7. Learning Theory

COMP0078: Supervised Learning

Mark Herbster

21 November 2022

University College London
Department of Computer Science
LT22v-draft

Acknowledgements

I would like to thank both Shai Shalev-Shwartz and John Shawe-Taylor for permission to use a number of their slides. Please see *Shai Shalev-Shwartz's Lecture Notes (2 & 3) for Understanding Machine Learning 2014* for the original slides.

Plan

- Introduction
- PAC
- Agnostic PAC

Part I
Introduction

Review and Notation

We recall the supervised learning framework introduced in the first lecture. As convenient we will adopt the notation of *Understanding Machine Learning* [SS14] since this lecture closely follows their presentation. For simplicity we consider the 2-class setting.

[SS14] adopts the following notation for probabilities. If $\mathcal{D}$ is distribution over $\mathcal{Z}$ then if $A \subseteq \mathcal{Z}$ then $\mathcal{D}(A)$ denotes the probability that if z is drawn $\mathcal{D}$ that $z \in A$.

Model : Data is sampled IID from a distribution $\mathcal{D}$ (previously P) over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{0, 1\}$. The *expected error* (AKA *generalisation error*) of a function $h : \mathcal{X} \to \mathcal{Y}$ is

$$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x, y) : h(x) \neq y\}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

in our previous notation $L_{\mathcal{D}}(h) = \mathcal{E}(h) = \int [h(x) \neq y] dP(x, y)$.

The *empirical error* of h with respect to a data set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is denoted $L_S(h) = \sum_{i=1}^{m} \frac{1}{m} [h(x_i) \neq y_i]$ (previously $\mathcal{E}_{\text{emp}}(S, h)$).

Validation Set Bound

Theorem

Select an h then for any $\delta \in (0, 1)$. With probability at least $1 - \delta$ over the random sample V of size m from $\mathcal{D}$ we have that

$$L_{\mathcal{D}}(h) \leq L_V(h) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}\,.$$

Interpretation and Application

The generalisation error of a function h may be bounded by its empirical error. Thus in practice we may select predictor h based on a training set S then we may bound its generalisation error by *validating* on a separate set of data V.

Validation Set Bound (Proof 1)

We will use a concentration inequality.

Idea

The law of large numbers states that the empirical average of an IID sequence of random variables $\frac{1}{m}\sum_{i=1}^{m} Z_i$ will converge to the mean $\mu = E[Z_i]$. Concentration inequalities quantify how the empirical average deviates from the (true) mean when m is finite.

Hoeffding's Inequality

Let $Z_1, Z_2, \ldots, Z_m$ be IID Bernoulli random variables where for all i the $\mathbb{P}[Z_i = 1] = p$ and $\forall i : \mathbb{P}[Z_i = 0] = 1 - p$. Let $\bar{Z} := \frac{1}{m}\sum_{i=1}^{m} Z_i$ then for any $\epsilon > 0$,

$$\mathbb{P}[\bar{Z} > p + \epsilon] \leq \exp(-2m\epsilon^2)$$
$$\mathbb{P}[\bar{Z} < p - \epsilon] \leq \exp(-2m\epsilon^2)$$

Validation Set Bound (Proof 2)

Given h we will bound

$$L_{\mathcal{D}}(h) - L_V(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] - \frac{1}{m}\sum_{i=1}^{m}[h(x_i) \neq y_i]$$

1. Define $Z_i := [h(x_i) \neq y_i]$.
2. $Z_1, \ldots, Z_m$ are statistically independent.
3. $\forall i : \mathbb{P}[Z_i] = L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$
4. Applying Hoeffding gives

$$\mathbb{P}\left[L_{\mathcal{D}}(h) - L_V(h) > \epsilon\right] \leq \exp(-2\epsilon^2 m)$$

5. Set $\delta := \exp(-2\epsilon^2 m)$ and solving gives the theorem.
6. **Note:** if we use both the upper and lower bounds in Hoeffding then alternately

$$|L_{\mathcal{D}}(h) - L_V(h)| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}\,.$$

□

Observations

The above is a nice result with minimal assumptions. However it has drawbacks,

1. The validation bound gives a way to estimate a confidence interval for our generalisation error. However the data V used for validating cannot have been used for training thus it is expensive in data.
2. Often, we have only few examples. Can we choose a model and assess its expected error directly (without splitting the training data)?
3. Rather than only derive confidence bounds about a predictor h we would like instead would like to analyse predictors derived from machine learning algorithms – thus additionally deriving insights into ML algorithms.
4. **Learning theory** studies conditions which ensure the **predictivity** of a learning algorithm:
   - The expected error is close to the empirical error
   - The expected error decreases when the number of samples increases
5. In this lecture we will limit ourselves to the study of Empirical Risk Minimisation.

Theories of learning

- Basic approach of Statistical Learning Theory (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 as well as algorithms that aim at optimising the predictions.
- Each theory has strengths and weaknesses – the better it captures the details of real world experience, the better the theory and the better the chances of it making accurate predictions and driving good algorithms.

General statistical considerations

- Statistical models (not including Bayesian) begin with an assumption that the data is generated by an underlying distribution $\mathcal{D}$ 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'

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

generated identically and independently (i.i.d.) according to the distribution $\mathcal{D}$.

Generalisation of a learner

- Assume that we have a learning algorithm $\mathcal{A}$ that chooses a hypothesis (function) $\mathcal{A}_{\mathcal{H}}(S)$ from a hypothesis (function) space $\mathcal{H}$ in response to the training set S.
- We are interested in the study of $L_{\mathcal{D}}(\mathcal{A}_{\mathcal{H}}(S))$ i.e., the generalisation error of our algorithm $\mathcal{A}_{\mathcal{H}}(S)$ which is a random variable dependent on the draw of S.
- In particular we will study empirical risk minimization

$$\mathrm{ERM}_{\mathcal{H}}(S) := \arg\min_{h \in \mathcal{H}} L_S(h)$$

- **Note:** There may be many possible empirical risk minimisers we assume $\mathrm{ERM}_{\mathcal{H}}(S)$ outputs an arbitrary one.

