Learning Theory

John Shawe-Taylor
Centre for Computational Statistics
and Machine Learning
University College London
jst@cs.ucl.ac.uk

September, 2009

STRUCTURE

PART A

1. General Statistical Considerations
2. Basic PAC Ideas and proofs
3. Real-valued Function Classes and the Margin

PART B

1. Rademacher complexity and Main Theory
2. Applications to classification
3. Conclusions

Aim:

- Some thoughts on why theory
- Basic Techniques with some deference to history
- Insights into proof techniques and statistical learning approaches
- Concentration inequalities and relation to Rademacher approach

What won't be included:

- The most general results
- Complete History
- Analysis of Bayesian inference
- Most recent developments, eg PAC-Bayes, local Rademacher complexity, etc.

PART A

Theories of learning

- Basic approach of SLT is to view learning from a statistical viewpoint.

- Aim of any theory is to model real/ artificial phenomena so that we can better understand/ predict/ exploit them.

- SLT is just one approach to understanding/ predicting/ exploiting learning systems, others include Bayesian inference, inductive inference, statistical physics, traditional statistical analysis.

Theories of learning cont.

- Each theory makes assumptions about the phenomenon of learning and based on these derives predictions of behaviour
  - every predictive theory automatically implies a possible algorithm
  - simply optimise the quantities that improve the predictions
- Each theory has strengths and weaknesses
  - generally speaking more assumptions, potentially more accurate predictions
  - BUT depends on capturing the right aspects of the phenomenon

General statistical considerations

- Statistical models (not including Bayesian) begin with an assumption that the data is generated by an underlying distribution P typically not given explicitly to the learner.

- If we are trying to classify cancerous tissue from healthy tissue, there are two distributions, one for cancerous cells and one for healthy ones.

General statistical considerations cont.

- Usually the distribution subsumes the processes of the natural/artificial world that we are studying.

- Rather than accessing the distribution directly, statistical learning typically assumes that we are given a 'training sample' or 'training set'

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$$

generated identically and independently (i.i.d.) according to the distribution P.

Generalisation of a learner

- Assume that we have a learning algorithm $\mathcal{A}$ that chooses a function $\mathcal{A}_{\mathcal{F}}(S)$ from a function space $\mathcal{F}$ in response to the training set S.

- From a statistical point of view the quantity of interest is the random variable:

$$\epsilon(S, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{(\mathbf{x},y)}\left[\ell(\mathcal{A}_{\mathcal{F}}(S), \mathbf{x}, y)\right],$$

  where ℓ is a 'loss' function that measures the discrepancy between $\mathcal{A}_{\mathcal{F}}(S)(\mathbf{x})$ and y.

Generalisation of a learner

- For example, in the case of classification ℓ is 1 if the two disagree and 0 otherwise, while for regression it could be the square of the difference between $\mathcal{A}_{\mathcal{F}}(S)(\mathbf{x})$ and y.

- We refer to the random variable $\epsilon(S, \mathcal{A}, \mathcal{F})$ as the generalisation of the learner.

- Note is random because of the dependence on the training set S.

Example of Generalisation I

- We consider the Breast Cancer dataset from the UCI repository.

- Use the simple Parzen window classifier: weight vector is

$$\mathbf{w}^+ - \mathbf{w}^-$$

where $\mathbf{w}^+$ is the average of the positive training examples and $\mathbf{w}^-$ is average of negative training examples. Threshold is set so hyperplane bisects the line joining these two points.

Example of Generalisation II

- Given a size m of the training set, by repeatedly drawing random training sets S we estimate the distribution of

$$\epsilon(S, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{(\mathbf{x},y)}\left[\ell(\mathcal{A}_{\mathcal{F}}(S), \mathbf{x}, y)\right],$$

  by using the test set error as a proxy for the true generalisation.

- We plot the histogram and the average of the distribution for various sizes of training set – initially the whole dataset gives a single value if we use training and test as the all the examples, but then we plot for training set sizes:

$$342, 273, 205, 137, 68, 34, 27, 20, 14, 7.$$

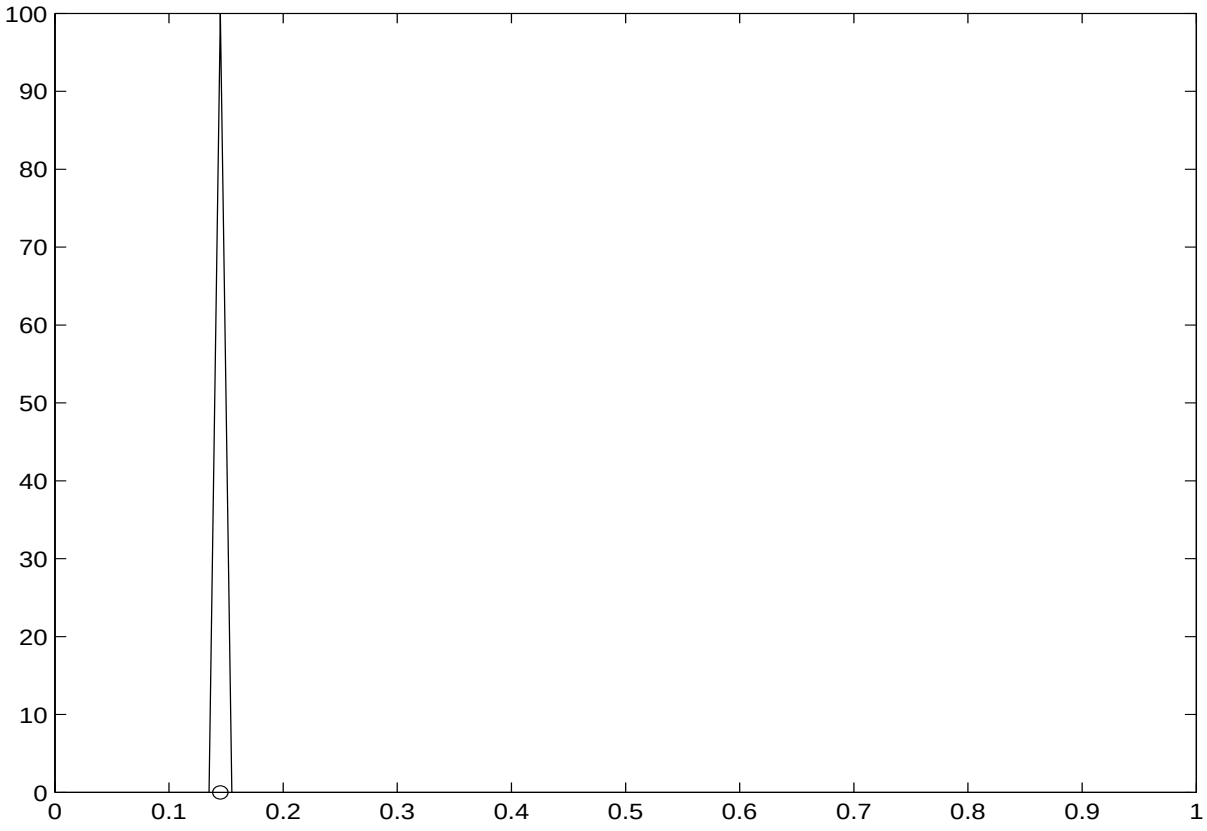
Example of Generalisation III

- Since the expected classifier is in all cases the same:

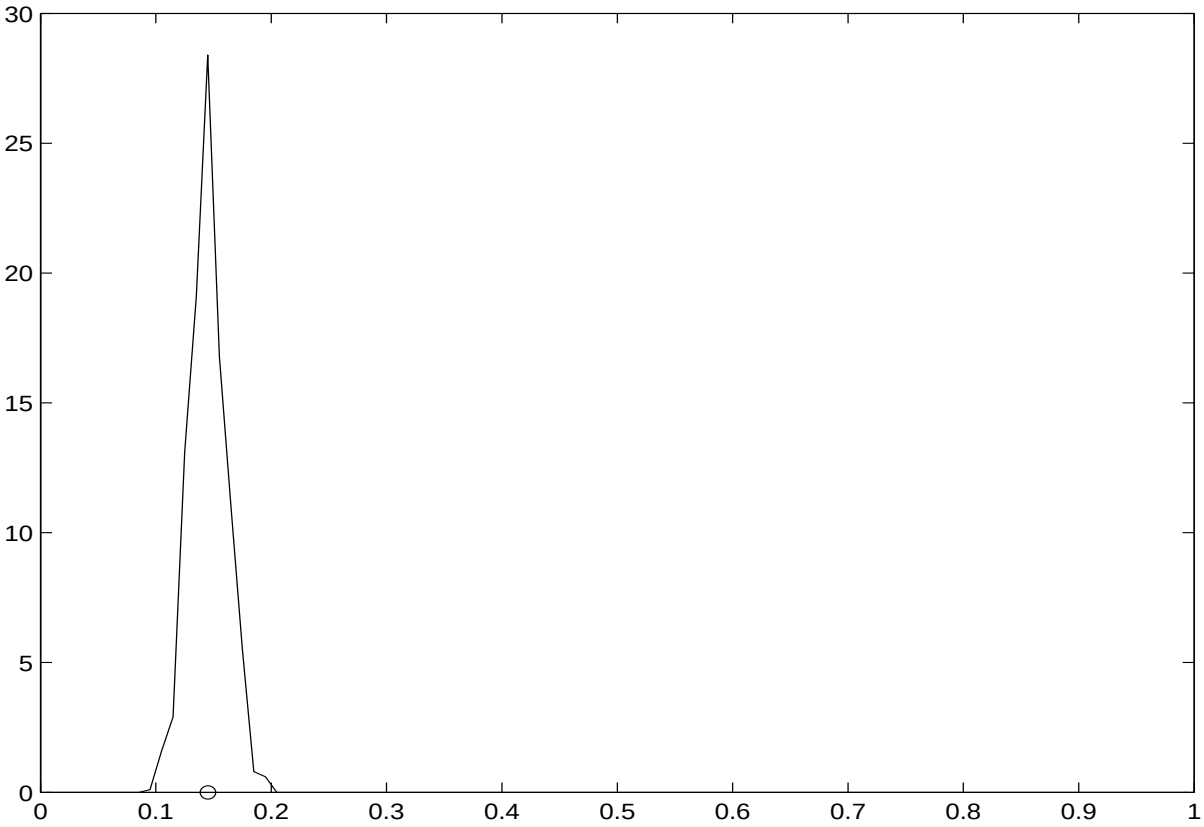
$$\begin{aligned}\mathbb{E}\left[\mathcal{A}_{\mathcal{F}}(S)\right] &= \mathbb{E}_S\left[\mathbf{w}_S^+ - \mathbf{w}_S^-\right] \\ &= \mathbb{E}_S\left[\mathbf{w}_S^+\right] - \mathbb{E}_S\left[\mathbf{w}_S^-\right] \\ &= \mathbb{E}_{y=+1}\left[\mathbf{x}\right] - \mathbb{E}_{y=-1}\left[\mathbf{x}\right],\end{aligned}$$

  we do not expect large differences in the average of the distribution, though the non-linearity of the loss function means they won't be the same exactly.

