Large Margin Classification

John Shawe-Taylor
Department of Computer Science
Royal Holloway and Bedford New College
University of London
john@dcs.rhbnc.ac.uk

Bob Williamson
Department of Engineering
Australian National University
Bob.Williamson@anu.edu.au

July 6, 1999

STRUCTURE

1. Basic PAC Ideas
2. Basic Margin Ideas
3. Their Exploitation
4. And Extension
5. Conclusions

Aim:

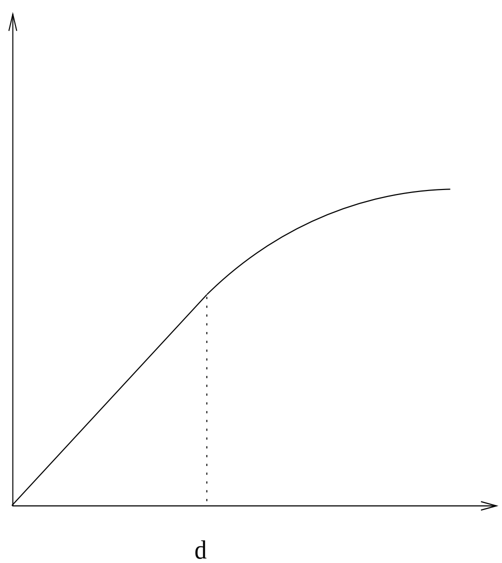
- Basic Techniques
- Overview of (some of) state of the art
- Where it fits in the grander scheme

We won't be covering

- Detailed proofs
- The most general results
- History
- Algorithms
- Other views of margins (e.g. Statistical Physics)

PAC BOUNDS AS A STARTING POINT

- Let H be a set of $\{-1, 1\}$ valued functions.
- The growth function $B_H(m)$ is the maximum cardinality of the set of functions H when restricted to m points.
- Consider a plot of the log of the growth function $\log_2(B_H(m))$ as a function of m:



Vapnik Chervonenkis dimension

- The Vapnik-Chervonenkis dimension is the point at which the graph stops being linear:

$$\text{VCdim}(H) = \max\{m \quad : \quad \text{for some } x^1, \ldots, x^m, \text{ for all } b \in \{-1, 1\}^m, \exists h_b \in H, h_b(x^i) = b_i\}$$

- For linear functions $\mathcal{L}$ in $\mathbb{R}^n$, $\text{VCdim}(\mathcal{L}) = n + 1$.

- Sauer's Lemma:

$$B_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d,$$

where $m \geq d = \text{VCdim}(H)$.

Basic Statistical Result

- We want to bound the probability that the training examples can mislead us about one of the functions we are considering using:

$$P^m\{\mathbf{X} \in X^m : \exists h \in H : \text{err}_{\mathbf{X}}(h) = 0, \text{err}_P(h) \geq \epsilon\}$$
$$\leq 2P^{2m}\{\mathbf{XY} \in X^{2m} : \exists h \in H \text{ such that } \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2\}$$
$$\leq 2B_H(2m)P^{2m}\{\mathbf{XY} \in X^{2m} : \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2\}$$
$$\leq 2B_H(2m)2^{-\epsilon m/2} \leq \delta$$

- inverting gives

$$\epsilon = \epsilon(m, H, \delta) = \frac{2}{m}\left(d \log \frac{2m}{d} + \log \frac{2}{\delta}\right)$$

i.e. with probability $1 - \delta$ over m random examples a consistent hypothesis has error less than ϵ.

Lower bounds

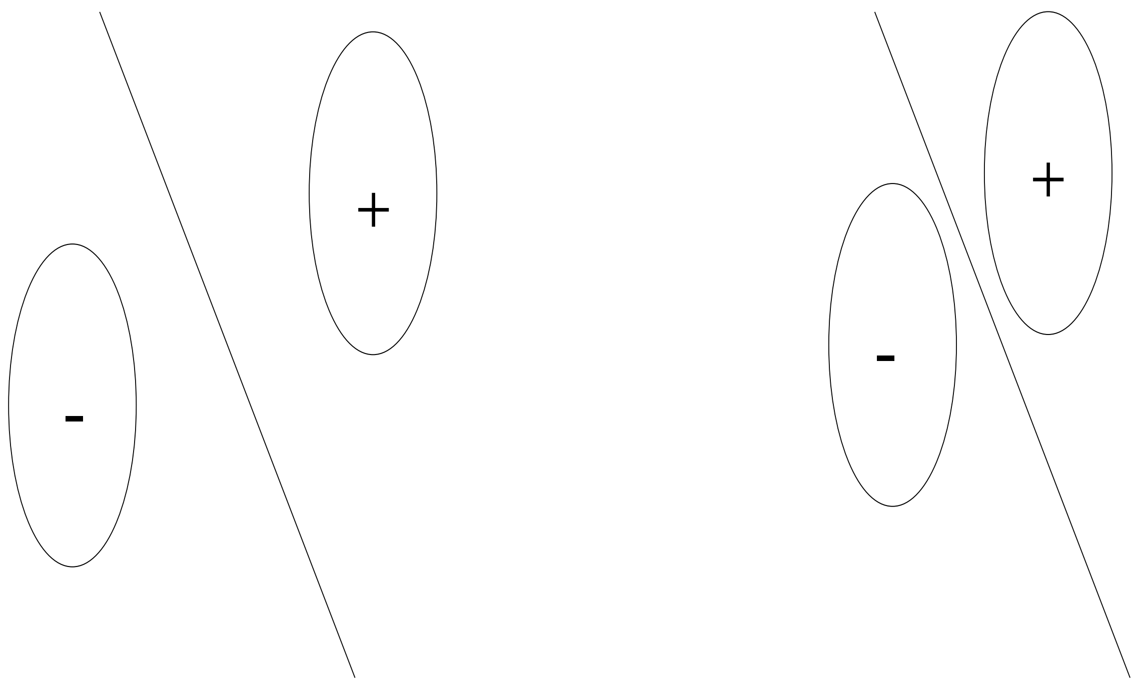
- VCdim *Characterises* Learnability in PAC setting: there exist distributions such that with probability at least δ over m random examples, the error of h is at least

$$\max\left(\frac{d-1}{32m}, \frac{1}{m}\log\left(\frac{1}{\delta}\right)\right).$$

Criticisms of PAC Theory

Numbers are crazy

More importantly does not give the right insight: says that learning to classify two gaussian clouds with difference in means 10, var 1 is just as hard as with diff of means 5 variance 1, whereas one's intuition suggests the former should be easier:



Hence, standard PAC does not always suggest new algorithms.