Example of Generalisation I

- We consider the Breast Cancer dataset from the UCI repository.
- Use the simple Parzen window classifier described by Bernhard Schölkopf: weight vector is

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

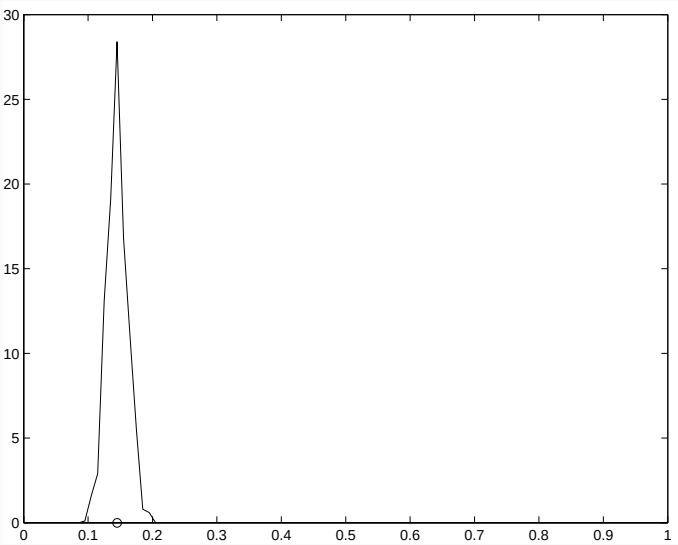
where $\boldsymbol{w}^+$ is the average of the positive training examples and $\boldsymbol{w}^-$ is average of negative training examples. Threshold is set so hyperplane bisects the line joining these two points.

Example of Generalisation II

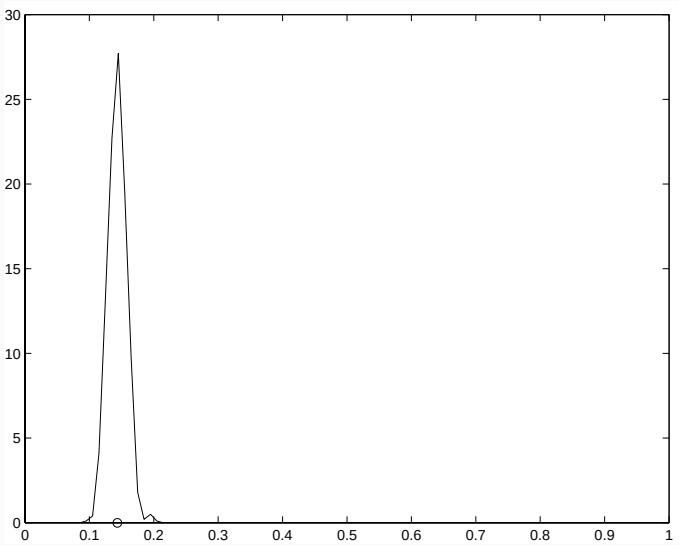
- Given a size m of the training set, by repeatedly drawing random training sets S we estimate the distribution of $L_D(\mathcal{A}_{\mathcal{H}}^{\text{parzen}}(S))$ by using the test set error as a proxy for the true generalisation.
- We plot the histogram and the average of the distribution for various sizes of training set (342, 273, 205, 137, 68, 34, 27, 20, 14, 7)