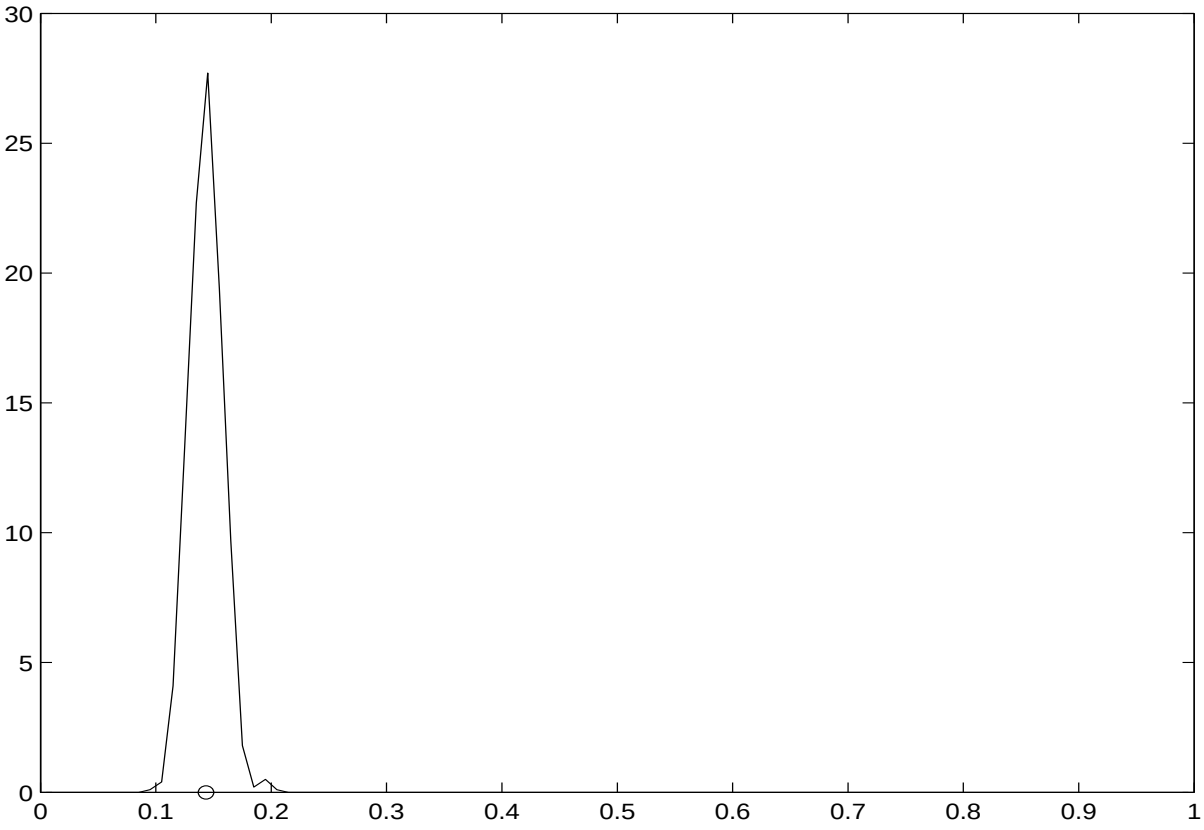
Error distribution: full dataset



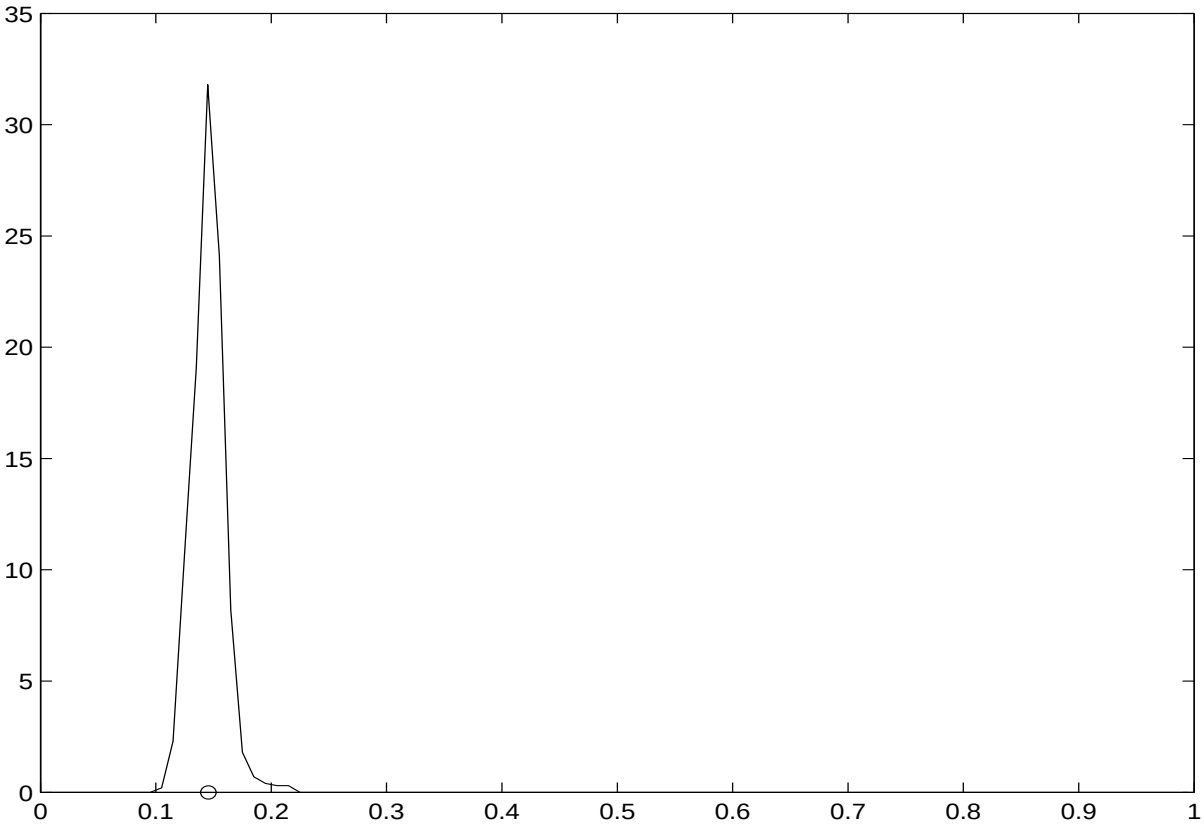
Error distribution: dataset size: 342



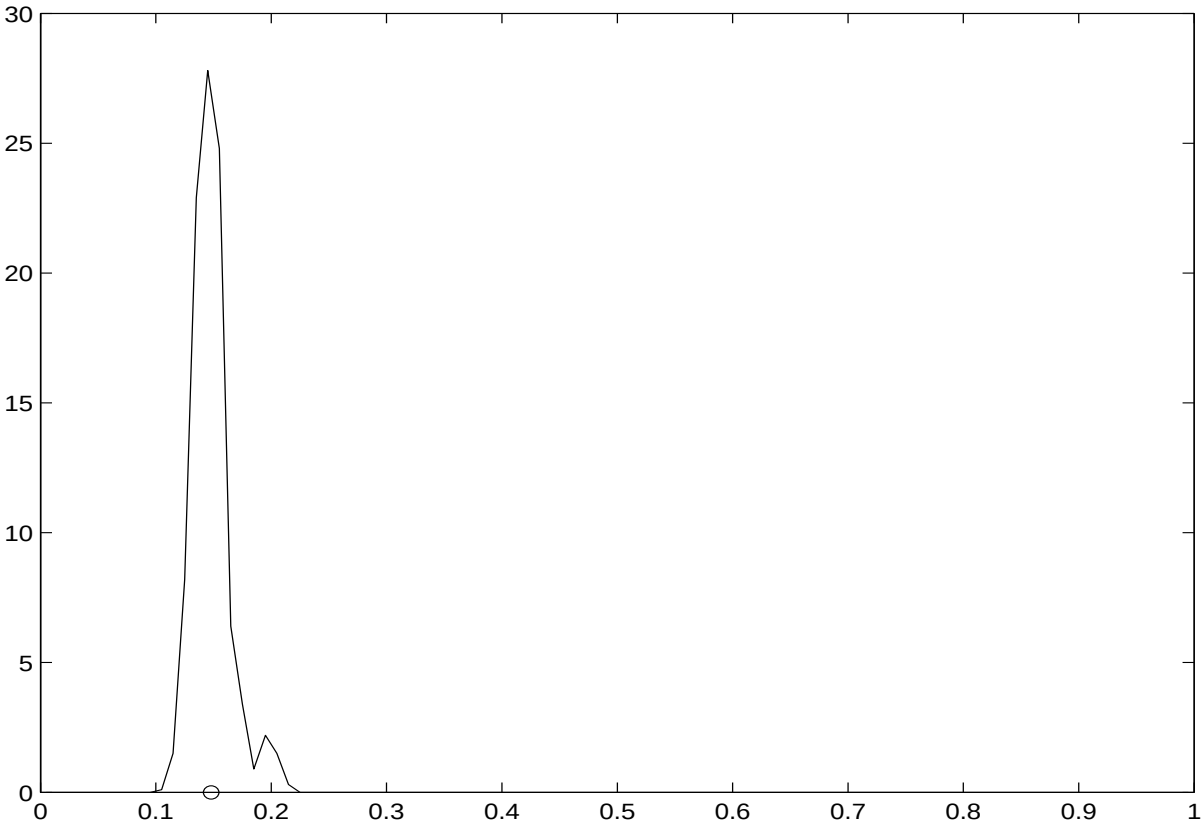
Error distribution: dataset size: 273



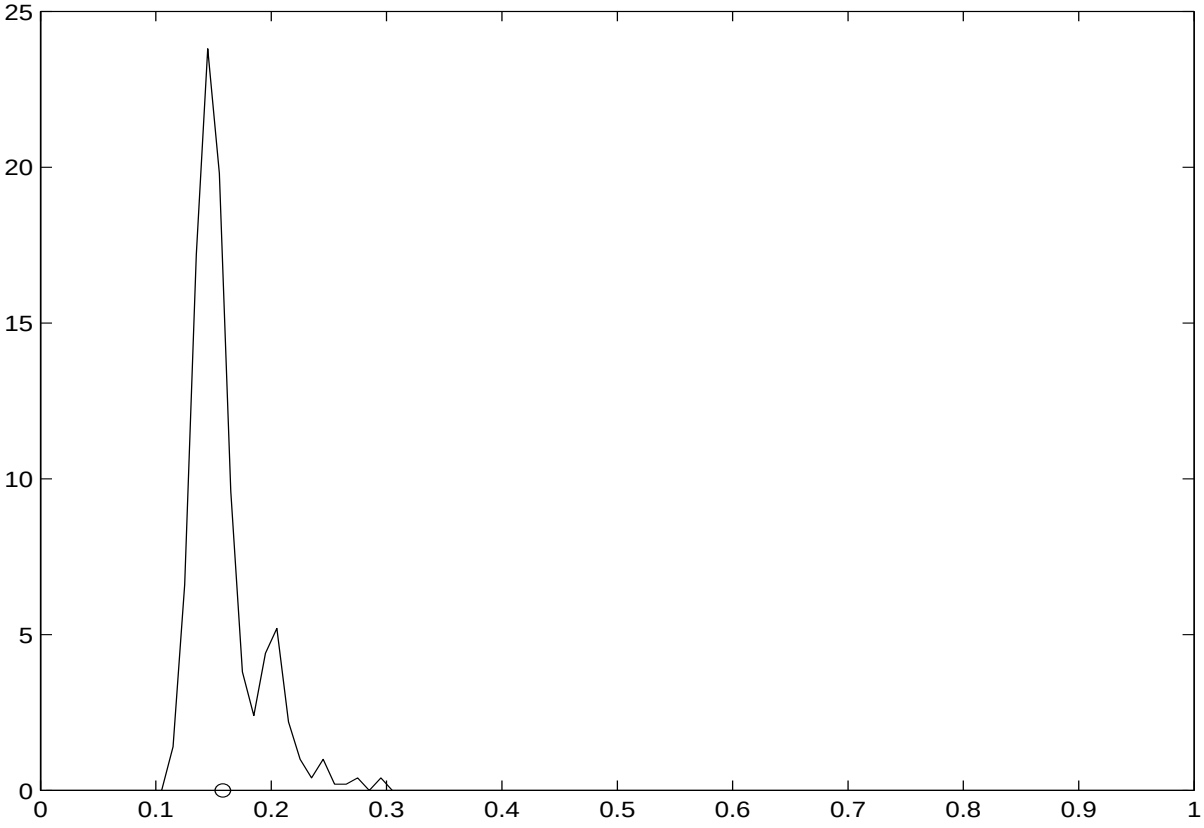
Error distribution: dataset size: 205



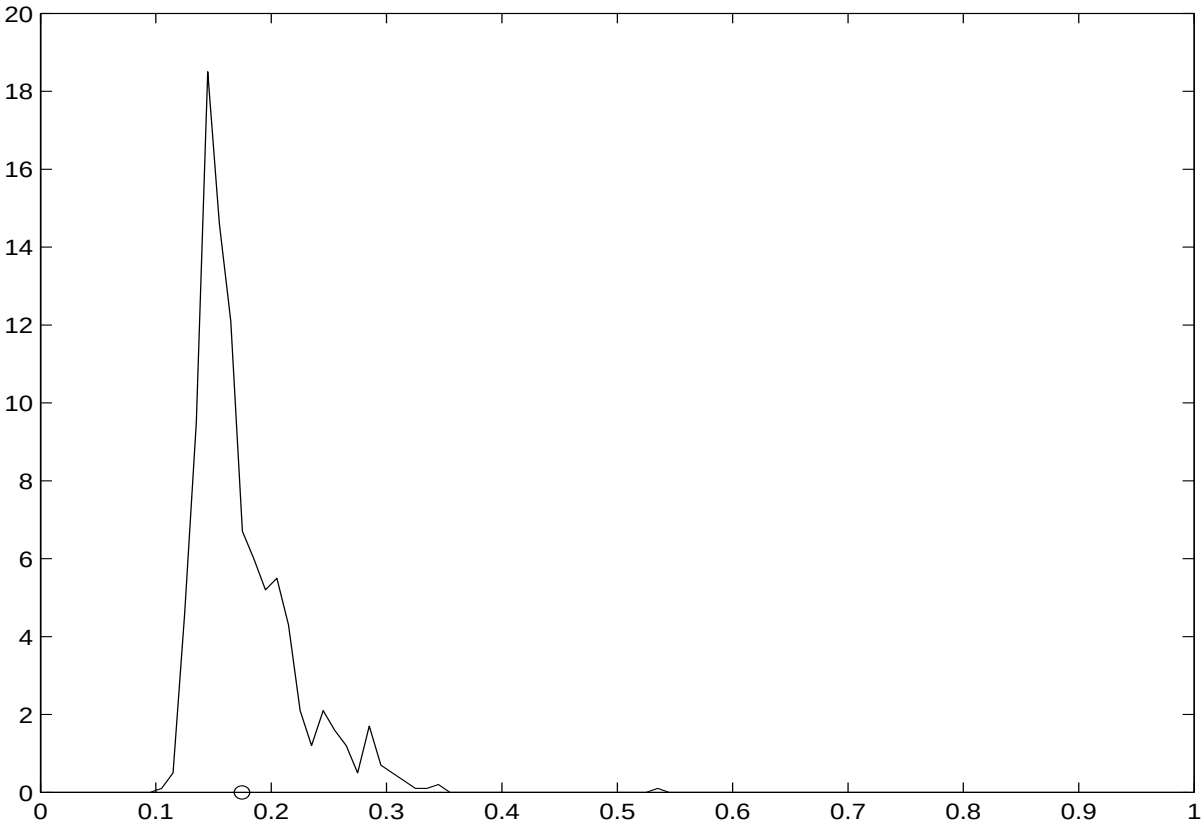
Error distribution: dataset size: 137



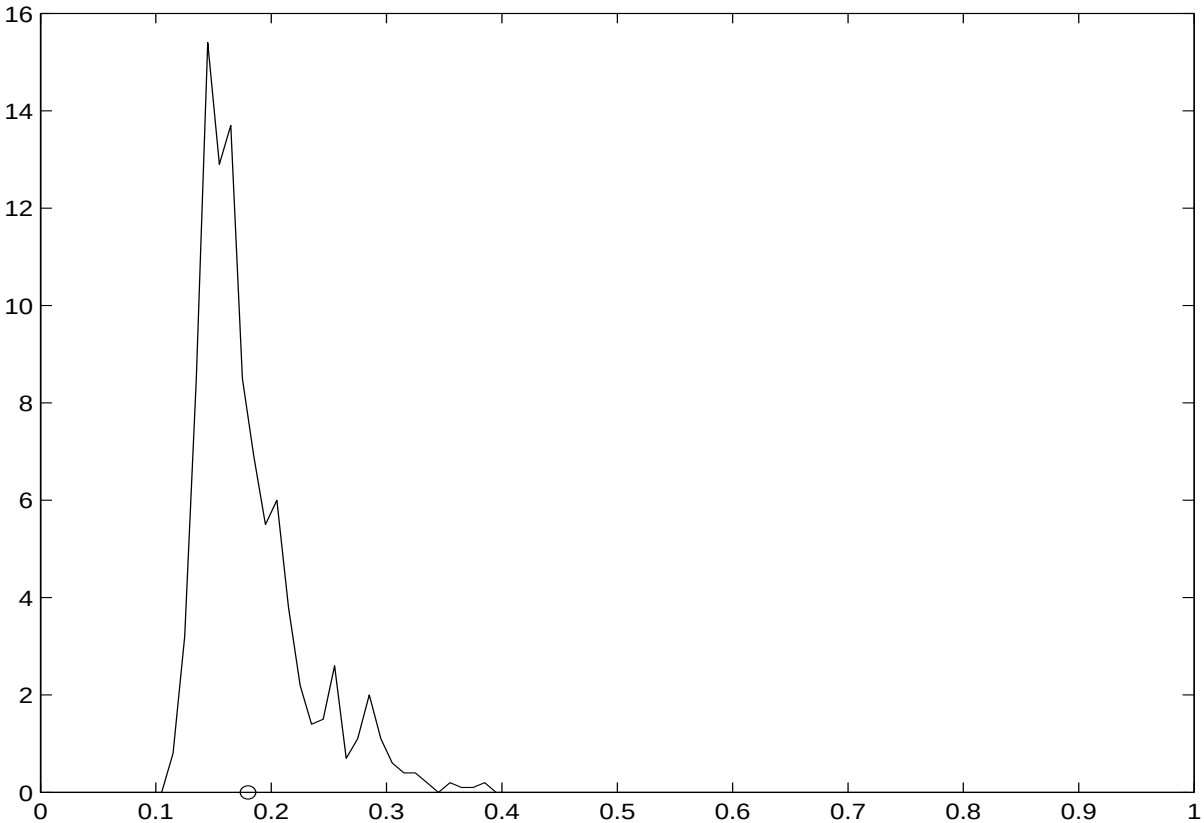
Error distribution: dataset size: 68



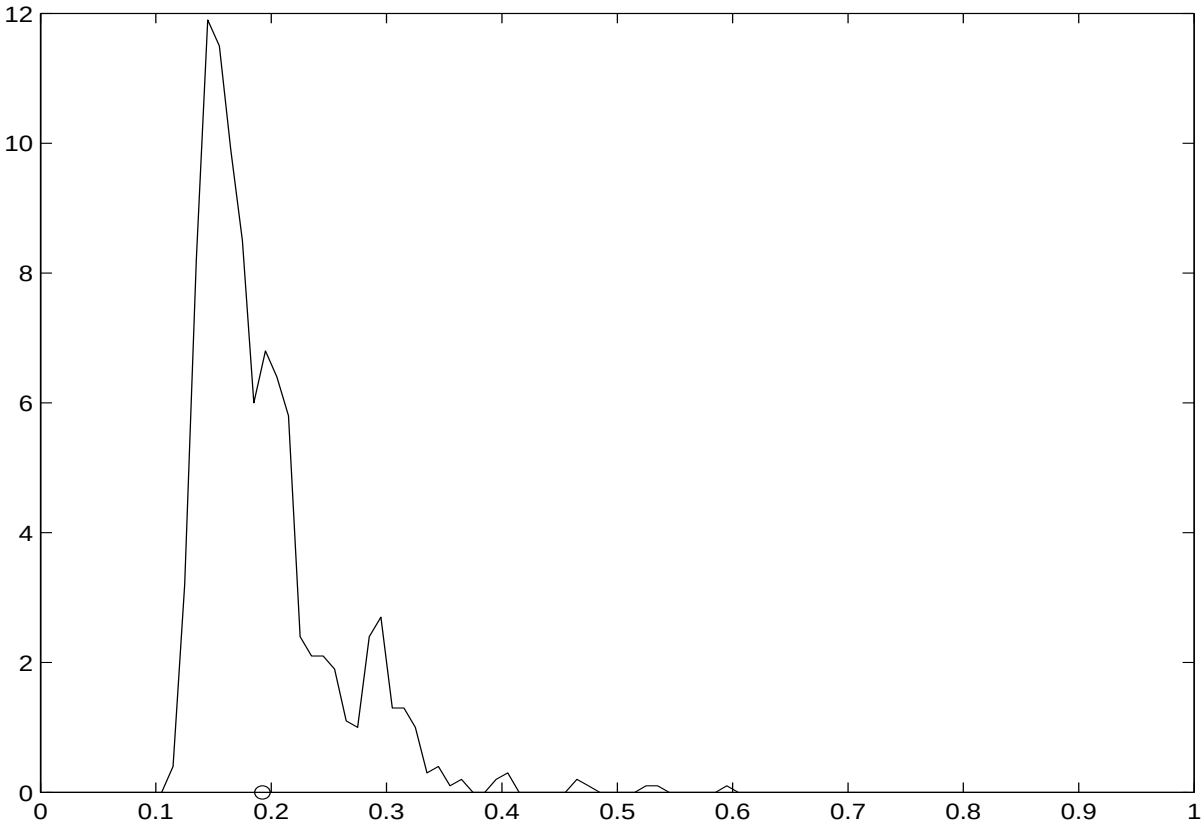
Error distribution: dataset size: 34



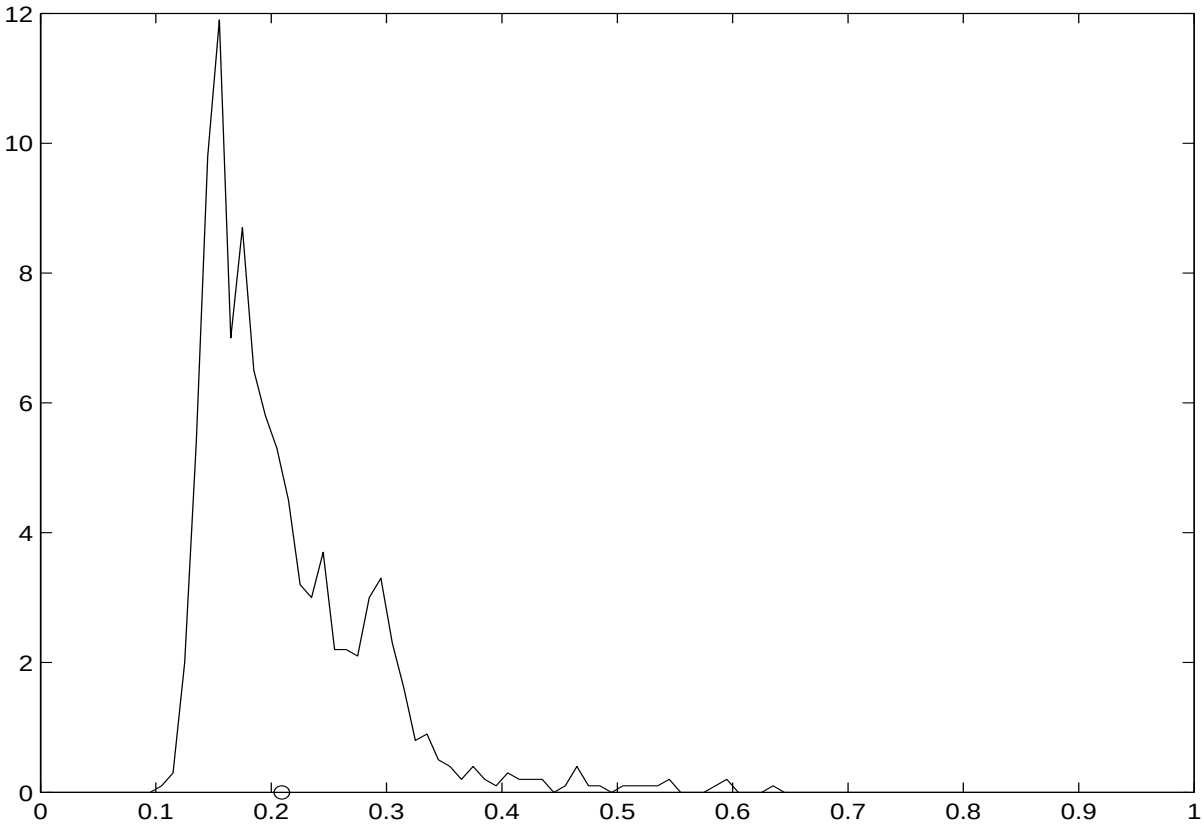
Error distribution: dataset size: 27



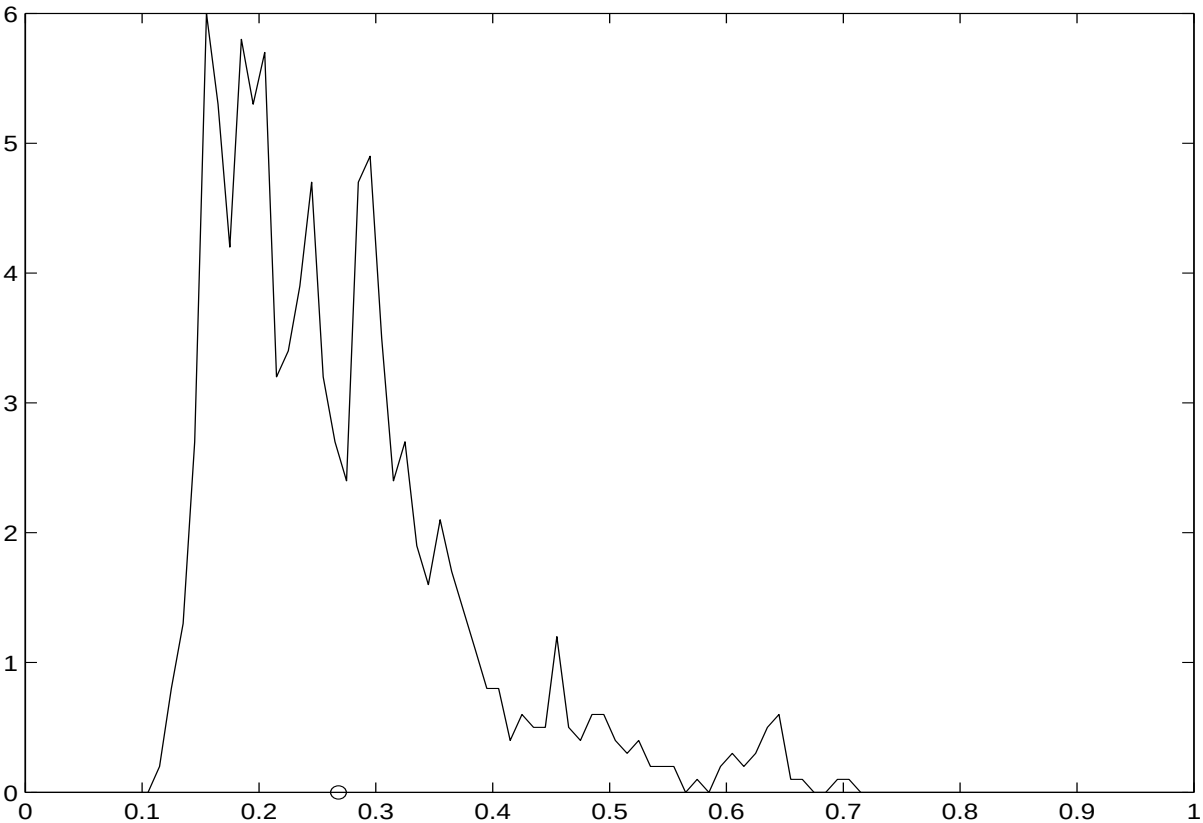
Error distribution: dataset size: 20



Error distribution: dataset size: 14



Error distribution: dataset size: 7



Bayes risk and consistency

- Traditional statistics has concentrated on analysing

$$\mathbb{E}_S\left[\epsilon(S, \mathcal{A}, \mathcal{F})\right].$$

- For example consistency of a classification algorithm $\mathcal{A}$ and function class $\mathcal{F}$ means

$$\lim_{m\to\infty} \mathbb{E}_S\left[\epsilon(S, \mathcal{A}, \mathcal{F})\right] = f_{\text{Bayes}},$$

where

$$f_{\text{Bayes}}(\mathbf{x}) = \begin{cases} 1 & \text{if } P(\mathbf{x}, 1) > P(\mathbf{x}, 0), \\ 0 & \text{otherwise.} \end{cases}$$

is the function with the lowest possible risk, often referred to as the Bayes risk.

Expected versus confident bounds

- For a finite sample the generalisation $\epsilon(S, \mathcal{A}, \mathcal{F})$ has a distribution depending on the algorithm, function class and sample size m.

- Traditional statistics as indicated above has concentrated on the mean of this distribution – but this quantity can be misleading, eg for low fold cross-validation.

Expected versus confident bounds cont.

- Statistical learning theory has preferred to analyse the tail of the distribution, finding a bound which holds with high probability.

- This looks like a statistical test – significant at a 1% confidence means that the chances of the conclusion not being true are less than 1% over random samples of that size.

- This is also the source of the acronym PAC: probably approximately correct, the 'confidence' parameter δ is the probability that we have been misled by the training set.

Probability of being misled in classification

- Aim to cover a number of key techniques of SLT. Basic approach is usually to bound the probability of being misled and set this equal to δ.

- What is the chance of being misled by a single bad function f, i.e. training error $\mathrm{err}_S(f) = 0$, while true error is bad $\mathrm{err}(f) > \epsilon$?