Support Vector Machines (SVM)

One example of PAC failure is in analysing SVMs: linear functions in very high dimensional feature spaces. Two key ingredients:

1. kernel trick to avoid explicit feature space map:

$$\Phi(\mathbf{x}) = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots)$$

   Kernel gives the inner product of two feature vectors (all we need for both learning algorithm and function evaluation) without computing them:

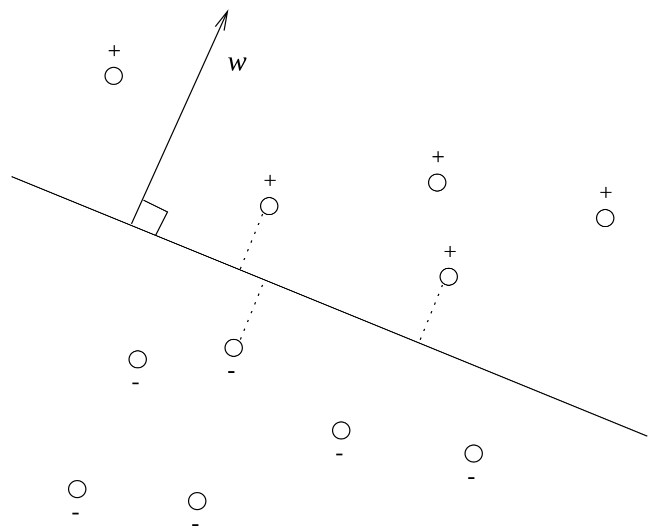
$$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \rangle$$

   e.g. Gaussian Kernel $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ corresponds to an infinite dimensional feature space so that PAC result does not apply: and YET very impressive performance, because ...

Margin in SVMs

2. Maximise the margin:

   – An example of maximising the margin in two dimensions:



   – Using linear functions (with unit weight vectors – space $\mathcal{L}$) to classify inputs from $\mathbb{R}^d$ into two classes:

$$h(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i=1}^{d} w_i x_i - b\right] = \operatorname{sgn}\left[f(\mathbf{x})\right],$$

given $f(\mathbf{x}) = \langle w \cdot \mathbf{x} \rangle - b$
Note that $|f(\mathbf{x})|$ is the distance from hyperplane.

Maximal Margin hyperplane

- Margin of a point $(\mathbf{x}, y)$ is $yf(\mathbf{x})$. Positive if correctly classified.

- Margin of f on training set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is

$$\gamma = m(f) = \min_i \{y_i f(\mathbf{x}_i)\}$$

positive if data correctly separated.

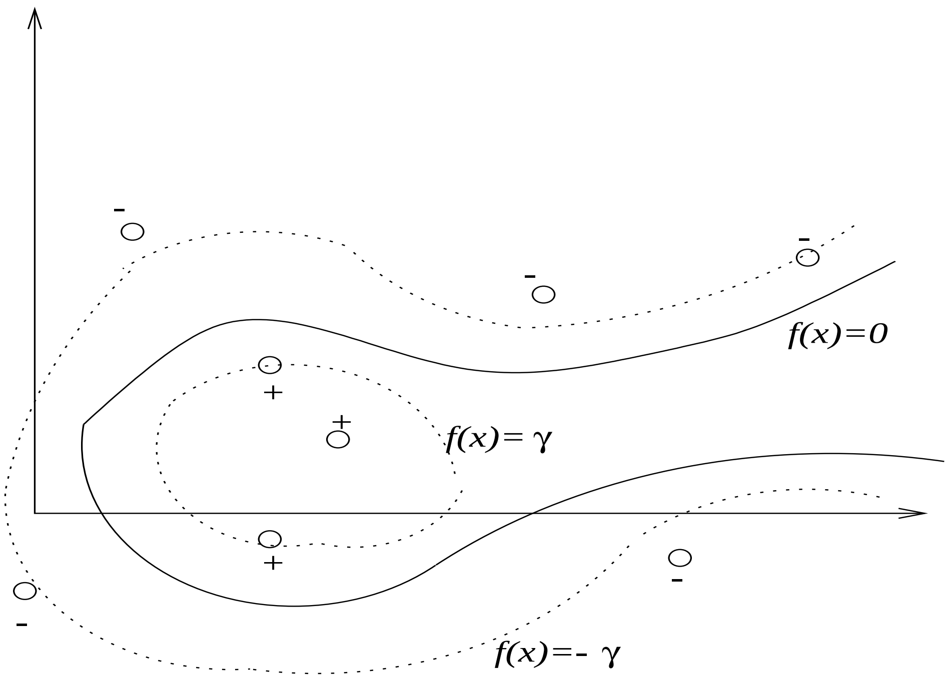
- We want a bound of the form $\epsilon = \epsilon(m, \mathfrak{L}, \delta, \gamma)$, i.e. with probability $1 - \delta$ over m random examples a margin γ hypothesis has error less than ϵ.

BASIC MARGIN IDEAS

1. The idea of the margin
2. Linear classifiers
3. General Real-valued classes (thresholded)
4. Boosting
5. Summary

The Idea of a margin

- General Margin: $yf(\mathbf{x})$:



- Intuitively having a margin gives immunity to noise

- Statistical Physics have analysed in terms of classification boundary

Distribution of margin values

- Given a training set $\mathbf{X}$ and f, we have a set of margin values:

$$M = \{y_i f(\mathbf{x}_i)\}$$

- Maximum margin algorithm maximises

$$\min M.$$

- Will look at other measures:
  - percentiles,
  - norm of the vector containing the amounts by which points fail to meet a target margin γ.

Some Definitions — Metric Spaces

The ℓ_p^d norms are:

For $0 < p < \infty$, $\|\mathbf{x}\|_{\ell_p^d} := \|\mathbf{x}\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$;

For $p = \infty$, $\|\mathbf{x}\|_{\ell_\infty^d} := \|\mathbf{x}\|_\infty = \max_{j=1,\ldots,d} |x_j|$.

(Note no normalization)

For $0 < p < \infty$, $\ell_p = \ell_p^\infty$.

More notation...

Given m points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \ell_p^d$, we use the shorthand

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m).$$

Suppose $\mathcal{F}$ is a class of functions $f : \mathbb{R}^d \to \mathbb{R}$.

The ℓ_∞^d norm *with respect to* $\mathbf{X}$ of $f \in \mathcal{F}$ is defined as

$$\|f\|_{\ell_\infty^{\mathbf{X}}} := \max_{i=1,\ldots,m} |f(\mathbf{x}_i)| = \|(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))\|_{\ell_\infty^{\mathbf{X}}}.$$

Likewise

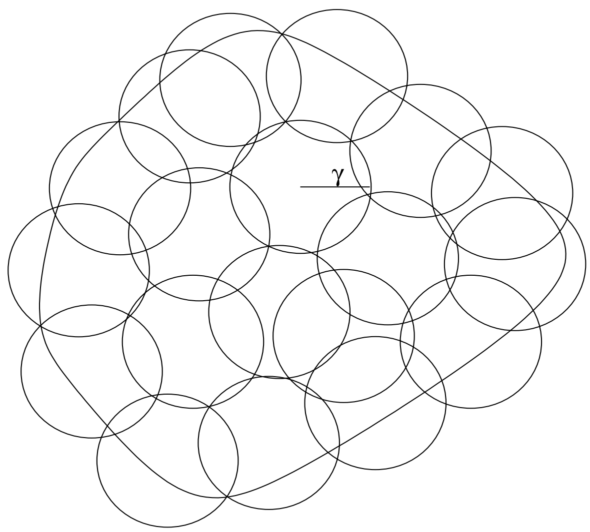
$$\|f\|_{\ell_p^{\mathbf{X}}} = \|(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))\|_{\ell_p^m}.$$

Covering Numbers

$\mathcal{F}$ a class of real functions defined on X and $\|\cdot\|_d$ a norm on $\mathcal{F}$, then

$$\mathcal{N}(\gamma, \mathcal{F}, \|\cdot\|_d)$$

is the smallest size set U_γ such that
for any $f \in \mathcal{F}$ there is a $u \in U_\gamma$ such that $\|f - u\|_d < \gamma$.