Error distributions – 1

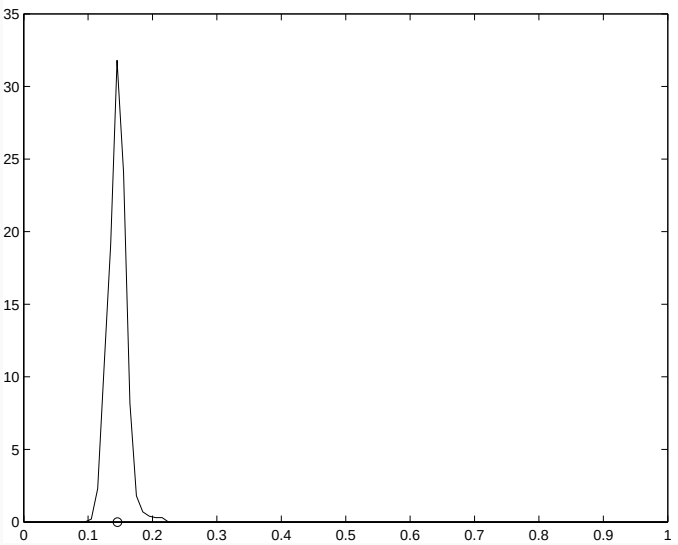
dataset size: 342



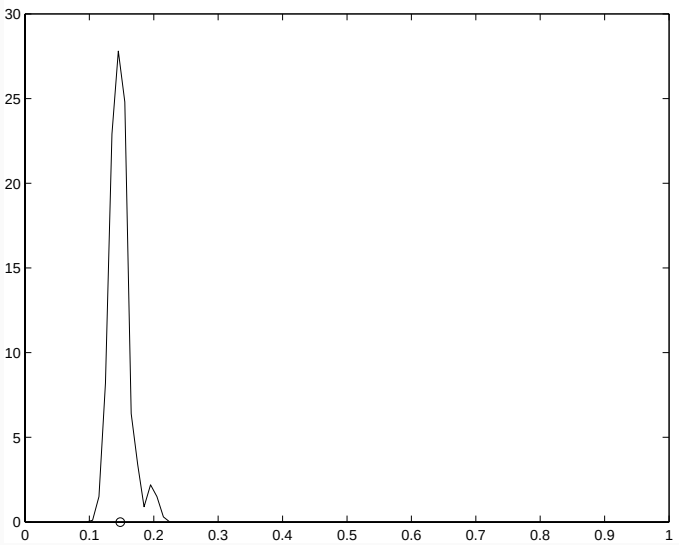
dataset size: 273



dataset size: 205

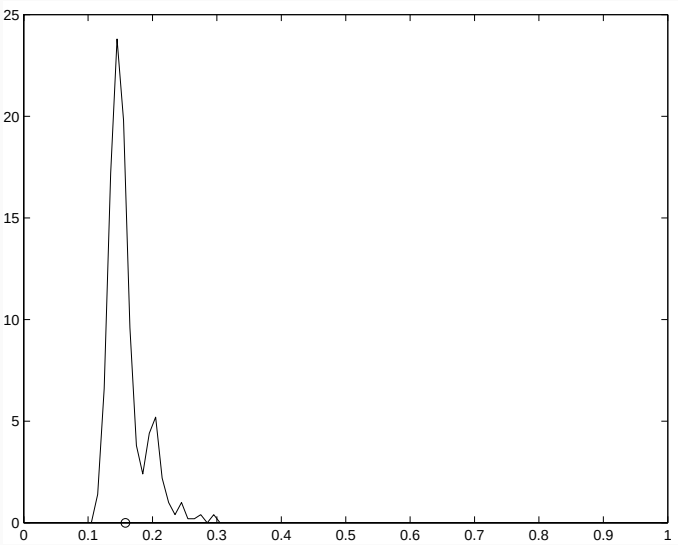


dataset size: 137

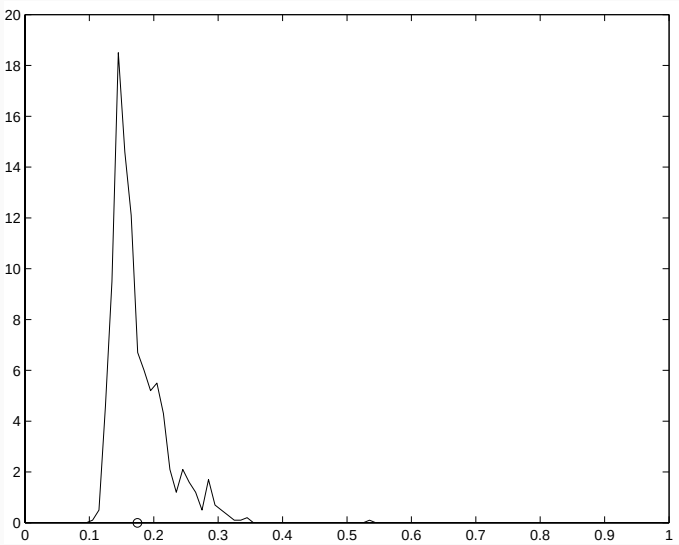


Error distributions – 2

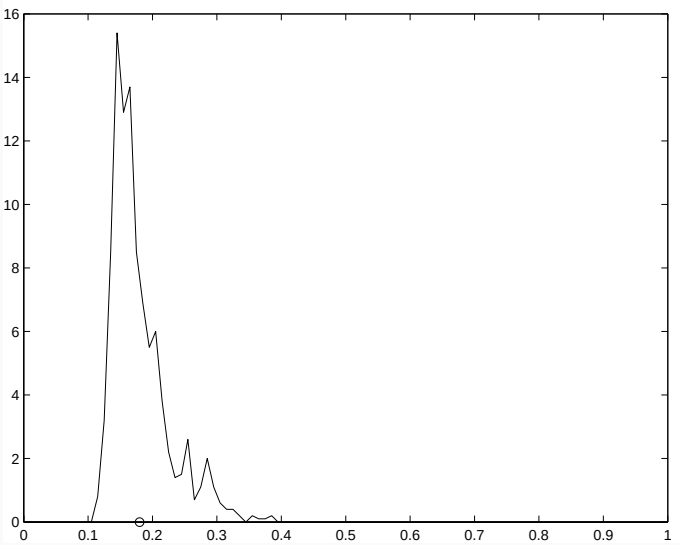
dataset size: 68



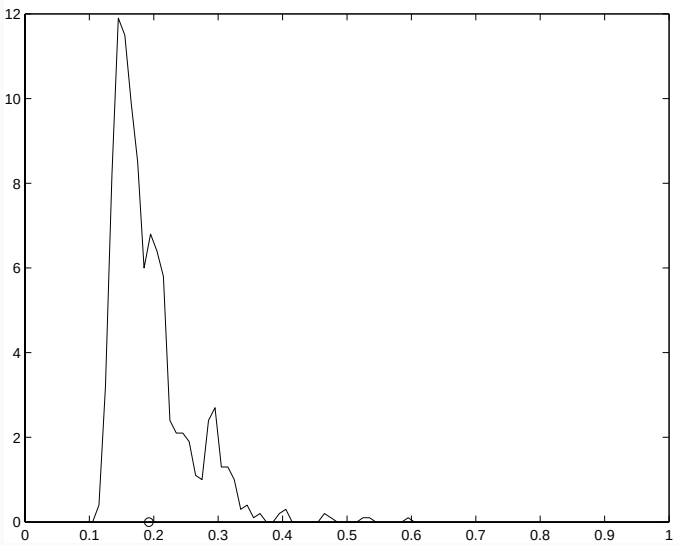
dataset size: 34



dataset size: 27

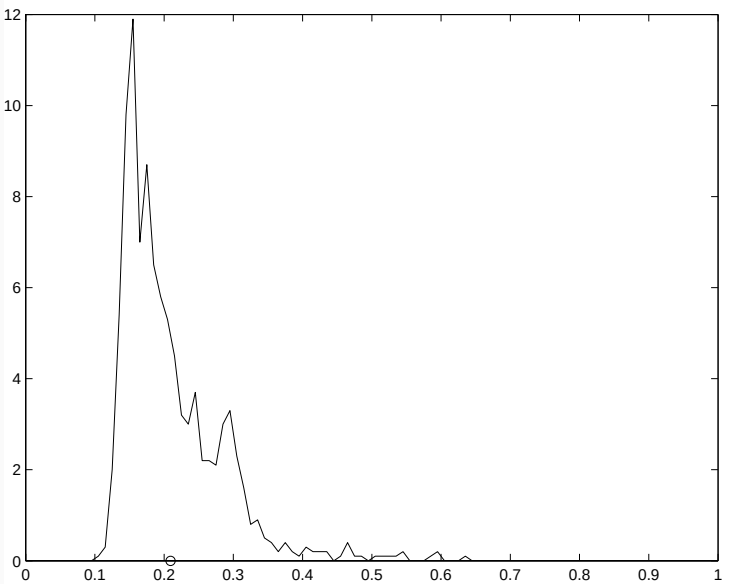


dataset size: 20

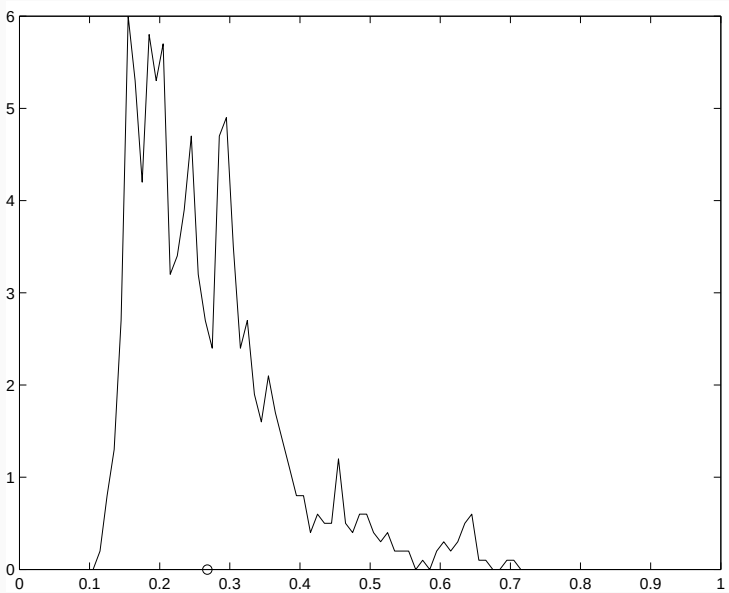


Error distributions – 3

dataset size: 14



dataset size: 7



Bayes risk and consistency

- Traditional statistics has concentrated on analysing

$$\lim_{m \to \infty} E_{S_m}[L_{\mathcal{D}}(\mathcal{A}(S_m)].$$

  where S_m denotes a training set of size m.
- For example consistency of a classification algorithm is the function with the lowest possible risk, often referred to as the Bayes risk.
- Recall our discussion of *kNN*.