$$\begin{aligned} P_S\left\{\mathrm{err}_S(f) = 0, \mathrm{err}(f) > \epsilon\right\} &= (1 - \mathrm{err}(f))^m \\ &\leq (1-\epsilon)^m \\ &\leq \exp(-\epsilon m). \end{aligned}$$

  so that choosing $\epsilon = \ln(1/t)/m$ ensures probability less than t.

Finite or Countable function classes

If we now consider a function class

$$\mathcal{F} = \{f_1, f_2, \ldots, f_n, \ldots\}$$

and make the probability of being misled by f_n less than $q_n\delta$ with

$$\sum_{n=1}^{\infty} q_n \leq 1,$$

then the probability of being misled by one of the functions is bounded by

$$P_S\left\{\exists f_n : \mathrm{err}_S(f_n) = 0, \mathrm{err}(f_n) > \frac{1}{m}\ln\left(\frac{1}{q_n\delta}\right)\right\} \leq \delta.$$

This uses the so-called union bound – the probability of the union of a set of events is at most the sum of the individual probabilities.

Finite or Countable function classes result

- The bound translates into a theorem: given $\mathcal{F}$ and q, with probability at least $1-\delta$ over random m samples the generalisation error of a function $f_n \in \mathcal{F}$ with zero training error is bounded by

$$\mathrm{err}(f_n) \leq \frac{1}{m}\left(\ln\left(\frac{1}{q_n}\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

Some comments on the result

- We can think of the term $\ln\left(\frac{1}{q_n}\right)$ as the complexity / description length of the function f_n.

- Note that we must put a prior weight on the functions. If the functions are drawn at random according to a distribution p_n, the expected generalisation will be minimal if we choose our prior $q = p$.

- This is the starting point of the PAC-Bayes analysis.

What if uncountably many functions?

- We need a way to convert an infinite set to a finite one.
- Key idea is to replace measuring performance on a random test point with measuring on a second 'ghost' sample
- In this way the analysis is reduced to a finite set of examples and hence a finite set of classification functions.
- This step is often referred to as the 'double sample trick'

Double sample trick

The result has the following form:

$$P^m\{\mathbf{X} \in X^m : \exists h \in H : \mathrm{err}_{\mathbf{X}}(h) = 0, \mathrm{err}(h) \geq \epsilon\}$$
$$\leq 2P^{2m}\{\mathbf{XY} \in X^{2m} : \exists h \in H :$$
$$\mathrm{err}_{\mathbf{X}}(h) = 0, \mathrm{err}_{\mathbf{Y}}(h) \geq \epsilon/2\}$$

If we think of the first probability as being over $\mathbf{XY}$ the result concerns three events:

$$A(h) := \{\mathrm{err}_{\mathbf{X}}(h) = 0\}$$