For generalization bounds we need the γ-*growth function*,

$$\mathcal{N}^m(\gamma, \mathcal{F}) := \sup_{\mathbf{X} \in X^m} \mathcal{N}(\gamma, \mathcal{F}, \ell_\infty^{\mathbf{X}}).$$

Second statistical result

- We want to bound the probability that the training examples can mislead us about one of the functions with margin bigger than fixed γ:

$$P^m\{\mathbf{X} \in X^m : \exists f \in \mathcal{F} : \text{err}_{\mathbf{X}}(f) = 0, m(f) \geq \gamma, \text{err}_P(f) \geq \epsilon\}$$
$$\leq 2P^{2m}\{\mathbf{XY} \in X^{2m} : \exists f \in \mathcal{F} \text{ such that}$$
$$\text{err}_{\mathbf{X}}(f) = 0, m(f) \geq \gamma, \text{err}_{\mathbf{Y}}(f) \geq \epsilon/2\}$$
$$\leq 2\mathcal{N}^{2m}(\gamma/2, \mathcal{F})P^{2m}\{\mathbf{XY} \in X^{2m} : \text{ for fixed } f'$$
$$\text{err}_{\mathbf{X}}(f') = 0, \text{err}_{\mathbf{Y}}(f') \geq \epsilon/2\}$$
$$\leq 2\mathcal{N}^{2m}(\gamma/2, \mathcal{F})2^{-\epsilon m/2} \leq \delta$$

- inverting gives
$$\epsilon = \epsilon(m, \mathcal{F}, \delta, \gamma) = \frac{2}{m}\left(\log_2 \mathcal{N}^{2m}(\gamma/2, \mathcal{F}) + \log_2 \frac{2}{\delta}\right)$$

  i.e. with probability $1 - \delta$ over m random examples a margin γ hypothesis has error less than ϵ. Must apply for finite set of γ ('do SRM over γ').

Bounding the covering numbers

Have the following correspondences with the standard VC case (easy slogans):

Growth function	–	γ-growth function
Vapnik Chervonenkis dim	–	Fat shattering dim
Sauer's Lemma	–	Alon *et al.*

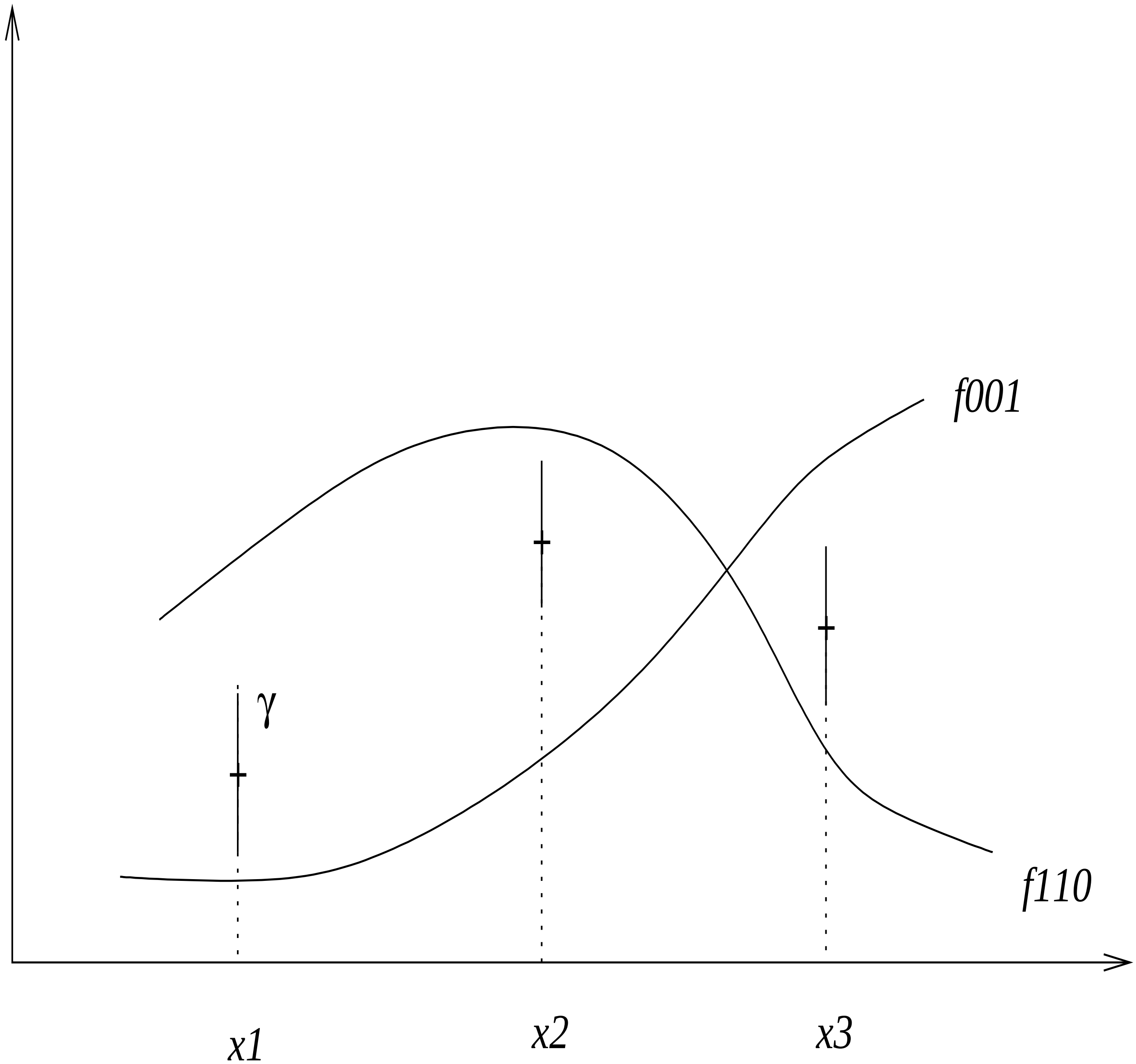
- set of points X is γ*-shattered by* $\mathcal{H}$ if for all binary vectors b indexed by X, there is a function $f_b \in \mathcal{H}$ satisfying

$$f_b(x) \begin{cases} \geq r_x + \gamma & \text{if } b_x = 1 \\ \leq r_x - \gamma & \text{otherwise,} \end{cases}$$

for some numbers r_x.