Expected versus confident bounds

- For a finite sample the generalisation $L_{\mathcal{D}}(\mathcal{A}(S_m)$ 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.
- Observe in the previous example on the breast cancer data the mean generalisation error only shifted moderately compared to the movement of the “tail” of the distribution.
- 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.

Part II
PAC Learning

Realisability Assumption

Realisablity Assumption: (We will later remove this assumption)

- We will assume that there exists some true function f^* so that for all $x \in \mathcal{X}$ we have $f^*(x) = y$ (i.e., there exists a classifier with zero error).
- An implication of this assumption is that we can treat $\mathcal{D}$ now as only a distribution over $\mathcal{X}$ (as opposed to $\mathcal{X} \times \mathcal{Y}$.
- To indicate the dependence on f^* we will now use the notation

$$L_{\mathcal{D},f^*}(h) = \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f^*(x)]$$

- Can we find an algorithm $\mathcal{A}$ so that $h = \mathcal{A}(S)$ s.t. $L_{\mathcal{D},f^*}(h)$ is small?

Can only be Approximately correct

- Claim: We can't hope to find h s.t. $L_{(\mathcal{D},f^*)}(h) = 0$
- Proof: for every $\epsilon \in (0,1)$ take $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{D}(\{x_1\}) = 1 - \epsilon$, $\mathcal{D}(\{x_2\}) = \epsilon$
- The probability not to see x_2 at all among m i.i.d. examples is $(1-\epsilon)^m \approx e^{-\epsilon m}$
- So, if $\epsilon \ll 1/m$ we're likely not to see x_2 at all, but then we can't know its label
- Relaxation: We'd be happy with $L_{(\mathcal{D},f^*)}(h) \leq \epsilon$, where ϵ is user-specified

Can only be Probably correct

- Recall that the input to the learner is randomly generated
- There's always a (very small) chance to see the same example again and again
- Claim: No algorithm can guarantee $L_{(\mathcal{D},f^*)}(h) \leq \epsilon$ for sure
- Relaxation: We'd allow the algorithm to fail with probability δ, where $\delta \in (0, 1)$ is user-specified
  Here, the probability is over the random choice of examples

Probably Approximately Correct (PAC) learning

- The learner doesn't know $\mathcal{D}$ and f^*.
- The learner receives accuracy parameter ϵ and confidence parameter δ
- The learner can ask for training data, S, containing $m(\epsilon, \delta)$ examples (that is, the number of examples can depend on the value of ϵ and δ, but it can't depend on $\mathcal{D}$ or f^*)
- Learner should output a hypothesis h s.t. with probability of at least $1 - \delta$ it holds that $L_{\mathcal{D},f^*}(h) \leq \epsilon$
- That is, the learner should be **P**robably (with probability at least $1 - \delta$) **A**pproximately (up to accuracy ϵ) **C**orrect

No Free Lunch and Prior Knowledge

No Free Lunch

- Suppose that $|\mathcal{X}| = \infty$
- For any finite $C \subset \mathcal{X}$ take $\mathcal{D}$ to be uniform distribution over C
- If number of training examples is $m \leq |C|/2$ the learner has no knowledge on at least half the elements in C
- Formalizing the above, it can be shown that:

Theorem (No Free Lunch)

Fix $\delta \in (0, 1), \epsilon < 1/2$. For every learner A and training set size m, there exists $\mathcal{D}, f^*$ such that with probability of at least δ over the generation of a training data, S, of m examples, it holds that $L_{\mathcal{D},f^*}(A(S)) \geq \epsilon$.

Remark: $L_{\mathcal{D},f^*}(\text{random guess}) = 1/2$, so the theorem states that you can't be better than a random guess. **Note:** This is stronger result in expectation than given in lecture 1. See Problem 5.3 in [SS14].

Prior Knowledge

- Give more knowledge to the learner: the target f comes from some hypothesis class, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$
- The learner knows $\mathcal{H}$
- Is it possible to PAC learn now ?
- Of course, the answer depends on $\mathcal{H}$ since the No Free Lunch theorem tells us that for $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ the problem is not solvable ...

PAC Learning of Finite Hypothesis Classes

Learning Finite Classes

- Assume that $\mathcal{H}$ is a finite hypothesis class
  - E.g.: $\mathcal{H}$ is all the functions from $\mathcal{X}$ to $\mathcal{Y}$ that can be implemented using a Python program of length at most b
- Use the Consistent learning rule:
  - Input: $\mathcal{H}$ and $S = (x_1, y_1), \ldots, (x_m, y_m)$
  - Output: any $h \in \mathcal{H}$ s.t. $\forall i,\ y_i = h(x_i)$
- This is also called Empirical Risk Minimization (ERM)

$\mathrm{ERM}_{\mathcal{H}}(S)$

- Input: training set $S = (x_1, y_1), \ldots, (x_m, y_m)$
- Define the empirical risk: $L_S(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|$
- Output: any $h \in \mathcal{H}$ that minimizes $L_S(h)$

Learning Finite Classes

Theorem

Fix $\epsilon, \delta \in (0, 1)$. *If*

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$$

then for every $\mathcal{D}, f^*$, *with probability of at least* $1 - \delta$ *(with respect to the randomly sampled training set* S *of size* m*),*

$$L_{\mathcal{D},f^*}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \epsilon\,.$$

Interpretation

By rearranging we have,

$$L_{\mathcal{D},f^*}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m}\,.$$

Thus ERM in the realisable case generalisation error decreases linearly in the number of samples m and increases logarithmically in the size of the hypothesis class.