$$B(h) := \{\mathrm{err}(h) \geq \epsilon\}$$
$$C(h) := \{\mathrm{err}_{\mathbf{Y}}(h) \geq \epsilon/2\}$$

Double sample trick II

It is clear that

$$\begin{aligned} P^{2m}(C(h)|A(h)\&B(h)) &= P^{2m}(C(h)|B(h)) \\ &> 0.5 \end{aligned}$$

for reasonable m by a binomial tail bound.

Double sample trick II

Hence, we have

$$
\begin{aligned}
P^{2m}\{\mathbf{XY} \in X^{2m} \quad &: \quad \exists h \in H : A(h)\&C(h)\} \geq \\
P^{2m}\{\mathbf{XY} \in X^{2m} \quad &: \quad \exists h \in H : A(h)\&B(h)\&C(h)\} = \\
P^{2m}\{\mathbf{XY} \in X^{2m} \quad &: \quad \exists h \in H : A(h)\&B(h)\} \\
& \qquad P(C(h)|A(h)\&B(h))
\end{aligned}
$$

It follows that

$$
\begin{aligned}
P^{m}\{\mathbf{X} \in X^{m} \quad &: \quad \exists h \in H : A(h)\&B(h)\} \leq \\
& 2P^{2m}\{\mathbf{XY} \in X^{2m} : \exists h \in H : A(h)\&C(h)\}
\end{aligned}
$$

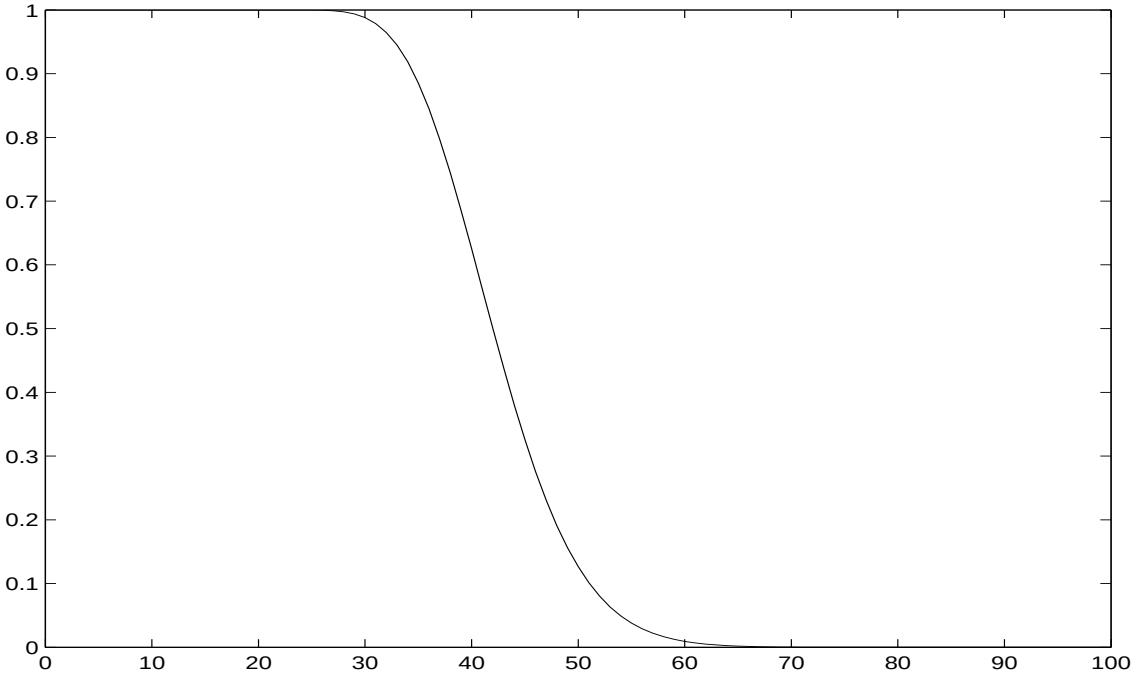
the required result.

How many functions on a finite sample?

- 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 – note that this cannot be larger than 2^m, i.e. $\log_2(B_H(m)) \leq m$
- For the statistics to work we want the number of functions to be much smaller than this as we will perform a union bound over this number.

Examining the growth function

Consider a plot of the ratio of the growth function $B_H(m)$ to 2^m for linear functions in a 20 dimensional space:



Vapnik Chervonenkis dimension

- The Vapnik-Chervonenkis dimension is the point at which the ratio stops being equal to 1:

$$\text{VCdim}(H) = \max\{m \quad : \quad \text{for some } \mathbf{x}_1, \ldots, \mathbf{x}_m, \text{ for all } b \in \{-1, 1\}^m, \exists h_b \in H, h_b(\mathbf{x}_i) = b_i\}$$

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

Sauer's Lemma

- Sauer's Lemma (also due to VC and Shelah):

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

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

- This is surprising as the rate of growth of $B_H(m)$ up to $m = d$ is exponential – but thereafter it is polynomial with exponent d.

- Furthermore, there is no assumption made about the functions being linear – this is a completely general result.

Basic Theorem of SLT

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

$$P^m\{\mathbf{X} \in X^m : \exists h \in H : \text{err}_{\mathbf{X}}(h) = 0, \text{err}(h) \geq \epsilon\}$$

$$\rightarrow \text{double sample trick} \rightarrow$$

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

$$\rightarrow \text{union bound} \rightarrow$$

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

Final ingredient is known as symmetrisation.