- The *fat shattering dimension* $\text{Fat}_{\mathcal{H}}$ of the set $\mathcal{H}$ at scale γ is the size of the largest γ-shattered set.

γ-shattering three points



γ-Growth function and Fat-shattering

The key result relating the two is the following for functions with range $[0, 1]$:

Theorem 1. *(Alon, Ben-David, Cesa-Bianchi and Haussler)*

$$\log_2 \mathcal{N}^m(\gamma, \mathcal{H}) \leq 1 + k \log_2 \left(\frac{2em}{k\gamma}\right) \log_2 \left(\frac{4m}{\gamma^2}\right)$$

where $k = \mathit{Fat}_{\mathcal{H}}(\gamma/4) \leq em$

- Similar form to Sauer's lemma except for extra $\log$ factor. Not known if this is necessary.

- (difficult) proof works by discretising the range and turning it into a combinatorial result.

- gives bound on large margin generalization in terms of $k = \mathrm{Fat}_{\mathcal{H}}(\gamma/8) \leq em$:

$$\epsilon(m, \mathcal{H}, \delta, \gamma) = \frac{2}{m}\left(k \log_2 \frac{8em}{k} \log_2(32m) + \log_2 \frac{8m}{\delta}\right)$$

Bounding Fat for linear functions

Let $\mathcal{H}$ be the set of linear functions with unit weight vectors restricted to inputs in a ball of radius R, then

$$\text{Fat}_{\mathcal{H}}(\gamma) \leq \frac{R^2}{\gamma^2}.$$

Proof: (Gurvits, Bartlett) Let $S = \{x^1, \ldots, x^m\}$ be a set of points that are γ shattered.

- For any $S_0 \subseteq S$, $\|\sum S_0 - \sum(S - S_0)\| \geq |S|\gamma$, follows from the fact that for w realising this split with margin γ, we have

$$\left\langle w \cdot \sum S_0 - \sum(S - S_0)\right\rangle \geq |S|\gamma.$$

- Hence, suffices to find an S_0 such that

$$\sqrt{|S|}R \geq \left\|\sum S_0 - \sum(S - S_0)\right\|.$$

continued ...

- For some $S_0 \subseteq S$,

$$\left\| \sum S_0 - \sum (S - S_0) \right\| \leq \sqrt{|S|} R.$$

Consider S_0 defined by a uniformly random $\{-1, 1\}$ vector b. Then for the expected value of the norm, we have:

$$\begin{aligned}
E \left\| \sum S_0 \right. & - \left. \sum (S - S_0) \right\|^2 = E \left\| \sum_{i=1}^{m} b_i x^i \right\|^2 \\
& = E \left\langle \sum_{i=1}^{m} b_i x^i \cdot \sum_{j=1}^{m} b_j x^j \right\rangle \\
& = \sum_{i=1}^{m} E \left\| b_i x^i \right\|^2 + \sum_{i \neq j} E \left\langle b_i x^i \cdot b_j x^j \right\rangle \\
& \leq |S| R^2
\end{aligned}$$

Generalization of SVMs

For distribution with support in ball of radius R, (eg Gaussian Kernels $R = 1$) and margin γ, have bound:

$$\epsilon(m, \mathfrak{L}, \delta, \gamma) = \frac{2}{m}\left(k \log_2 \frac{8em}{k} \log_2(32m) + \log_2 \frac{8m}{\delta}\right)$$

where $k = \frac{64R^2}{\gamma^2}$.

- Apparently contradicts lower bound for eg Gaussian kernels where VC $= \infty$.
- Hence, quality of the bound must be distribution dependent since there exist distributions which force high error
- BUT bound holds independently of the distribution – just won't be good – is good if we are *lucky*
- γ measures the benigness of the distribution relative to learning task

Agnostic results

- The next result to be obtained was an ‘agnostic’ result, ie with training errors, except that the errors are now ‘margin’ errors:

Theorem 2. *(Bartlett) With probability at least $1-\delta$, every linear classifier $f \in \mathcal{F}$ has error no more than*

$$b/m + \sqrt{\frac{c}{m}\left(\frac{R^2}{\gamma^2}\log^2 m + \log(1/\delta)\right)}$$

where b is the number of labelled training examples with margin less than γ.

- Measure of the distribution of margin values is γ its b/m percentile. Bound involves the square root of the ratio of the fat shattering dimension and sample size.

- A result involving the norm of the slack variables will be mentioned later.

Examples

The result is more general that just SVMs. In order to apply we simply need classes for which we know a bound on the fat shattering dimension.

- For $\mathcal{H}$ single hidden layer neural networks, with linear output node and input dim n, Gurvits and Koiran showed (B bounds the 1-norm of the output weights, but no limit on their number!):

$$\text{Fat}_{\mathcal{H}}(\gamma) \le O\left(\frac{B^2n^2}{\gamma^2}\log\frac{B^2n^2}{\gamma}\right),$$

- Generalised by Bartlett to $\mathcal{F} =$ neural networks with L layers, V the 1-norm of weights into each layer, and B the Lipschitz constant for the activation function (provided $V \ge 1/(2B)$, and $\gamma \le 16VB$):

$$\text{Fat}_{\mathcal{F}_L}(\gamma) \le \frac{1}{6}\left(\frac{48}{\gamma}\right)^{2L}(2VB)^{L(L+1)}\log(2n+2),$$

Boosting and the margin

- SVMs were not the only learning system that seemed to contradict traditional views of generalization.

- Adaboost combines weak learners in a weighted majority voting scheme. The t-th weak learner h_t (output in $\{-1, 1\}$) is trained in an altered distribution $D_t(i)$ to give error ϵ_t. The distribution is updated:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(x_i) = y_i \\ \exp(\alpha_t) & \text{otherwise,} \end{cases}$$

  where Z_t is a normalisation and $\alpha_t = 0.5 \ln((1 - \epsilon_t)/\epsilon_t)$. The final hypothesis is the sign of

$$f(x) = \sum_t \alpha_t h_t(x)$$

- Practical experiments showed that continuing to add new weak learners after correct classification of the training set had been achieved could further improve test set performance.

Boosting the margin

- Plots showed that the margin of $f(x)$ on the training set continued to grow as more weak learners were added.
- The reason is that the distribution $D_t(i)$ is a function of the margin of the current hypothesis

$$f_t(x) = \sum_{j=1}^{t} \alpha_j h_j(x)$$

$$\begin{aligned} D_t(i) &= \frac{1}{m} \prod_{j=1}^{t} \frac{\exp(-y_i \alpha_j h_j(x_i))}{Z_j} \\ &= \frac{1}{m \prod_{j=1}^{m} Z_j} \exp\left(-y_i \sum_{j=1}^{t} \alpha_j h_j(x_i)\right) \\ &= \frac{1}{m \prod_{j=1}^{m} Z_j} \exp(-y_i f_t(x_i)) \end{aligned}$$

Boosting and fat shattering

- So boosting weights points as an exponential function of their margin and hence increases the margin of the cumulative hypothesis.