Proof

- Let $S|_x = (x_1, \ldots, x_m)$ be the instances of the training set
- We would like to prove:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \leq \delta$$

- Let $\mathcal{H}_B$ be the set of "bad" hypotheses,

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f^*)}(h) > \epsilon\}$$

- Let M be the set of "misleading" samples,

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

- Observe:

$$\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

Proof (Cont.)

Lemma (Union bound)

For any two sets A, B and a distribution $\mathcal{D}$ we have

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B) \ .$$

- We have shown:
  $\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$
- Therefore, using the union bound

$$\begin{aligned}
&\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \\
&\qquad \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \\
&\qquad \leq |\mathcal{H}_B| \max_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})
\end{aligned}$$

Proof (Cont.)

- Observe:

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \;=\; (1 - L_{\mathcal{D},f^*}(h))^m$$

- If $h \in \mathcal{H}_B$ then $L_{\mathcal{D},f^*}(h) > \epsilon$ and therefore

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \;<\; (1 - \epsilon)^m$$

- We have shown:

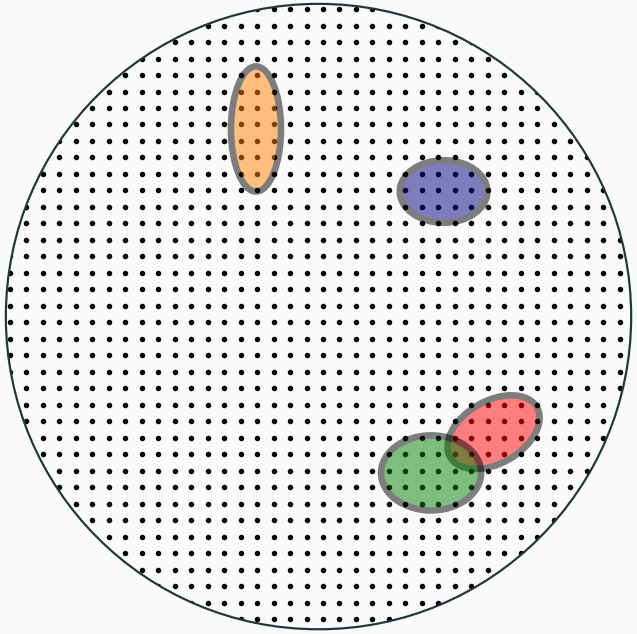
$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \;<\; |\mathcal{H}_B|\,(1 - \epsilon)^m$$

- Finally, using $1 - \epsilon \leq e^{-\epsilon}$ and $|\mathcal{H}_B| \leq |\mathcal{H}|$ we conclude:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f^*)}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \;<\; |\mathcal{H}|\,e^{-\epsilon\, m}$$

- The right-hand side would be $\leq \delta$ if $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$. □

Illustrating the use of the union bound



- Each point is a possible sample $S|_x$. Each colored oval represents misleading samples for some $h \in \mathcal{H}_B$. The probability mass of each such oval is at most $(1-\epsilon)^m$. But, the algorithm might err if it samples $S|_x$ from any of these ovals.

The Fundamental Theorem of Learning Theory

PAC learning

Definition (PAC learnability)

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- for every $\epsilon, \delta \in (0,1)$
- for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f^* : \mathcal{X} \to \{0,1\}$

when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f^*, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f^*)}(h) \leq \epsilon$.

$m_{\mathcal{H}}$ is called the sample complexity of learning $\mathcal{H}$

PAC learning

Leslie Valiant, Turing award 2010

For transformative contributions to the theory of computation, including the theory of probably approximately correct (PAC) learning, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing.

What is learnable and how to learn?

- We have shown:

Corollary

Let $\mathcal{H}$ be a finite hypothesis class.

- *$\mathcal{H}$ is PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$*
- *This sample complexity is obtained by using the $\mathrm{ERM}_{\mathcal{H}}$ learning rule*

- What about infinite hypothesis classes?
- What is the sample complexity of a given class?
- Is there a generic learning algorithm that achieves the optimal sample complexity ?

What is learnable and how to learn?

The fundamental theorem of statistical learning:

- The sample complexity is characterized by the VC dimension
- The ERM learning rule is a generic (near) optimal learner

Chervonenkis

Vapnik

The VC dimension — Motivation

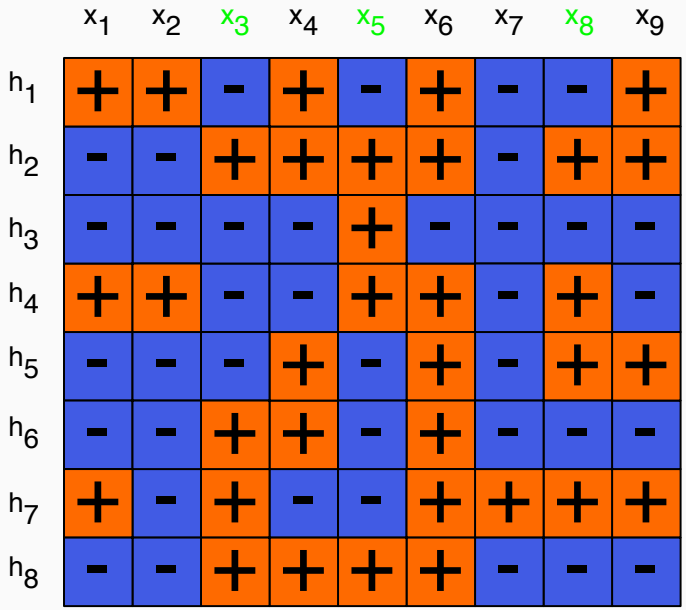
- Suppose we got a training set $S = (x_1, y_1), \ldots, (x_m, y_m)$
- We try to explain the labels using a hypothesis from $\mathcal{H}$
- Then, ooops, the labels we received are incorrect and we get the same instances with different labels, $S' = (x_1, y'_1), \ldots, (x_m, y'_m)$
- We again try to explain the labels using a hypothesis from $\mathcal{H}$
- If this works for us, no matter what the labels are, then something is fishy ...
- Formally, if $\mathcal{H}$ allows all functions over some set C of size m, then based on the No Free Lunch, we can't learn from, say, $m/2$ examples