Symmetrisation

- Consider generating a $2m$ sample S. Since the points are generated independently the probability of generating the same set of points in a different order is the same.

- Consider a fixed set Σ of permutations and each time we generate a sample we randomly permute it with a uniformly chosen element of Σ – gives probability distribution P_{Σ}^{2m}

Symmetrisation cont.

- Any event has equal probability under P^{2m} and P_Σ^{2m}, so that

$$P^{2m}(A) = P_\Sigma^{2m}(A) = \mathbb{E}^{2m}\left[P_{\sigma\sim\Sigma}(A)\right]$$

- Consider particular choice of Σ the permutations that swap/leave unchanged corresponding elements of the two samples $\mathbf{X}$ and $\mathbf{Y}$ – 2^m such permutations.

Completion of the proof

$$
\begin{aligned}
P^{2m}\{\mathbf{XY} \in X^{2m} : \mathrm{err}_{\mathbf{X}}(h) = 0, \mathrm{err}_{\mathbf{Y}}(h) \geq \epsilon/2\} \\
&\leq \mathbb{E}^{2m}\left[P_{\sigma\sim\Sigma}\{\mathrm{err}_{\mathbf{X}}(h) = 0, \mathrm{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \text{ for } \sigma(\mathbf{XY})\}\right] \\
&\leq \mathbb{E}^{2m}\left[2^{-\epsilon m/2}\right] \\
&= 2^{-\epsilon m/2}
\end{aligned}
$$

- Setting the right hand side equal to $\delta/(2B_H(2m))$ and inverting gives the bound on ϵ.

Final result

- Assembling the ingredients gives the result: with probability at least $1-\delta$ of random m samples the generalisation error of a function $h \in H$ chosen from a class H with VC dimension d with zero training error is bounded by

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

- Note that we can think of d as the complexity / capacity of the function class H.

- The bound does not distinguish between functions in H.

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

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

Non-zero training error

- Very similar results can be obtained for non-zero training error.
- The main difference is the introduction of a square root to give a bound of the form

$$\epsilon(m, H, k, \delta) = k + O\left(\sqrt{\frac{d}{m}\log\frac{2em}{d}} + \sqrt{\frac{1}{m}\log\frac{2}{\delta}}\right)$$

  for k training errors, which is significantly worse than in the zero training error case.
- PAC-Bayes bounds now interpolate between these two.

Structural Risk Minimisation

- The way to differentiate between functions using the VC result is to create a hierarchy of classes:

$$H_1 \subseteq H_2 \subseteq \cdots \subseteq H_d \subseteq \cdots \subseteq H_K.$$

  of increasing complexity/VC dimension.

- Can now find the function in each class with minimum empirical error k_d and choose between the classes by minimising over the choice of d:

$$\epsilon(m, d, \delta) = k_d + O\left(\sqrt{\frac{d}{m}\log\frac{2em}{d}} + \sqrt{\frac{1}{m}\log\frac{2K}{\delta}}\right)$$

  which bounds the generalisation by a further application of the union bound over the K classes.

Criticisms of PAC Theory

- The theory is certainly valid and the lower bounds indicate that it is not too far out – so can't criticise as stands
- Criticism is that it doesn't accord with experience of those applying learning.
- Mismatch between theory and practice.
- For example

Support Vector Machines (SVM)

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

1. kernel trick means we can work in an infinite dimensional feature space ($\Rightarrow$ infinite VC dimension) so that VC result does not apply:

2. and YET very impressive performance

Support Vector Machines cont.

- SVM seeks linear function in a feature space defined implicitly via a kernel κ:

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

- For example the 1-norm SVM seeks $\mathbf{w}$ to solve

$$\begin{array}{ll} \min_{\mathbf{w},b,\gamma,\xi} & \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} & y_i \left(\langle \mathbf{w}, \phi\left(\mathbf{x}_i\right) \rangle + b \right) \geq 1 - \xi_i, \, \xi_i \geq 0, \\ & i = 1, \ldots, m. \end{array}$$

Margin in SVMs

- Intuition behind SVMs is that maximising the margin makes it possible to obtain good generalisation despite the high VC dimension

- The lower bound implies that we must be taking advantage of a benign distribution, since we know that in the worst case generalisation will be bad.

- Structural Risk Minimisation cannot be applied to a hierarchy determined by the margin of different classifiers, since the margin is not known until we see the data, while the SRM hierarchy must be chosen in advance.

Margin in SVMs cont

- Hence, we require a theory that can give bounds that are sensitive to serendipitous distributions – in particular we conjecture that the margin is an indication of such 'luckiness'.

- The proof approach will rely on using real-valued function classes. The margin gives an indication of the accuracy with which we need to approximate the functions when applying the statistics.

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 $\|f - u\|_d < \gamma$.

Covering Numbers cont.

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

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

where $\ell_\infty^{\mathbf{x}}$ gives the distance between two functions as the maximum difference between their outputs on the sample.

Covering numbers for linear functions

- Consider the set of functions:

$$\left\{ \mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x}) \rangle \quad : \quad \mathbf{w} = \frac{1}{Z} \sum_{i=1}^{m} z_i \phi(\mathbf{x}_i), z_i \in \mathbb{Z}, \right.$$
$$\left. Z = \sum_{i=1}^{m} z_i \neq 0, \sum_{i=1}^{m} |z_i| \leq \frac{8R^2}{\gamma^2} \right\}$$

- This is an explicit cover that approximates the output of any norm 1 linear function on the sample

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$$

to within $\gamma/2$ on the sample.

Covering numbers for linear functions

- We convert the $\gamma/2$ approximation on the sample problem into a classification problem, which is solvable with a margin of $\gamma/2$.

- It follows that if we use the perceptron algorithm to find a classifier, we will find a function satisfying the $\gamma/2$ approximation with just $8R^2/\gamma^2$ updates.

- This gives a sparse dual representation of the function. The covering is chosen as the set of functions with small sparse dual representations.

- Gives a bound on the size of the cover

$$\log_2 \mathcal{N}^{2m}(\gamma/2, \mathcal{F}) \leq k \log_2 \frac{e(2m+k-1)}{k}, \text{ where } k = \frac{8R^2}{\gamma^2}.$$

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:

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

Second statistical result cont.

- 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.*

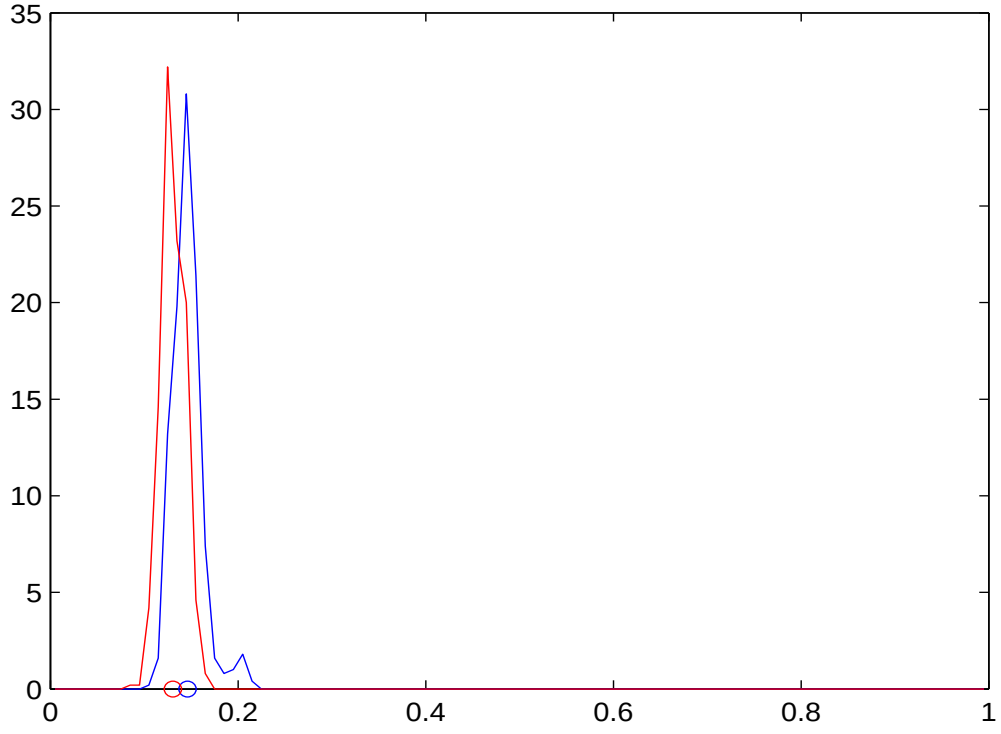
Generalization of SVMs

- Since SVMs are linear functions, can apply linear function bound.
- Hence for distribution with support in ball of radius R, (eg Gaussian Kernels $R = 1$) and margin γ, have bound:

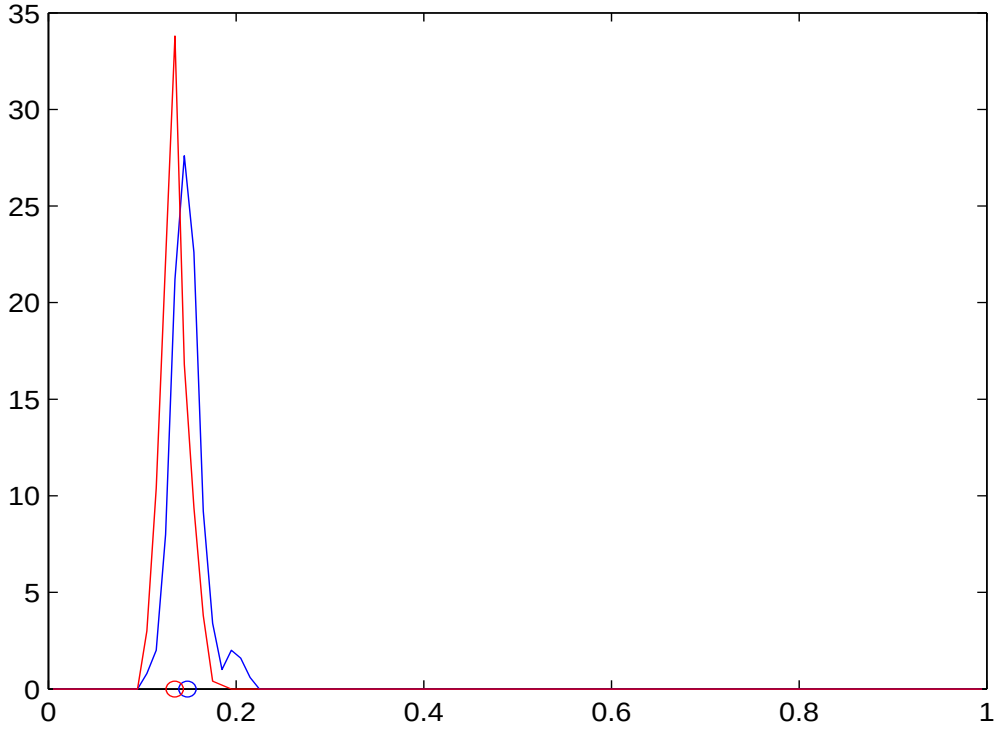
$$\epsilon(m, \mathcal{L}, \delta, \gamma) = \frac{2}{m}\left(k \log_2 \frac{e(2m + k - 1)}{k} + \log_2 \frac{m}{\delta}\right)$$

where $k = \frac{8R^2}{\gamma^2}$.