- The following theorem shows that for weak learners from a low VC class the set of boosted functions has bounded fat shattering dimension independently of the number of boosting stages – implying that the margin bounds can be applied.

Theorem 3. *(Schapire* et al.*) There is a constant c so that for all classes H, the class of convex combinations of functions from H,*

$$\mathcal{F} = \left\{ x \mapsto \sum_{i=1}^{N} w_i f_i(x) : f_i \in H,\ w_i > 0,\ \sum_i w_i = 1 \right\}$$

satisfies

$$\mathit{Fat}_{\mathcal{F}}(\gamma) \leq c \frac{\mathit{VCdim}(H)}{\gamma^2} \log(1/\gamma),$$

Summary of this section

- The PAC model fails to explain the performance of SVMs and Boosting
- The new bounds rely on the margin as an indication of luckiness of the distribution generating the data and its relation to the target hypothesis
- The two algorithms are able to exploit this fortuitous relation that appears to be very common in real-world applications
- Technically the new bounds have a similar flavour to the classical PAC bounds, with the following correspondence:

Growth function	–	γ-growth function
Vapnik Chervonenkis dim	–	Fat shattering dim
Sauer's Lemma	–	Alon *et al.*

BREAK — 5 minutes

(STRUCTURE)

1. Basic PAC Ideas
2. Basic Margin Ideas
3. **Their Exploitation**
4. **And Extension**
5. **Conclusions**

THEIR EXPLOITATION

Large margin ideas allow “linear” classes to perform much better.

Advantage: linear classes *much* easier to analyze.

We will now exploit the linear nature of LM classes.

1. Calculation of covering numbers

2. SV machines

3. Convex combinations

Calculation of Covering Numbers — Entropy Numbers and Operators

Entropy numbers ϵ_n are the functional inverse of the covering numbers $\mathcal{N}(\epsilon) = \mathcal{N}(\epsilon, \mathcal{F}, d)$.

The nth *entropy number of a set* $M \subset E$, for $n \in \mathbb{N}$, is

$$\epsilon_n(M) := \inf\{\epsilon > 0 \quad : \quad \text{there exists an } \epsilon\text{-cover for } M \text{ in } E \text{ containing } n \text{ or fewer points}\}$$

Example: $\mathcal{N}(\epsilon) \sim \epsilon^{-d} \;\Rightarrow\; \epsilon_n \sim n^{-1/d}$

Entropy Numbers of Operators — 1

Function class as an image of an operator

Entropy Numbers of Operators — 2

Consider bounded linear operators T between the normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$, i.e. operators such that the image of the (closed) unit ball

$$U_E := \{x \in E \colon \|x\|_E \leq 1\}$$

is bounded.

The smallest such bound is called the *operator norm*,

$$\|T\| := \sup_{x \in U_E} \|Tx\|_F.$$

Entropy Numbers of Operators — 3

The *entropy numbers of an operator* $T \in \mathfrak{L}(E, F)$ are defined as

$$\epsilon_n(T) := \epsilon_n(T(U_E)) = \epsilon_n(T(U_E), F)$$

Meaning of entropy number of T

We have $\|T\| = \epsilon_1(T) \geq \epsilon_2(T) \geq \cdots$ and $\epsilon_{kl}(ST) \leq \epsilon_k(S)\epsilon_l(T)$

Dyadic Entropy Numbers

The *dyadic entropy numbers of an operator* are defined by

$$e_n(T) := \epsilon_{2^{n-1}}(T), \quad n \in \mathbb{N}.$$

The dyadic entropy numbers are the functional inverse of $\log \mathcal{N}(\epsilon)$.

Properties:

$\|T\| \geq e_1(T) \geq e_2(T) \geq \cdots \geq 0.$

$\forall k, l \in \mathbb{N}$, $e_{k+l-1}(ST) \leq e_k(S) e_l(T)$.

Replace e by s (singular values) and same theorem holds.

(Why mathematicians say e_k are "s-numbers")

An example

The identity operator from $\ell_{p_1}^m$ to $\ell_{p_2}^m$ is defined by

$$\begin{aligned} \mathrm{id}\colon \ell_{p_1}^m &\to \ell_{p_2}^m \\ \mathrm{id}\colon x &\mapsto x \end{aligned}$$

What id does

Thus the $\epsilon_n(\mathrm{id}\colon \ell_{p_1}^m \to \ell_{p_2}^m)$ is the smallest value of ϵ such that n $\ell_{p_2}^m$ balls of radius ϵ cover the $U_{\ell_{p_1}^m}$ (the unit ball in $\ell_{p_1}^m$).

Examples of Entropy Numbers (cont)

Let $0 < p_1 \leq p_2 \leq \infty$. Then

$$e_k(\mathrm{id}\colon \ell_{p_1}^m \to \ell_{p_2}^m) \leq$$

$$c \begin{cases} 1 & \text{if } 1 \leq k \leq \log m \\ (k^{-1} \log(1 + \frac{m}{k}))^{1/p_1 - 1/p_2} & \text{if } \log m \leq k \leq m \\ 2^{-k/m} m^{1/p_2 - 1/p_1} & \text{if } k \geq m \end{cases}$$

for $k \in \mathbb{N}$ where c is a positive constant independent of m and k depends on p_1 and p_2.

The constants can be determined explicitly, e.g.:

$$e_{k+1}(\mathrm{id}\colon \ell_2^m \to \ell_\infty^m) \leq 1.86 \left(\frac{\log(\frac{m}{k} + 1)}{k} \right)^{1/2}$$

Proof: Clever counting of how many square boxes needed to cover a round ball.