The VC dimension — Formal Definition

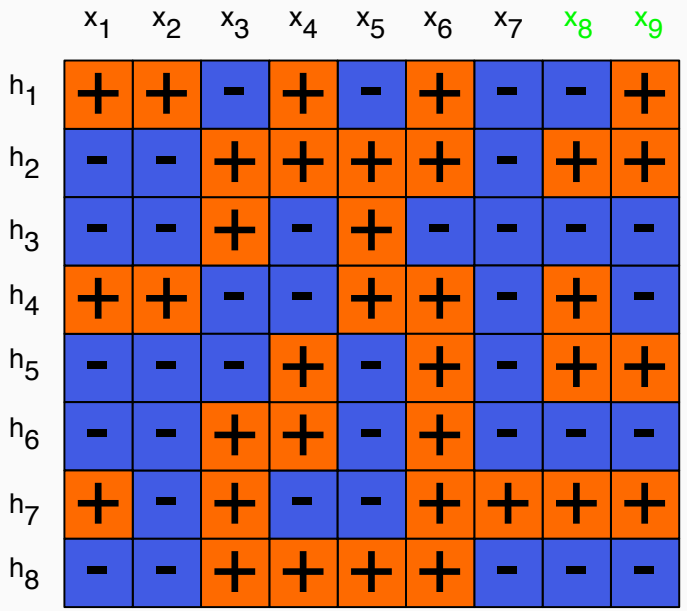
- Let $C = \{x_1, \ldots, x_{|C|}\} \subset \mathcal{X}$
- Let $\mathcal{H}_C$ be the restriction of $\mathcal{H}$ to C, namely, $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\}$ where $h_C : C \to \{-1, 1\}$ is s.t. $h_C(x_i) = h(x_i)$ for every $x_i \in C$
- Observe: we can represent each h_C as the vector $(h(x_1), \ldots, h(x_{|C|})) \in \{\pm 1\}^{|C|}$
- Therefore: $|\mathcal{H}_C| \leq 2^{|C|}$
- We say that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| = 2^{|C|}$
- $\mathrm{VCdim}(\mathcal{H}) = \sup\{|C| \ : \ \mathcal{H} \text{ shatters } C\}$
- That is, the VC dimension is the maximal size of a set C such that $\mathcal{H}$ gives no prior knowledge w.r.t. C

VC-Dimension Illustrated

• We may visualise a hypothesis space $\mathcal{H}$ as a sign matrix (potentially infinite. The rows representing hypotheses and the columns instances. The VC-dimension $\mathrm{VCdim}(\mathcal{H})$ is determined by the maximal $2^{\mathrm{VCdim}(\mathcal{H})} \times \mathrm{VCdim}(\mathcal{H})$ submatrix in which each row is distinct, i.e, all $2^{\mathrm{VCdim}(\mathcal{H})}$ possible sign patterns of length $\mathrm{VCdim}(\mathcal{H})$ are present.



$\mathrm{VCdim}(\mathcal{H}) = 3$



$\mathrm{VCdim}(\mathcal{H}) = 2$

VC dimension — Examples

To show that $\mathrm{VCdim}(\mathcal{H}) = d$ we need to show that:

1. There exists a set C of size d which is shattered by $\mathcal{H}$.
2. Every set C of size $d + 1$ is not shattered by $\mathcal{H}$.

VC dimension — Examples

Threshold functions: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathbb{R}\}$

- Show that $\{0\}$ is shattered
- Show that any two points cannot be shattered

Assume for this lecture that

$$\text{sign}(z) := \begin{cases} 1 & z \geq 0 \\ -1 & z < 0 \end{cases}$$

This differ from the usual definition where $\text{sign}(0) = 0$.

VC dimension — Examples

Intervals: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_{a,b} : a < b \in \mathbb{R}\}$, where $h_{a,b}(x) = 1$ iff $x \in [a, b]$

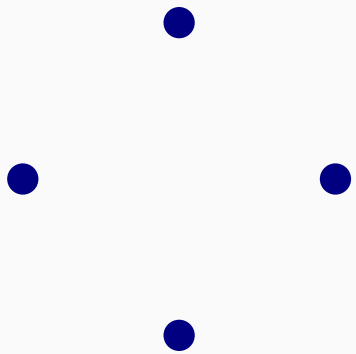
- Show that $\{0, 1\}$ is shattered
- Show that any three points cannot be shattered

VC dimension — Examples

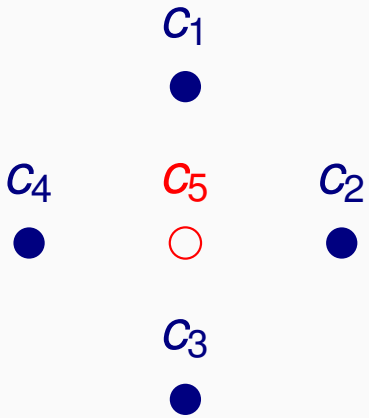
Axis aligned rectangles: $\mathcal{X} = \mathbb{R}^2$,
$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 < a_2 \text{ and } b_1 < b_2\}$, where
$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = 1$ iff $x_1 \in [a_1, a_2]$ and $x_2 \in [b_1, b_2]$

Show:

Shattered



Not Shattered



VC dimension — Examples

Finite classes:

- Show that the VC dimension of a finite $\mathcal{H}$ is at most $\log_2(|\mathcal{H}|)$.
- Show that there can be arbitrary gap between $\mathrm{VCdim}(\mathcal{H})$ and $\log_2(|\mathcal{H}|)$

VC dimension — Examples

Halfspaces: $\mathcal{X} = \mathbb{R}^n$, $\mathcal{H} = \{\mathbf{x} \mapsto \text{sign}(\langle \boldsymbol{w}, \mathbf{x} \rangle) : \boldsymbol{w} \in \mathbb{R}^n\}$

- Show that $\{\eta_1, \ldots, \eta_n\}$ is shattered
- Show that any $n + 1$ points cannot be shattered

VC dimension (Large Margin Halfspaces)

Define the inner product space of bounded square summable sequences $\ell_2 := \{\mathbf{x} \in \mathbb{R}^\infty : \sum_{i=1}^\infty x_i^2 < \infty\}$ with inner product $\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i=1}^\infty x_i x_i'$.

Definition

Given an $X \subset \ell_2$ and a $\Lambda \in (0, \infty)$ then define,

$$\mathcal{H}_{X,\Lambda} = \{\mathbf{x} \mapsto \text{sign}(\langle \boldsymbol{w}, \mathbf{x} \rangle) : \mathbf{x} \in X, \boldsymbol{w} \in \ell_2, \|\boldsymbol{w}\| \leq \Lambda, |\langle \boldsymbol{w}, \mathbf{x} \rangle| \geq 1\}$$

- Observe that $1/\|\boldsymbol{w}\|$ is the margin.