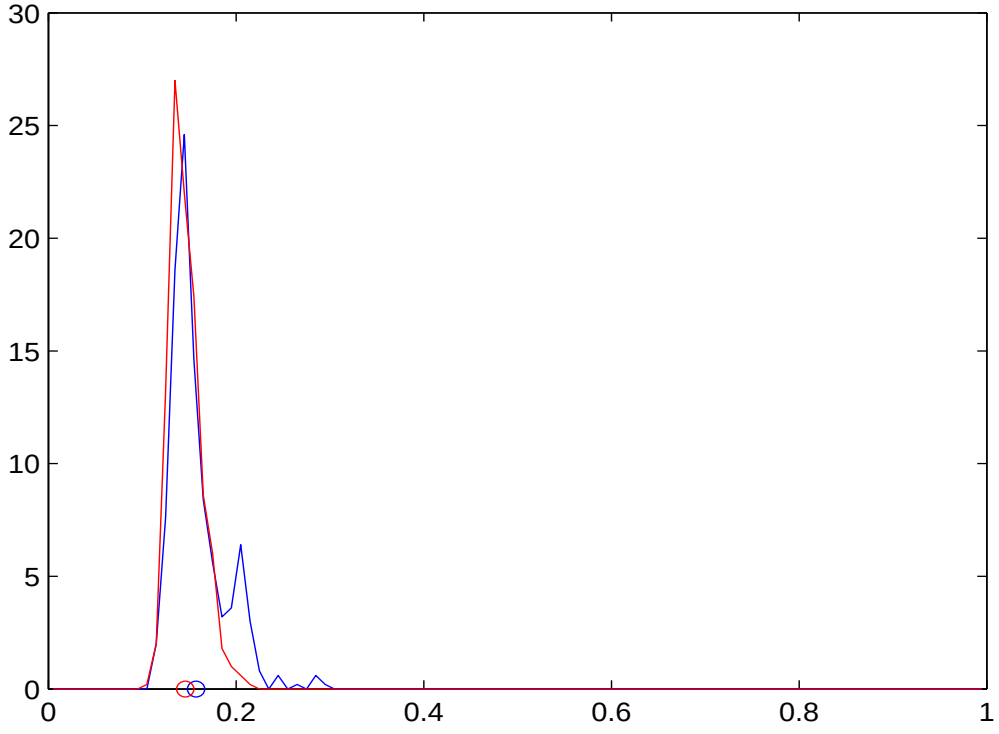
Error distribution: dataset size: 205



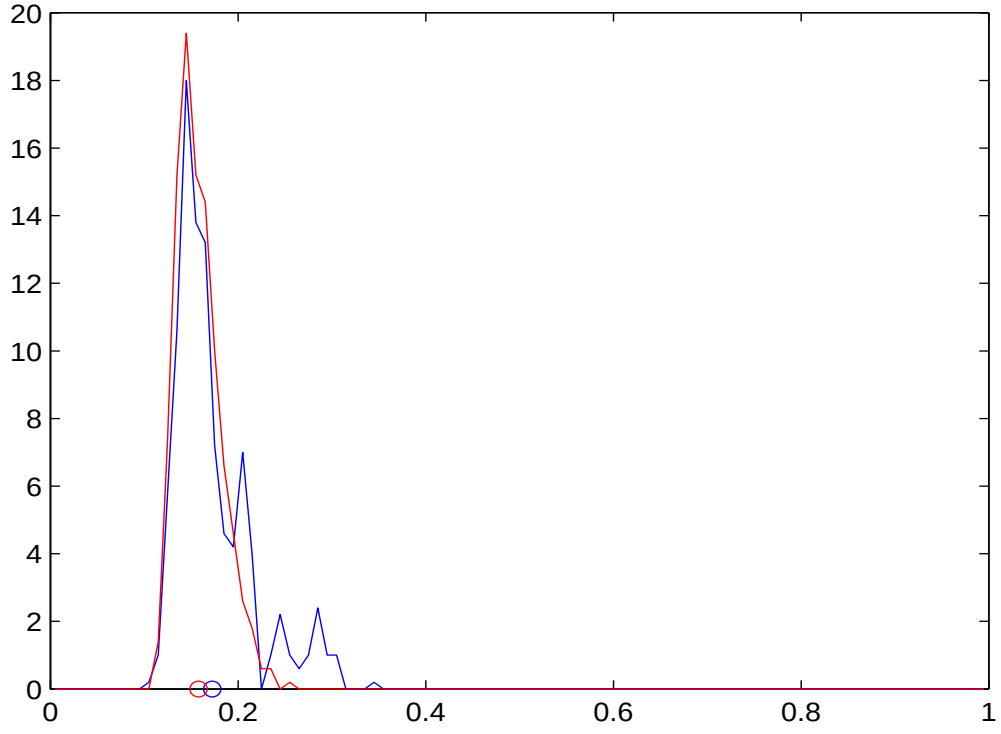
Error distribution: dataset size: 137



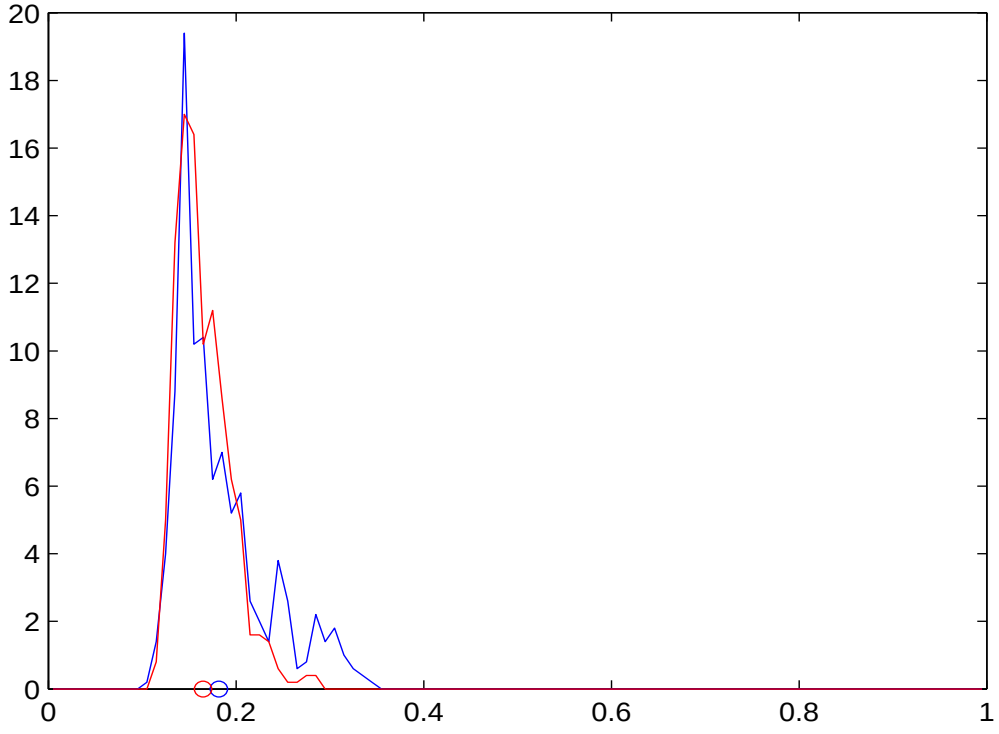
Error distribution: dataset size: 68



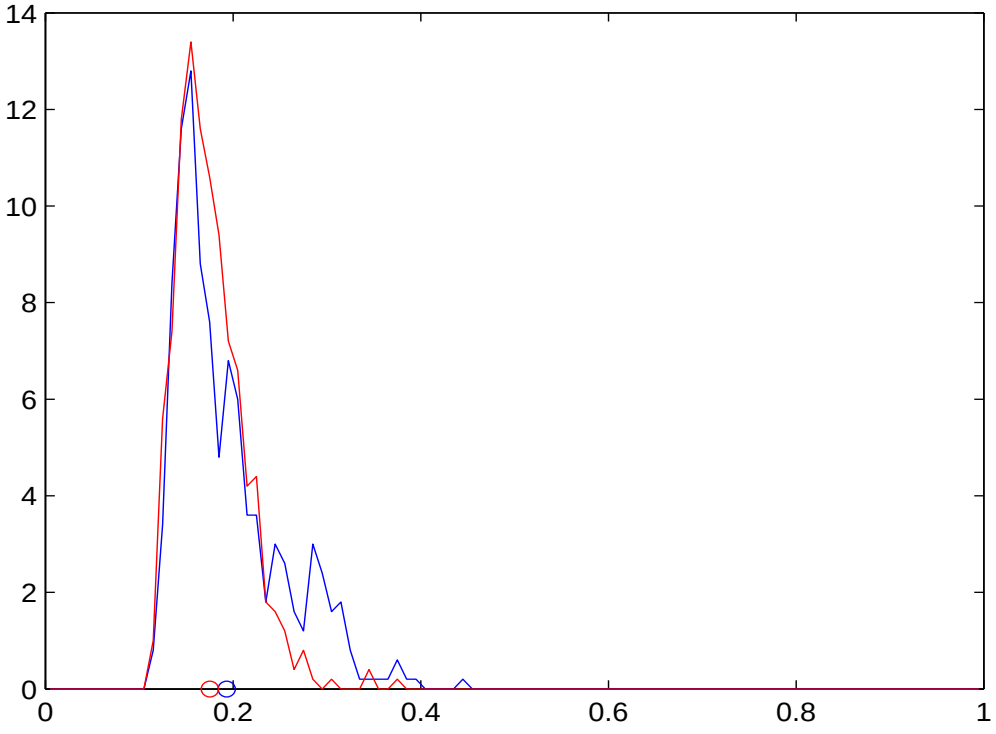
Error distribution: dataset size: 34



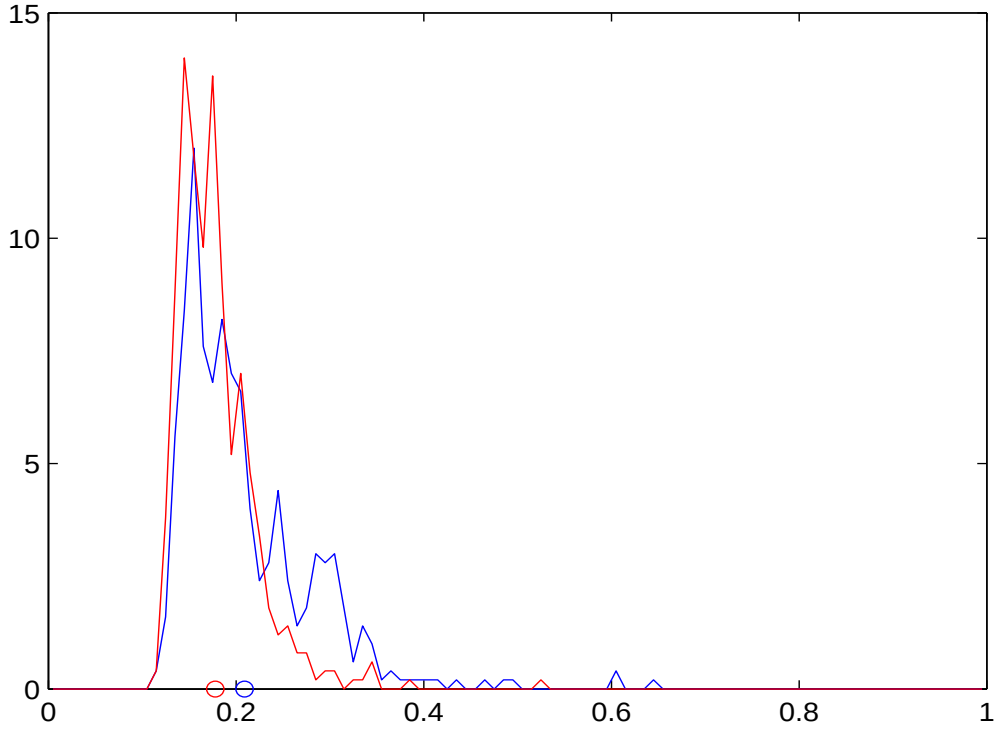
Error distribution: dataset size: 27



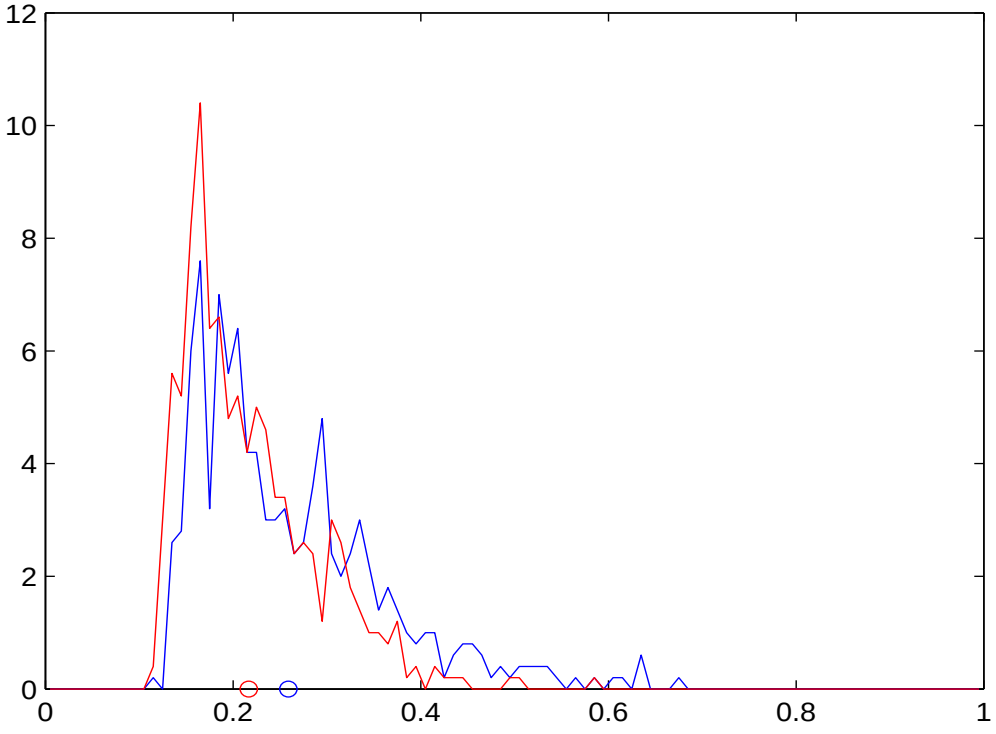
Error distribution: dataset size: 20



Error distribution: dataset size: 14



Error distribution: dataset size: 7



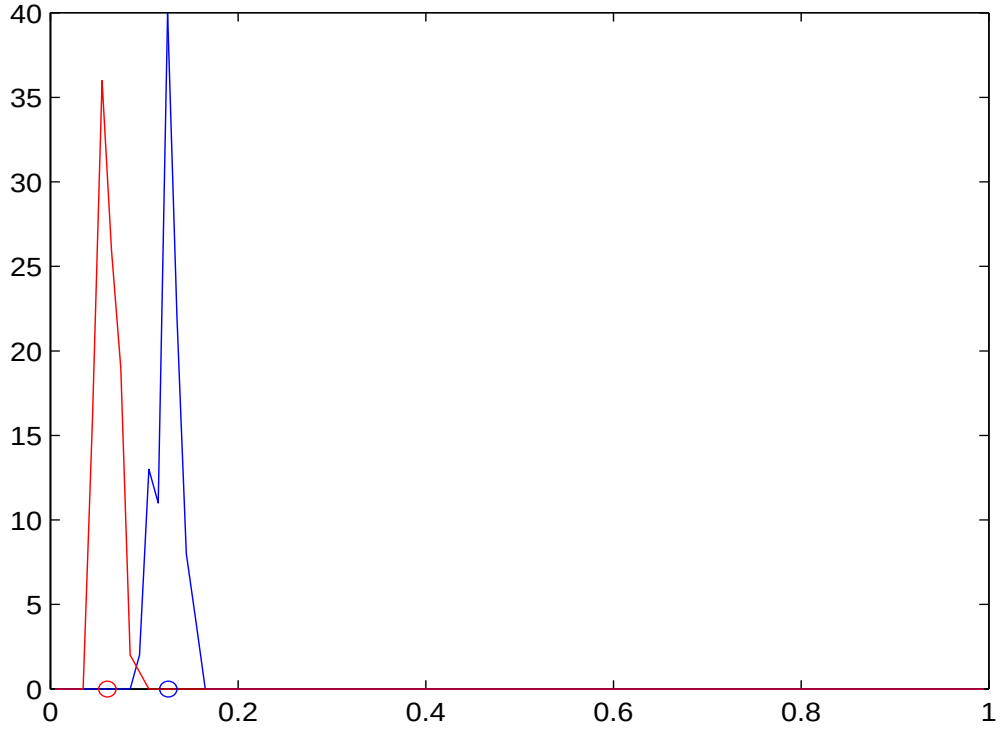
Using a kernel

- Can consider much higher dimensional spaces using the kernel trick

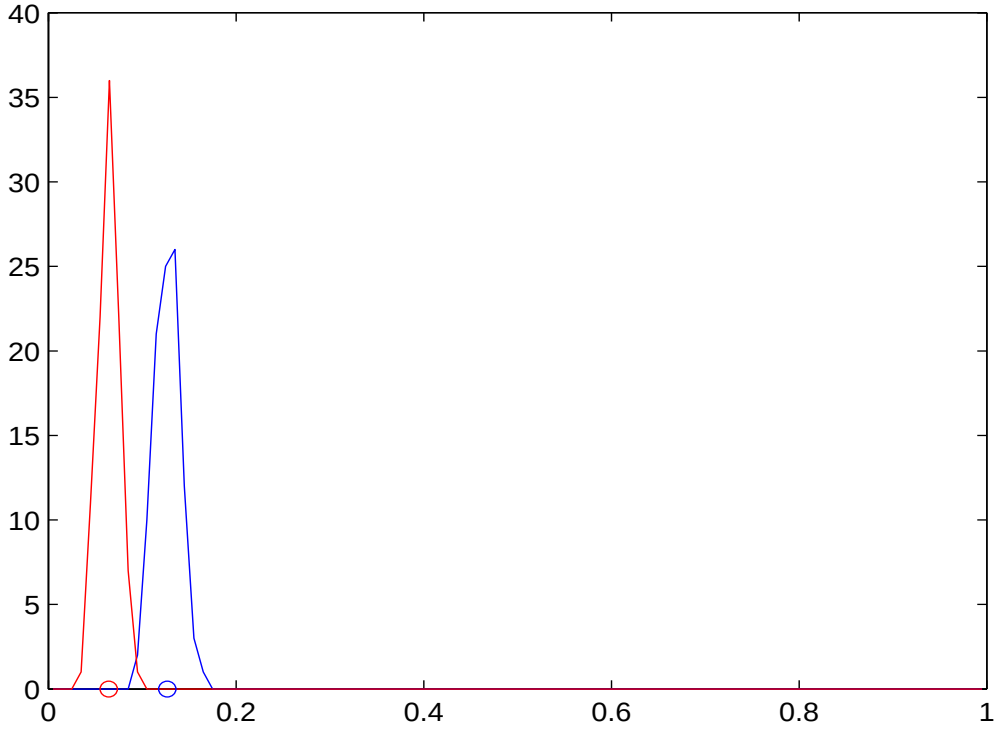
- Can even work in infinite dimensional spaces, eg using the Gaussian kernel:

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right)$$

Error distribution: dataset size: 342



Error distribution: dataset size: 273



Conclusions

- Outline of philosophy and approach of SLT
- Central result of SLT
- Touched on covering number bounds for margin based analysis
- Application to analyse SVM learning

STRUCTURE

PART A

1. General Statistical Considerations
2. Basic PAC Ideas and proofs
3. Real-valued Function Classes and the Margin

PART B

1. Rademacher complexity and Main Theory
2. Applications to classification
3. Conclusions

PART B

Concentration inequalities

- Statistical Learning is concerned with the reliability or stability of inferences made from a random sample.

- Random variables with this property have been a subject of ongoing interest to probabilists and statisticians.

Concentration inequalities cont.

- As an example consider the mean of a sample of m 1-dimensional random variables $X_1, \ldots, X_m$:

$$S_m = \frac{1}{m} \sum_{i=1}^{m} X_i.$$

- Hoeffding's inequality states that if $X_i \in [a_i, b_i]$

$$P\{|S_m - \mathbb{E}[S_m]| \geq \epsilon\} \leq 2 \exp\left(- \frac{2m^2\epsilon^2}{\sum_{i=1}^{m} (b_i - a_i)^2} \right)$$

  Note how the probability falls off exponentially with the distance from the mean and with the number of variables.

Concentration for SLT

- We are now going to look at deriving SLT results from concentration inequalities.

- Perhaps the best known form is due to McDiarmid (although he was actually re-presenting previously derived results):

McDiarmid's inequality

Theorem 1. *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i,$$

for $1 \leq i \leq n$. Then for all $\epsilon > 0$,

$$P\left\{f\left(X_1, \ldots, X_n\right) - \mathbb{E}f\left(X_1, \ldots, X_n\right) \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

- Hoeffding is a special case when $f(x_1, \ldots, x_n) = S_n$

Using McDiarmid

- By setting the right hand side equal to δ, we can always invert McDiarmid to get a high confidence bound: with probability at least $1 - \delta$

$$f(X_1, \ldots, X_n) < \mathbb{E} f(X_1, \ldots, X_n) + \sqrt{\frac{\sum_{i=1}^n c_i^2}{2} \log \frac{1}{\delta}}$$

- If $c_i = c/n$ for each i this reduces to

$$f(X_1, \ldots, X_n) < \mathbb{E} f(X_1, \ldots, X_n) + \sqrt{\frac{c^2}{2n} \log \frac{1}{\delta}}$$

Rademacher complexity

- Rademacher complexity is a new way of measuring the complexity of a function class. It arises naturally if we rerun the proof using the double sample trick and symmetrisation but look at what is actually needed to continue the proof:

Rademacher proof beginnings

For a fixed $f \in \mathcal{F}$ we have

$$\mathbb{E}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right).$$

where $\mathcal{F}$ is a class of functions mapping from Z to $[0, 1]$ and $\hat{\mathbb{E}}$ denotes the sample average.

We must bound the size of the second term. First apply McDiarmid's inequality to obtain ($c_i = 1/m$ for all i) with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right) \leq \mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