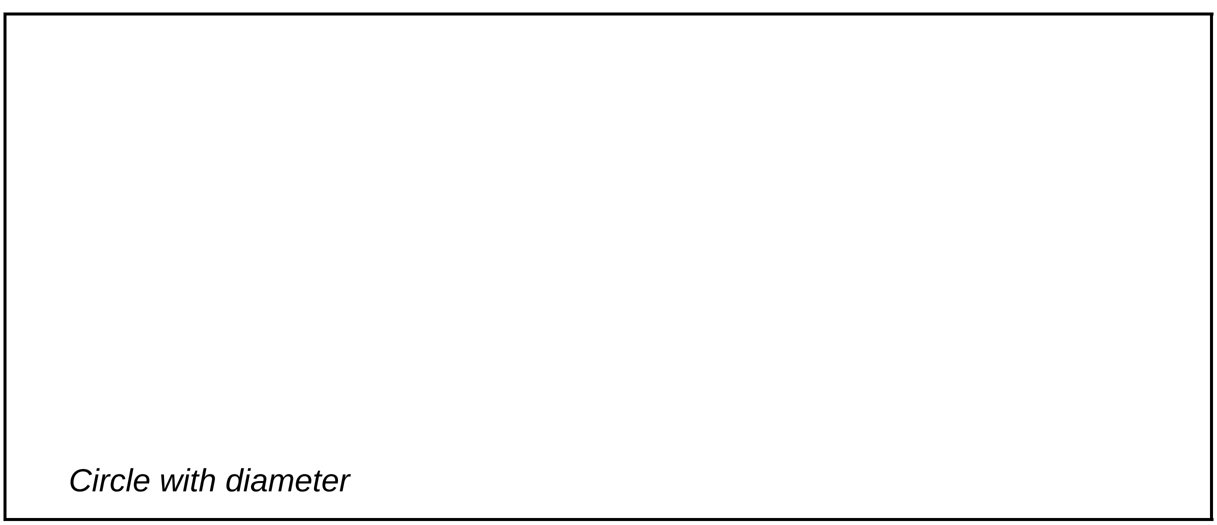
Examples of Entropy Numbers — $T\colon H \to \ell_\infty^m$

The Maurey-Carl theorem states: If H is a Hilbert space, then

$$e_k(T : H \to \ell_\infty^m) \leq 26\|T\| \left(\frac{\log(\frac{m}{k} + 1)}{k} \right)^{1/2}.$$

Significance: does not depend on dimension of H.

Intuition: Projections in H don't increase norms.



Application of Maurey's Theorem

$\mathcal{F}_{2,2} := \{\langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \ell_2,\ \|\mathbf{w}\|_2 \leq 1,\ \mathbf{x} \in \ell_2,\ \|\mathbf{x}\|_2 \leq 1\}.$

This is the class of functions MM algorithms work with.

There exists $c > 0$ such that for all $n \in \mathbb{N}$, and all $\epsilon > 0$,

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{2,2}) \leq \begin{cases} \frac{c \log(m)}{\epsilon^2} & \epsilon \geq \frac{1}{\sqrt{m}} \\ cm \log(\frac{1}{m\epsilon}) & \epsilon < \frac{1}{\sqrt{m}}. \end{cases}$$

By comparison, Alon et al. plus $\mathrm{fat}_{\mathcal{F}_{2,2}}(\gamma) \leq 1/\gamma^2$ implies

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{2,2}) \leq \frac{c}{\epsilon^2} \log^2(m)$$

Variations

$$\mathcal{F}_{p_1,p_2}{}^{M} := \{\langle \mathbf{w}, \mathbf{x}\rangle,\ \|\mathbf{w}\|_{\ell_{p_1}}^{M} \le 1,\ \|\mathbf{x}\|_{\ell_{p_2}}^{M} \le 1\}.$$

For $p \neq 2$, $U_{\ell_p^m}$ is not rotationally invariant. This means that projections of ℓ_p^m onto a subspace are no longer norm 1.

Result is a dependence on M (dimension of $\mathbf{w}$).

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{1,\infty}^{M}) \le \frac{21.6 m^{1/2} \log^{1/2}(m) \log^{1/2}(M)}{\epsilon}.$$

Cf. EG algorithm mistake bounds.

SV Machines

This machinery can be used to understand covering numbers of SV classes.

Basic idea. Feature space map

$$\Phi(\mathbf{x}) = (\sqrt{\lambda_1}\psi_1(\mathbf{x}), \sqrt{\lambda_2}\psi_2(\mathbf{x}), \ldots)$$

$\lambda_1 \geq \lambda_2 \geq \cdots$

Thus $\Phi(\mathcal{X})$ is not a ball; it is a squashed ellipse.

Introduce a scaling operator A which turns the ellipse into a ball in order to analyze. Take account of A.

Illustration of SV idea

See Ying Guo’s talk for more.

Convex Combinations of Various Types

Suppose $p > 0$, and S is a set. The p-convex hull of F is

$$\begin{aligned} \mathrm{co}_p(S) &= \bigcup_{n \in \mathbb{N}} \left\{ \sum_{i=1}^{n} \alpha_i f_i \colon f_1, \ldots, f_n \in F, \right. \\ & \left. \alpha_1, \ldots, \alpha_n \in \mathbb{R},\ \sum_{i=1}^{n} |\alpha_i|^{p} \leq 1 \right\} \end{aligned}$$

(e.g.) 1-convex hull of Heavisides on $[0,1]$ is the set of functions of bounded variation.

Carl (plus many others) have many results on $\epsilon_n(\mathrm{co}_1(S))$. If $\mathcal{N}(\epsilon, H) \sim \left(\frac{1}{\epsilon}\right)^d$ for some $d \in \mathbb{N}$. Then

$$\mathcal{N}(\epsilon, \mathrm{co}_1(H)) \sim \left(\frac{1}{\epsilon}\right)^{\frac{2d}{2+d}}$$

Cf. $\frac{1}{\epsilon}^2 \log^2(1/\epsilon)$ via simple Maurey plus Alon et al.

When $p < 1$

Graphical illustration of p-convex hull

Suppose $\mathcal{N}(\epsilon, H) \sim \left(\frac{1}{\epsilon}\right)^d$ for some $d \in \mathbb{N}$. Then

$$\log \mathcal{N}(\epsilon, \mathrm{co}_p(H)) \sim c(p) d \left(\frac{1}{\epsilon}\right)^{\frac{2p}{2-p}} \log\left(\frac{1}{\epsilon}\right)$$

For $p = 1$ this is $O((1/\delta)^2 \log(1/\delta))$.

Right rate is $O((1/\delta)^{\frac{2d}{d+2}})$. Difference negligible for large d.

Summary

- Margin analysis has added impetus to get better bounds for covering numbers

- Viewing classes as images of linear operators allows use of bag of existing theory.

- Could equally well state results in terms of fat, but little incentive to do so.

- Viewpoint suggests new algorithms and explains effect of others (e.g. LP machines).

AND THEIR EXTENSION

1. LP Machines
2. Decision Trees
3. Margin Distribution
4. General Data-Dependent Hierarchies in SRM
5. Different Learning Problems

LP Machines

Use kernels, but not necessarily Mercer kernels.

Need to assume induces “trace-class operator” ($\sum_i |\lambda_i| \leq \infty$)

Can then obtain bounds on ϵ-covering numbers of

$$\mathrm{co}_\Lambda \mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R}^d \,\middle|\, f(x) = \sum_i \alpha_i k(x_i, x) \right.$$
$$\left. \text{with } \alpha_i \in \mathbb{R}^d,\ \sum_i \|\alpha_i\|_{\ell_1^d} \leq \Lambda,\ x_i \in \mathcal{X} \right\}$$

Roughly speaking, replace A scaling operator from SV case by A^2.

This is the class of functions used in “Linear Programming Machines” (Mangasarian and others).

Use the same statistical result as in SVM analysis.