Theorem

$$\text{VCdim}(\mathcal{H}_{X,\Lambda}) \leq \Lambda^2 \max_{\mathbf{x} \in X} \|\mathbf{x}\|^2$$

- Prove theorem (Hint : think "online learning")
- Show that for every $\Lambda \in (0, \infty)$ there exists an X such that $\text{VCdim}(\mathcal{H}_{X,\Lambda}) = \lfloor \Lambda^2 \max_{\mathbf{x} \in X} \|\mathbf{x}\|^2 \rfloor$.

The Fundamental Theorem of Statistical Learning

Theorem (The Fundamental Theorem of Statistical Learning)

Let $\mathcal{H}$ be a hypothesis class of binary classifiers. Then, there are absolute constants C_1, C_2 such that the sample complexity of PAC learning $\mathcal{H}$ is

$$C_1 \frac{\mathrm{VCdim}(\mathcal{H}) + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{\mathrm{VCdim}(\mathcal{H}) \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Furthermore, this sample complexity is achieved by the ERM learning rule.

Proof of the lower bound – main ideas

- Suppose $\mathrm{VCdim}(\mathcal{H}) = d$ and let $C = \{x_1, \ldots, x_d\}$ be a shattered set
- Consider the distribution $\mathcal{D}$ supported on C s.t.

$$\mathcal{D}(\{x_i\}) = \begin{cases} 1 - 4\epsilon & \text{if } i = 1 \\ 4\epsilon/(d-1) & \text{if } i > 1 \end{cases}$$

- If we see m i.i.d. examples then the expected number of examples from $C \setminus \{x_1\}$ is $4\epsilon m$
- If $m < \frac{d-1}{8\epsilon}$ then $4\epsilon m < \frac{d-1}{2}$ and therefore, we have no information on the labels of at least half the examples in $C \setminus \{x_1\}$
- Best we can do is to guess, but then our error is $\geq \frac{1}{2} \cdot 2\epsilon = \epsilon$

Proof of the upper bound – main ideas

- Recall our proof for finite class:
  - For a single hypothesis, we've shown that the probability of the event: $L_S(h) = 0$ given that $L_{(\mathcal{D},f^*)}(h) > \epsilon$ is at most $e^{-\epsilon m}$
  - Then we applied the union bound over all "bad" hypotheses, to obtain the bound on ERM failure: $|\mathcal{H}|\, e^{-\epsilon m}$
- If $\mathcal{H}$ is infinite, or very large, the union bound yields a meaningless bound

Proof of the upper bound – main ideas

- The two samples trick: show that

$$\mathbb{P}_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H}_B : L_S(h) = 0]$$
$$\leq 2\mathbb{P}_{S,T\sim\mathcal{D}^m}[\exists h \in \mathcal{H}_B : L_S(h) = 0 \text{ and } L_T(h) \geq \epsilon/2]$$

- Symmetrization: Since S, T are i.i.d., we can think on first sampling $2m$ examples and then splitting them to S, T at random
- If we fix h, and $S \cup T$, the probability to have $L_S(h) = 0$ while $L_T(h) \geq \epsilon/2$ is $\leq e^{-\epsilon m/4}$
- Once we fixed $S \cup T$, we can take a union bound only over $\mathcal{H}_{S\cup T}$ (i.e., the restriction of $\mathcal{H}$ to "columns" denoted by $S \cup T$) !

Proof of the upper bound – main ideas

- The growth function $\Pi_{\mathcal{H}}(m)$ is the maximum cardinality of the set of functions in $\mathcal{H}$ when restricted to m points ("columns"). Thus,

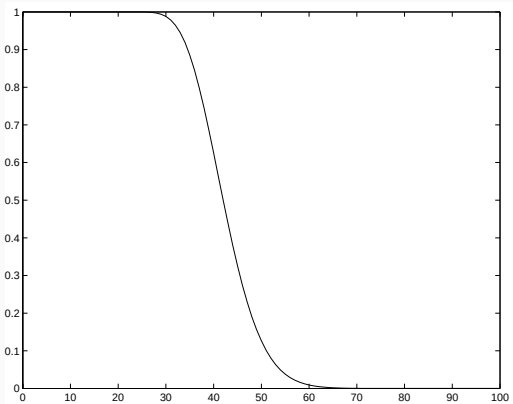
$$\Pi_{\mathcal{H}}(m) := \max_{C \subset \mathcal{X}} \{|\mathcal{H}_C| : |C| = m\}.$$

- Thus $\mathrm{VCdim}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$

Lemma: Sauer-Shelah-Perles[2]-Vapnik-Chervonenkis

Let $\mathcal{H}$ be a hypothesis class with $d = \mathrm{VCdim}(\mathcal{H}) < \infty$ then the growth function may be bounded as,

$$\Pi_{\mathcal{H}}(m) = \begin{cases} 2^m & m \leq d \\ \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d & m > d \end{cases}$$



$\Pi_{\mathcal{H}}(m)/2^m$ for linear functions in a 20 dimensional space.

• The VC-dimension is the point where the curve rapidly decreases. I.e, where we now have a polynomial growth in the number of distinct hypotheses as opposed to exponential.

• It is precisely this slowed growth that allows the bound to be proven.

Part III
Agnostic Pac

Relaxing the realizability assumption – Agnostic PAC learning

- So far we assumed that labels are generated by some $f^* \in \mathcal{H}$
- This assumption may be too strong
- Relax the realizability assumption by replacing the "target labeling function" with a more flexible notion, a data-labels generating distribution
- We now return to the assumption that $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$.
- We now use $L_{\mathcal{D}}(h)$ (as opposed to $L_{(\mathcal{D},f^*)}(h)$) :

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\})$$

- We redefine the "approximately correct" notion to

$$L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

- **Note:** We will only prove the result for finite $\mathcal{H}$ however, the proof for instead hypotheses classes with finite $\mathrm{VCdim}(\mathcal{H})$ carries over in the same way in the agnostic setting.