Deriving double sample result

- We can now move to the ghost sample by simply observing that $\mathbb{E}[h] = \mathbb{E}_{\tilde{S}}\left[\hat{\mathbb{E}}[h]\right]$:

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right)\right] =$$

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^{m} h(\tilde{\mathbf{z}}_i) - \frac{1}{m}\sum_{i=1}^{m} h(\mathbf{z}_i)\,\middle|\, S\right]\right]$$

Deriving double sample result cont.

Since the sup of an expectation is less than or equal to the expectation of the sup (we can make the choice to optimise for each $\tilde{S}$) we have

$$\mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] \leq$$

$$\mathbb{E}_S \mathbb{E}_{\tilde{S}} \left[\sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i) \right) \right]$$

Adding symmetrisation

Here symmetrisation is again just swapping corresponding elements – but we can write this as multiplication by a variable σ_i which takes values ± 1 with equal probability:

$$\mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] \leq$$

$$\begin{aligned} &\leq \mathbb{E}_{\sigma S \tilde{S}} \left[\sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i) \right) \right] \\ &\leq 2\mathbb{E}_{S\sigma} \left[\sup_{h \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\mathbf{z}_i) \right| \right] \\ &= R_m(\mathcal{F}), \end{aligned}$$

Rademacher complexity

where

$$R_m(\mathcal{F}) = \mathbb{E}_{S\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{m}\sum_{i=1}^{m}\sigma_i f\left(\mathbf{z}_i\right)\right|\right].$$

is known as the Rademacher complexity of the function class $\mathcal{F}$.

- Rademacher complexity is the expected value of the maximal correlation with random noise – a very natural measure of capacity.

- Note that the Rademacher complexity is distribution dependent since it involves an expectation over the choice of sample – this might seem hard to compute.

Main Rademacher theorem

Putting the pieces together gives the main theorem of Rademacher complexity: with probability at least $1-\delta$ over random samples S of size m, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

Empirical Rademacher theorem

- Since the empirical Rademacher complexity

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i f\left(\mathbf{z}_i\right) \right| \middle| \mathbf{z}_1, \ldots, \mathbf{z}_m \right]$$

  is concentrated, we can make a further application of McDiarmid to obtain with probability at least $1 - \delta$

$$\mathbb{E}_{\mathcal{D}}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

Relation to VC theorem

- For H a class of ± 1 valued functions with VC dimension d, we can upper bound $\hat{R}_m(H)$ using Hoeffding's inequality to upper bound

$$P\left\{\left|\sum_{i=1}^{m}\sigma_i f(\mathbf{x}_i)\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{\epsilon^2}{2m}\right)$$

  for a fixed function f, since the expected value of the sum is 0, and the maximum change by replacing a σ_i is 2.

Relation to VC theorem cont.

- By Sauer's lemma there are at most $(em/d)^d$ we can bound the probability that the sum is bounded by ϵ for all functions by

$$\left(\frac{em}{d}\right)^d 2\exp\left(-\frac{\epsilon^2}{2m}\right) =: \delta.$$

- Taking $\delta = m^{-1}$ and solving for ϵ gives

$$\epsilon = \sqrt{2md\ln\frac{em}{d} + 2m\ln(2m)}$$

Rademacher bound for VC class

- Hence we can bound

$$\begin{aligned}\hat{R}_m(H) &\leq \frac{2}{m}\left(\delta m + \epsilon(1-\delta)\right) \\ &\leq \frac{2}{m} + \sqrt{\frac{8(d\ln(em/d) + \ln(2m))}{m}}\end{aligned}$$

- This is equivalent to the PAC bound with non-zero loss, except that we could have used the growth function or VC dimension measured on the sample rather than the sup over the whole input space.