Decision Trees

- Perceptron decision trees have perceptrons at the decision nodes – usually no kernels involved – standard heuristic algorithm OC1.

- Run OC1 and replace hyperplanes with max marg hyperplanes implementing same split improves generalization.

- Can also put the margin as a criterion into the heuristic search for a split.

- Generalization bound in terms of margins $(\gamma_i)_{i=1}^K$ classifying an m sample from region of radius R: with probability greater than $1-\delta$ less than

$$\frac{130R^2}{m}\left(D'\log(4em)\log(4m)+\log\frac{(4m)^{K+1}\binom{2K}{K}}{(K+1)\delta}\right)$$

where $D'=\sum_{i=1}^K \frac{1}{\gamma_i^2}$.

A Bayesian Connection

- Can argue that integrating the posterior distribution over the function space leads to a large margin classifier:

$$P(y|x, D) = \int_{\lambda} f(x, \lambda) p(\lambda|D) dP(\lambda)$$

in the Hilbert space given by the functions

$$\mathcal{H} = \left\{ \mathbf{z} : \Lambda \to \mathbb{R} | \text{ such that } \int_{\lambda \in \Lambda} \mathbf{z}(\lambda)^2 dP(\lambda) < \infty \right\}$$

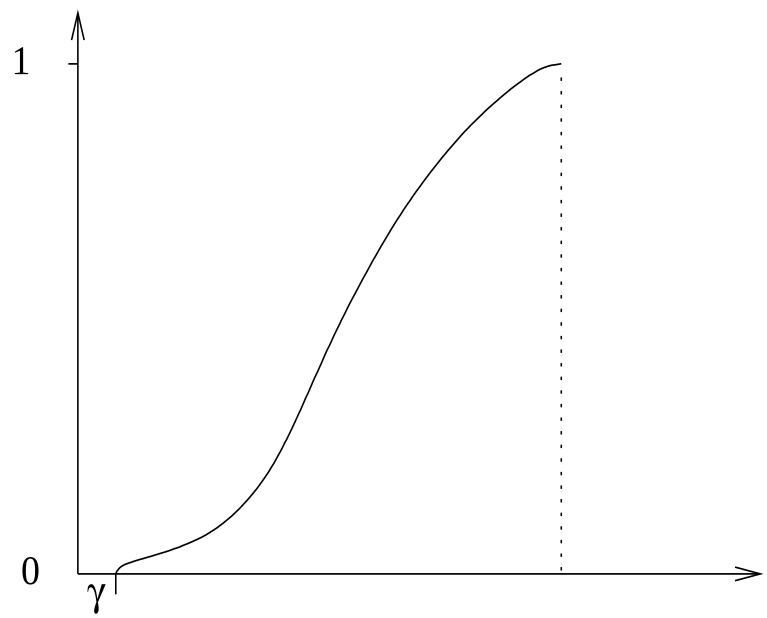
with the inner product

$$\langle \mathbf{z}_1 \cdot \mathbf{z}_2 \rangle = \int_{\lambda \in \Lambda} \mathbf{z}_1(\lambda) \mathbf{z}_2(\lambda) dP(\lambda).$$

- Relation to other important new learning technique – Gaussian processes

Margin Distribution

- Plots of the cumulative distribution of margin values frequently look something like this:



- Intuitively feels wrong to rely on the value of γ which may depend on only a small number of points and may be negative.

- Open question what is the "right" measure of the distribution for predicting generalization.

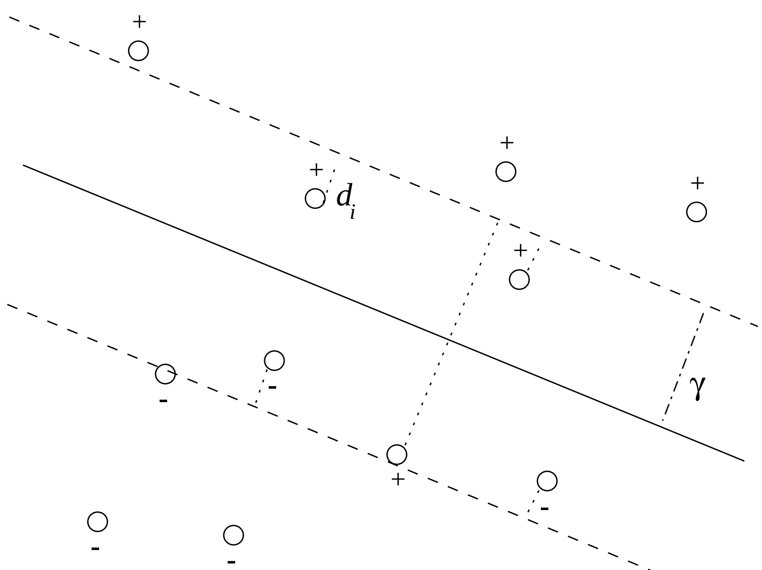
New measures of Margin Distribution

- Percentile result involves the square root and so is fully agnostic and correspondingly weaker.

- DOOM (Direct Optimization of Margin) implements a strategy for pushing the distribution with good results.

- Recent result shows a bound in terms of the quantity $D = \|d\|_2$, where: d_i is amount by which (x_i, y_i) fails to have margin γ

$$= \max\{0, \gamma - y_i f(x_i)\}$$



Generalization in terms of Margin Distribution

- Behaves like a class with fat shattering dimension:

$$k = \left\lfloor \frac{[(R+D)^2 + 2.25RD]}{\gamma^2} \right\rfloor$$

  which can be much smaller than $R^2/\gamma_{\min}^2$.

- Optimizing this bound corresponds to minimising the 2-norm of the slack variables – Cortes and Vapnik – now provably okay way to avoid NP-completeness of minimising number of training errors.

- Can also obtain bound in terms of 1-norm of slacks – box constraint algorithm.

- Generalization to non-linear classes at this conference.

General Data-Dependent Hierarchies

- We can view the margin criterion as breaking our set of hypotheses into different classes:

$$H_{\gamma_1} \subset H_{\gamma_2} \subset \ldots \subset H_{\gamma_i} \subset \ldots,$$

where H_{γ_i} are hypotheses with margin γ_i on the training set, with $\gamma_1 > \gamma_2 > \ldots > \gamma_i > \ldots$. At first sight we appear to be doing Structural Risk Minimisation over this hierarchy, i.e. choosing the first i for which H_{γ_i} has a consistent hypothesis.

- Problem is that hierarchy is ‘data-dependent’, which violates the SRM principle.

- Hence, margin analysis is *one* way of doing data-dependent SRM.

Other Data-Dependent Hierarchies

- Can bound generalization error of SVMs in terms of number h of support vectors since they form a compression scheme (Littlestone and Warmuth): with probability $1-\delta$,

$$\epsilon(m, \mathfrak{L}, \delta, h) = \frac{1}{m-h}\left(h \log_2 \frac{em}{h} + \log_2 \frac{m}{\delta}\right).$$

- Hence h is again an indication of a benign relation between distribution and target function.

