{2}{\ell}\sqrt{\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} w_i w_j K(x_i, x_j)}\sqrt{\sum_{i=1}^{\ell} K(x_i, x_i)}}_{\text{regularization term (complexity penalty)}} \to \min_w$$

Note: regularization term may be nontrivial...
in contrast with usual $\|w\|^2$ or L_p-norms

Weighted voting of classifiers

Weighted voting of classifiers (*boosting*, *bagging*, etc.):

$$a(x) = \operatorname{sign} \sum_{t=1}^{T} w_t b_t(x), \quad w_t \geqslant 0,$$

where $b_t(x)$ are *base classifiers* of VC-dimension h.

$b_t(x)$ can be learned independently (bagging) or subsequently (boosting). It is no matter for generalization!

From the property $\mathcal{R}(\text{conv}A) = \mathcal{R}(A)$ one obtain

$$\mathcal{R}(\text{conv}A) \leqslant \sqrt{\frac{2h}{\ell} \ln \frac{\ell e}{h}},$$

where h is VC-dimension of the set of base classifiers.

Rademacher Complexity bound for weighted voting

Theorem

For any a with probability at least $1 - \eta$

$$P(a) \leqslant \sum_{i=1}^{\ell} \mathscr{L}\big(M_i(w)\big) + 2\lambda\sqrt{\frac{2h}{\ell}\ln\frac{\ell e}{h}} + \mathscr{L}_{\max}\sqrt{\frac{2}{\ell}\ln\frac{1}{\eta}}.$$

Conclusions for weighted voting learning algorithms:

- a great variety of loss functions $\mathscr{L}(M)$ can be used;
- generalization of weighted voting does not depend on T;
- boosting maximizes margins M_i effectively, then minimizing the first term of the bound;
- one can use very simple base classifiers

Further reading

Further readings on margin-based generalization bounds

[1] *Koltchinskii V., Panchenko D.* Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics. 2002. Vol. 30, No. 1. Pp 1–50.

[2] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances. ESAIM: Probability and Statistics. 2005. No. 9. Pp. 323–375.

Conclusion (Part I)

- Generalization bounds give optimization problems to construct learning algorithms with better performance.
- Typically, this is ERM Approximation + Regularization.
- The better performance is not always successfully attained because of the looseness of the bounds.
- There are two reasons for the looseness:
  — nonuniform splitting of the set of classifiers;
  — similarity of classifiers.
- None of recent generalization bounds can treat both problems.

To be continued in 20 minutes...

We will consider a *combinatorial approach*, the first in Learning Theory that takes into account both splitting and similarity of a classifier set and sometimes gives exact generalization bounds.

Questions?

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

www.MachineLearning.ru/wiki (in Russian):

- Участник:Vokov
- Слабая вероятностная аксиоматика
- Расслоение и сходство алгоритмов (виртуальный семинар)