Application to large margin classification

- Rademacher complexity comes into its own for Boosting and SVMs.

Application to Boosting

- We can view Boosting as seeking a function from the class (H is the set of weak learners)

$$\left\{ \sum_{h \in H} a_h h(\mathbf{x}) : \sum_{h \in H} a_h \leq B \right\} = \mathrm{conv}_B(H)$$

  by minimising some function of the margin distribution.

- Adaboost corresponds to optimising an exponential function of the margin over this set of functions.

- We will see how to include the margin in the analysis later, but concentrate on computing the Rademacher complexity for now.

Rademacher complexity of convex hulls

Rademacher complexity has a very nice property for convex hull classes:

$$
\begin{aligned}
\hat{R}_m(\mathrm{conv}_B(H)) &= \frac{2}{m}\mathbb{E}_\sigma\left[\sup_{h_j\in H,\sum_j a_j\le B}\left|\sum_{i=1}^m \sigma_i \sum_j a_j h_j(\mathbf{x}_i)\right|\right] \\
&\le \frac{2}{m}\mathbb{E}_\sigma\left[\sup_{h_j\in H,\sum_j a_j\le B}\sum_j a_j\left|\sum_{i=1}^m \sigma_i h_j(\mathbf{x}_i)\right|\right] \\
&\le \frac{2}{m}\mathbb{E}_\sigma\left[\sup_{h_j\in H} B\left|\sum_{i=1}^m \sigma_i h_j(\mathbf{x}_i)\right|\right] \\
&\le B\hat{R}_m(H).
\end{aligned}
$$

Rademacher complexity of convex hulls cont.

- Hence, we can move to the convex hull without incurring any complexity penalty for $B = 1$!

Rademacher complexity for SVMs

- The Rademacher complexity of a class of linear functions with bounded 2-norm:

$$\left\{ \mathbf{x} \to \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) : \alpha' \mathbf{K} \alpha \leq B^2 \right\} \subseteq$$

$$\subseteq \{ \mathbf{x} \to \langle \mathbf{w}, \phi(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq B \}$$

$$= \mathcal{F}_B,$$

  where we assume a kernel defined feature space with

$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \kappa(\mathbf{x}, \mathbf{z}).$$

Rademacher complexity of $\mathcal{F}_B$

$$
\begin{aligned}
\hat{R}_m(\mathcal{F}_B) &= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}_B} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i f\left(\mathbf{x}_i\right) \right| \right] \\
&= \mathbb{E}_\sigma \left[\sup_{\|\mathbf{w}\| \leq B} \left| \left\langle \mathbf{w}, \frac{2}{m} \sum_{i=1}^{m} \sigma_i \phi\left(\mathbf{x}_i\right) \right\rangle \right| \right] \\
&\leq \frac{2B}{m} \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{m} \sigma_i \phi(\mathbf{x}_i) \right\| \right] \\
&= \frac{2B}{m} \mathbb{E}_\sigma \left[\left(\left\langle \sum_{i=1}^{m} \sigma_i \phi(\mathbf{x}_i), \sum_{j=1}^{m} \sigma_j \phi(\mathbf{x}_j) \right\rangle \right)^{1/2} \right]
\end{aligned}
$$

$$
\leq \frac{2B}{m} \left(\mathbb{E}_\sigma \left[\sum_{i,j=1}^{m} \sigma_i \sigma_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \right] \right)^{1/2} = \frac{2B}{m} \sqrt{\sum_{i=1}^{m} \kappa(\mathbf{x}_i, \mathbf{x}_i)}
$$

Applying to 1-norm SVMs

We take the following formulation of the 1-norm SVM:

$$\begin{array}{ll} \min_{\mathbf{w},b,\gamma,\xi} & -\gamma + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} & y_i\left(\langle \mathbf{w}, \phi\left(\mathbf{x}_i\right)\rangle + b\right) \geq \gamma - \xi_i,\ \xi_i \geq 0, \\ & i = 1, \ldots, m, \text{ and } \|\mathbf{w}\|^2 = 1. \end{array} \tag{1}$$

Note that

$$\xi_i = \left(\gamma - y_i g(\mathbf{x}_i)\right)_+,$$

where $g(\cdot) = \langle \mathbf{w}, \phi(\cdot)\rangle + b$.

- The first step is to introduce a loss function which upper bounds the discrete loss

$$P(y \neq \operatorname{sgn}(g(\mathbf{x}))) = \mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x}))\right],$$

where $\mathcal{H}$ is the Heaviside function.

Applying the Rademacher theorem

- Consider the loss function $\mathcal{A} : \mathbb{R} \to [0, 1]$, given by

$$\mathcal{A}(a) = \begin{cases} 1, & \text{if } a > 0; \\ 1 + a/\gamma, & \text{if } -\gamma \leq a \leq 0; \\ 0, & \text{otherwise.} \end{cases}$$

- By the Rademacher Theorem and since the loss function $\mathcal{A} - 1$ dominates $\mathcal{H} - 1$, we have that

$$\begin{aligned} \mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x})) - 1\right] &\leq \mathbb{E}\left[\mathcal{A}(-yg(\mathbf{x})) - 1\right] \\ &\leq \hat{\mathbb{E}}\left[\mathcal{A}(-yg(\mathbf{x})) - 1\right] + \\ &\quad \hat{R}_m((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned}$$

Empirical loss and slack variables

- But the function $\mathcal{A}(-y_i g(\mathbf{x}_i)) \leq \xi_i/\gamma$, for $i = 1, \ldots, \ell$, and so

$$\mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x}))\right] \leq \frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \hat{R}_m((\mathcal{A}-1)\circ\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

- The final missing ingredient to complete the bound is to bound $\hat{R}_m((\mathcal{A}-1)\circ\mathcal{F})$ in terms of $\hat{R}_m(\mathcal{F})$.

- This will require a more detailed look at Rademacher complexity.

Rademacher complexity bounds

- First simple observation:

$$\text{for } a \in \mathbb{R}, \quad \hat{R}_m(a\mathcal{F}) = |a|\hat{R}_m(\mathcal{F}),$$

since af is the function achieving the sup for some σ for $a\mathcal{F}$ iff f achieves the sup for $\mathcal{F}$.

- We are interested in bounding RC $\hat{R}_m(\mathcal{L} \circ \mathcal{F}) \leq 2L\hat{R}_m(\mathcal{F})$ for class $\mathcal{L} \circ \mathcal{F} = \{\mathcal{L} \circ f : f \in \mathcal{F}\}$, where $\mathcal{L}$ satisfies, $\mathcal{L}(0) = 0$ and

$$|\mathcal{L}(a) - \mathcal{L}(b)| \leq L|a - b|,$$

i.e. $\mathcal{L}$ is a Lipschitz function with constant $L > 0$.

Rademacher complexity bounds cont.

- In our case $\mathfrak{L} = \mathcal{A} - 1$ and $L = 1/\gamma$.
- By above it is sufficient to prove for case $L = 1$ only, since then

$$\hat{R}_m(\mathfrak{L} \circ \mathcal{F}) = L\hat{R}_m((\mathfrak{L}/L) \circ \mathcal{F}) \leq 2L\hat{R}_m(\mathcal{F})$$

Proof for contraction ($L = 1$)

Want to get rid of absolute value in RC:

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i f\left(\mathbf{z}_i\right) \right| \right]$$

so consider defining

$$\mathcal{L}^+ = \mathcal{L}, \quad \mathcal{L}^-(a) = -\mathcal{L}(-a).$$

Proof for contraction ($L = 1$)

Assume $\mathcal{F}$ is closed under negation. Now if for some σ sup achieved with f such that

$$\sum_{i=1}^{m} \sigma_i \mathcal{L}\left(f\left(\mathbf{z}_i\right)\right) < 0,$$

then

$$\left|\sum_{i=1}^{m} \sigma_i \mathcal{L}\left(f\left(\mathbf{z}_i\right)\right)\right| = \sum_{i=1}^{m} \sigma_i \mathcal{L}^{-}\left(-f\left(\mathbf{z}_i\right)\right),$$

Contraction proof cont.

- and so

$$\hat{R}_m(\mathcal{L} \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, \mathcal{N} \in \{\mathcal{L}^+, \mathcal{L}^-\}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \mathcal{N}\left(f\left(\mathbf{z}_i\right)\right) \right]$$

- if we further assume $0 \in \mathcal{F}$ we have

$$\begin{aligned} \hat{R}_m(\mathcal{L} \circ \mathcal{F}) &= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \mathcal{L}^+\left(f\left(\mathbf{z}_i\right)\right) \right] \\ &\quad + \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \mathcal{L}^-\left(f\left(\mathbf{z}_i\right)\right) \right] \end{aligned}$$

Contraction proof cont.

- Hence if we show the result without the factor of 2 for the complexity without absolute values the desired result will follow, since for classes closed under negation we have

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f\left(\mathbf{z}_i\right) \right].$$

Contraction proof cont.

- We show

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i \mathcal{L}(f(\mathbf{x}_i)) \right] \leq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sigma_1 f(\mathbf{x}_1) + \sum_{i=2}^{m} \sigma_i \mathcal{L}(f(\mathbf{x}_i)) \right]$$

and apply induction to obtain the full result.

Contraction proof cont.

- We take the permutations in pairs:

$$(1, \sigma_2, \ldots, \sigma_m) \text{ and } (-1, \sigma_2, \ldots, \sigma_m)$$

The result will follow if we show that for all $f, g \in \mathcal{F}$ we can find $f', g' \in \mathcal{F}$ such that ($f_i = f(\mathbf{x}_i)$ etc.)

$$\mathcal{L}(f_1) + \sum_{i=2}^{m} \sigma_i \mathcal{L}(f_i)) - \mathcal{L}(g_1) + \sum_{i=2}^{m} \sigma_i \mathcal{L}(g_i)$$

$$\leq f_1' + \sum_{i=2}^{m} \sigma_i \mathcal{L}(f_i')) - g_1' + \sum_{i=2}^{m} \sigma_i \mathcal{L}(g_i').$$

Contraction proof end

- If $f_1 \geq g_1$ take $f' = f$ and $g' = g$ to reduce to showing

$$\mathcal{L}(f_1) - \mathcal{L}(g_1) \leq f_1 - g_1 = |f_1 - g_1|$$

  which follows since $|\mathcal{L}(f_1) - \mathcal{L}(g_1)| \leq |f_1 - g_1|$.

- Otherwise $f_1 < g_1$ and we take $f' = g$ and $g' = f$ to reduce to showing

$$\mathcal{L}(f_1) - \mathcal{L}(g_1) \leq g_1 - f_1 = |f_1 - g_1|$$

  which again follows from

$$|\mathcal{L}(f_1) - \mathcal{L}(g_1)| \leq |f_1 - g_1|.$$

Final SVM bound

- Assembling the result we obtain:

$$P(y \neq \operatorname{sgn}(g(\mathbf{x}))) = \mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x}))\right]$$
$$\leq \frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{m\gamma}\sqrt{\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

- Note that for the Gaussian kernel this reduces to

$$P(y \neq \operatorname{sgn}(g(\mathbf{x}))) \leq \frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{\sqrt{m}\gamma} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

Final Boosting bound

- Applying a similar strategy for Boosting with the 1-norm of the slack variables we arrive at Linear programming boosting that minimises

$$\sum_h a_h + C \sum_{i=1}^{m} \xi_i,$$

where $\xi_i = (1 - y_i \sum_h a_h h(\mathbf{x}_i))_+$.

- with corresponding bound:

$$\begin{aligned} P(y \neq \operatorname{sgn}(g(\mathbf{x}))) &= \mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x}))\right] \\ &\leq \frac{1}{m} \sum_{i=1}^{m} \xi_i + \hat{R}(H) \sum_h a_h + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \end{aligned}$$

Conclusions

- Outline of philosophy and approach of SLT
- Central result of SLT
- Touched on covering number analysis for margin based analysis
- Moved to consideration of Rademacher complexity.
- Case of RC for classification giving bounds for two of the most effective classification algorithms: SVMs and Boosting

Where to find out more

Web Sites: `www.support-vector.net` (SV Machines)

`www.kernel-methods.net` (kernel methods)

`www.kernel-machines.net` (kernel Machines)

`www.neurocolt.com`

`www.pascal-network.org`