- This concept has been generalized to the notion of a luckiness function,

$L(f, \mathbf{X}, \mathbf{y}) =$how lucky function f is with data $\mathbf{X}, \mathbf{y}$.

Other Luckiness Functions

- Lugosi and Pintér algorithm which splits sample, finds cover on first half and chooses which element by labels of the second half.
- Volume of a ball that can be fitted into consistent region of version space – relation to Bayesian evidence. (ST and Williamson)
- VC dimension of the function class restricted to the sample. (ST, Bartlett, Williamson, Anthony)
- Covering numbers of the function class restricted to the sample. (ST and Williamson)
- Microchoice algorithms (Langford and Blum, presented on thursday)
- Your favourite intuition about collusions between distributions and target functions.

CONCLUSION

- Have shown how to "marginalize" a number of results/techniques.
- Margins analysis revitalizes linear methods: makes them competitive with harder to analyse non-linear methods.
- But can also marginalise non-linear (e.g. NN)
- Margin approach allows refinement of basic PAC ideas to take account of distribution. In general *luckiness* measures the serendipitous simplicity of the hypothesis and the distribution *together*.

It’s Exciting Because...

- Makes the “PAC” theory more practical (well closer to being practical)
- Makes a formal link between two seemingly different camps (PACmen and Bayesians)
- Explains recent algorithms (SV/Boost/SV Soft Margin)
- Suggests new algorithms (DOOM)
- Illustrates the power of the SRM principle: its apparent weakness (you need a bound to get an algorithm) is its strength: once you have a bound, you have an algorithm.
- Is thus a vital and lively field.

Where to find out more

Books:

B. Schölkopf *et al.* (Eds.), *Advances in Kernel Methods*, MIT Press 1999. (includes several overview papers)

Martin Anthony and Peter Bartlett, *Neural Network Learning: Theoretical Foundations*, To be published by Cambridge University Press, 1999.

Alex Smola *et al.* (Eds) *Large Margin Classifiers*, To be published by MIT Press 1999. [Based on NIPS'98 Worshop] (includes extensive introductory chapter)

Berd Carl and Irmtraud Stephani, *Entropy, Compactness and the Approximation of Operators*, Cambridge University Press, 1990.

Vladimir Vapnik, *Statistical Learning Theory*, John Wiley, 1998.

Papers and Websites

Papers:

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson and Martin Anthony, "Structural Risk Minimization over Data-Dependent Hierarchies", *IEEE Transactions on Information Theory*, **44**(5), 1926–1940 (1998).

Peter Bartlett, "The Sample Complexity of Pattern Classsification with Neural Networks: the Size of the Weights is more important than the size of the network", *IEEE Transactions on Information Theory*, **44**(2), 525–536 (1998).

Robert Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," Annals of Statistics **26**(5), 1651–1686 (1998).

Several Neurocolt reports

Web Sites:

`svm.first.gmd.de` (SV Machines)

`www.neurocolt.com` (Neurocolt: lots of TRs)

References

[1] M. Anthony and P. Bartlett. Function learning from interpolation. Technical Report 94–013, NeuroCOLT, 1994. An extended abstract appeared in EuroCOLT'95, Paul Vitanyi, ed., LNAI 904, Springer Verlag, 1995, 211–221.

[2] P. Bartlett, P. Long, and R. Williamson. Fat–Shattering and the Learnability of Real–Valued Functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

[3] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 43–54, Cambridge, MA, 1999. MIT Press.

[4] P.L. Bartlett. The sample complexity of pattern classsification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Information theory*, 44(2):525–536, 1998.

[5] P.L. Bartlett, S.R. Kulkarni, and S.E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 1997. (to appear).

[6] N. Cristianini and J. Shawe-Taylor. Bayesian voting schemes and large margin classifiers. In B. Schölkopf,

C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 55–68, Cambridge, MA, 1999. MIT Press.

[7] N. Cristianini, J. Shawe-Taylor, and P. Sykacek. Bayesian classifiers are large margin hyperplanes in a hilbert space. In J. Shavlik, editor, *Machine Learning: Proceedings of the Fifteenth International Conference*, San Francisco, CA, 1998. Morgan Kaufmann.

[8] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161, Cambridge, MA, 1997. MIT Press.

[9] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. In *to appear Proceedings of the Twelth Conference on Computational Learning Theory, COLT'99*, 1999.

[10] G. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *Advances in Neural Information Processing Systems*, volume 11, Cambridge, MA, 1999. MIT Press.

[11] L. Mason, P.L. Bartlett, and J. Baxter. Direct optimization of margins improves generalization in combined classifiers. In *Advances in Neural Information*

Processing Systems 12, Cambridge, MA, 1999. MIT Press. forthcoming.

[12] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.

[13] B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Support vector regression with automatic accuracy control. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of ICANN'98*, Perspectives in Neural Computing, pages 111 – 116, Berlin, 1998. Springer Verlag.

[14] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. Technical Report NC-TR-98-027, NeuroColt2, University of London, UK, 1998. submitted to Neural Computation.

[15] B. Schölkopf, A.J. Smola, P. Bartlett, and R. Williamson. Shrinking the tube — a new support vector regression algorithm. In *Neural Information Processing Systems 1998*, Boston, MA, 1999. MIT Press. forthcoming.

[16] J. Shawe-Taylor. Confidence estimates of classification accuracy on new examples. In *Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'97, Lecture Notes in Articifical Intelligence, 1208*, pages 260–271. Springer, 1997.

[17] J. Shawe-Taylor. Classification accuracy based on observed margin. *Algorithmica*, 22:157–172, 1998.

[18] J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. A framework for structural risk minimization. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76, New York, 1996. Association for Computing Machinery.

[19] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[20] J. Shawe-Taylor and N. Cristianini. Data-dependent structural risk minimisation for perceptron decision trees. In *Advances in Neural Information Processing Systems 10*, pages 336–342, 1998.

[21] J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *to appear Proceedings of the Twelth Conference on Computational Learning Theory, COLT'99*, 1999.

[22] J. Shawe-Taylor and Nello Cristianini. Robust bounds on generalization from the margin distribution. NeuroCOLT Technical Report NC-TR-1998-020, ESPRIT NeuroCOLT2 Working Group, http://www.neurocolt.com, 1998.

[23] J. Shawe-Taylor and R.C. Williamson. Generalization performance of classifiers in terms of observed covering

numbers. In *Proceedings of EuroCOLT'99*, 1999.

[24] A. J. Smola and B. Schölkopf. On a kernel–based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

[25] A.J. Smola and B. Schölkopf. From regularization operators to support vector kernels. In *Advances in Neural information processings systems 10*, pages 343–349, San Mateo, CA, 1998.

[26] A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[27] D. Wu, K.P. Bennett, N. Cristianini, and J. Shawe-Taylor. Large margin decision trees for induction and transduction. In *Proceedings of International Conference on Machine Learning, ICML'99*, 1999.