The General PAC Learning Problem

We wish to Probably Approximately Solve:

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \quad \text{where} \quad L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \ .$$

- Learner knows $\mathcal{H}$
- Learner receives accuracy parameter ϵ and confidence parameter δ
- Learner can decide on training set size m based on ϵ, δ
- Learner doesn't know $\mathcal{D}$ but can sample $S \sim \mathcal{D}^m$
- Using S the learner outputs some hypothesis $\mathcal{A}(S)$
- We want that with probability of at least $1 - \delta$ over the choice of S, the following would hold: $L_{\mathcal{D}}(\mathcal{A}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

Formal definition

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm, $\mathcal{A}$, with the following property: for every $\epsilon, \delta \in (0,1)$, $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, and distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$\mathcal{D}^m \left(\left\{ S \in (\mathcal{X} \times \mathcal{Y})^m : L_{\mathcal{D}}(\mathcal{A}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \right) \geq 1 - \delta$$

Learning via Uniform Convergence

Representative Sample

Definition (ϵ-representative sample)

A training set S is called ϵ-representative if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \ .$$

Representative Sample

Lemma

Assume that a training set S is $\frac{\epsilon}{2}$-representative. Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely any $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \, .$$

Proof: For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \tfrac{\epsilon}{2} \leq L_S(h) + \tfrac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon$$

Uniform Convergence is Sufficient for Learnability

Definition (uniform convergence)

$\mathcal{H}$ has the *uniform convergence property* if there exists a function $m_{\mathcal{H}}^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$, and every distribution $\mathcal{D}$,

$$\mathcal{D}^m \left(\{ S \in Z^m : S \text{ is } \epsilon\text{-representative} \} \right) \geq 1 - \delta$$

with $m \geq m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta)$.

Corollary

- *If $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then $\mathcal{H}$ is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.*
- *Furthermore, in that case, the $\text{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

Finite Classes are Agnostic PAC Learnable

We will prove the following:

Theorem

Assume $\mathcal{H}$ is finite. Then, $\mathcal{H}$ is agnostically PAC learnable using the $\mathrm{ERM}_{\mathcal{H}}$ algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

Interpretation (compare to realisable case see page 31)

By rearranging we have,

$$L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} .$$

Comparing to the realisable case generalisation error now decreases $\sqrt{m}$ rate as opposed to linearly.

Proof (cont.)

Proof: It suffices to show that $\mathcal{H}$ has the uniform convergence property with

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

- To show uniform convergence, we need:

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta .$$

- Using the union bound (compare to Page 33):

$$\begin{aligned} &\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) = \\ &\mathcal{D}^m(\cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \\ &\sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) . \end{aligned}$$

Proof (cont.)

This is the same argument we used on Page 8 (Validation Set Bound Proof)

- Recall: $L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$ and $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}[h(x_i) \neq y_i]$.
- Define $Z_i := [h(x_i) \neq y_i]$.
- $Z_1, \ldots, Z_m$ are statistically independent.

Recall:

Hoeffding's Inequality

Let $Z_1, Z_2, \ldots, Z_m$ be IID Bernoulli random variables where for all i the $\mathbb{P}[Z_i = 1] = p$ and $\forall i : \mathbb{P}[Z_i = 0] = 1 - p$. Let $\bar{Z} := \frac{1}{m}\sum_{i=1}^{m} Z_i$ then for any $\epsilon > 0$,

$$\mathbb{P}[\bar{Z} > p + \epsilon] \leq \exp(-2m\epsilon^2)$$
$$\mathbb{P}[\bar{Z} < p - \epsilon] \leq \exp(-2m\epsilon^2)$$

This implies:

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \;\leq\; 2\,\exp\left(-2\,m\,\epsilon^2\right)\;.$$

Proof (cont.)

We have shown:

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2\,|\mathcal{H}|\,\exp\left(-2\,m\,\epsilon^2\right)$$

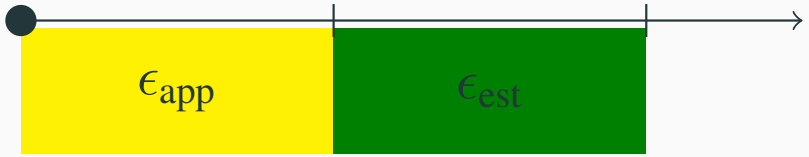
So, if $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$ then the right-hand side is at most δ as required.

□

Error Decomposition

- Let $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$. We can decompose the risk of h_S as:

$$L_{\mathcal{D}}(h_S) = \epsilon_{\mathrm{app}} + \epsilon_{\mathrm{est}}$$



- **The approximation error,** $\epsilon_{\mathrm{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$:
  - How much risk do we have due to restricting to $\mathcal{H}$
  - Doesn't depend on S
  - Decreases with the complexity (size, or VC dimension) of $\mathcal{H}$
- **The estimation error,** $\epsilon_{\mathrm{est}} = L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$:
  - Result of L_S being only an estimate of $L_{\mathcal{D}}$
  - Decreases with the size of S
  - Increases with the complexity of $\mathcal{H}$

- Bias/Complexity : Choosing $\mathcal{H}' \supset \mathcal{H}$ leads to decreased ϵ_{app} while ϵ_{est} is increased.

Summary

- PAC-learning provides guarantees on the generalisation error under worst-case distributional assumptions
- ERM is near optimal in this setting.
- NOT DISCUSSED : many interesting problems there are no efficient ERM algorithms. For example linear classification under 0-1 loss. k-literal disjunction learning.

Problems – 1

1. Problem 3.3 in [SS14]
2. Prove $(1 - x) \leq e^{-x}$ for $x \in \mathbb{R}$.
3. Problem 5.3 in [SS14]
4. Problem 6.2 in [SS14]
5. Problem 6.7 in [SS14]
6. On page 49 why did we formulate the VC-dimension of large margin classifiers with respect to a fixed input space X rather than for example $\{\mathbf{x} \in \ell_2 : \|\mathbf{x}\| \leq R\}$?
7. Prove the VC-dimension results on 49.

Suggested Readings

This lecture closely follows Chapters 3,4 and 6 from [SS14] (Chapter 5, for No Free Lunch). We omitted proofs for Saurer's Lemma and the generalisation of our results from finite hypothesis class to classes with bounded VC-dimension. Sauer's Lemma is proved relatively directly in 6.5.1. Extending from finite hypotheses classes to bounded VC-dimension introduces the powerful technique of the double sample trick, proofs via Rademacher Complexities are in 28.1 and 28.3 in [SS14]. Arguably there is more insight to be gained by looking at the proofs that more directly show the parallel to the finite hypothesis class case, I suggest for example Maria-Florina Balcan's notes *Lectures 5,6 and 7 : 8803 Machine Learning Theory 2010*.

Useful references

1. *Understanding Machine Learning from Theory to Algorithms*, S. Shalev-Shwartz and S. Ben-David, Cambridge University Press (2014)