CS:E4830 Kernel Methods in Machine Learning

Lecture 5 : Introductory Learning Theory - II

Rohit Babbar

6th February, 2019

Generalization in Machine Learning

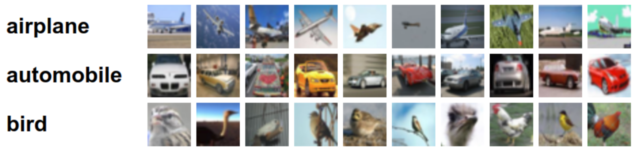


Figure: Some examples from three of the **ten classes** from CIFAR-10 dataset (others being **cat**, **deer**, **dog**, **frog**, **horse**, **ship**, **truck**) are shown above. Dataset contains 50,000 training and 10,000 test images for a total of 6,000 images per class

Consider the following scenario :

- Train a deep net on the above dataset, and test on the test set
  - What are the training error and test errors?
- Now, keep the training set images same but randomly shuffle their labels
  - Keep the test set same as the previous case
  - Train a deep net on the training set with randomized labels, and test on the test set,
  - What are the training and test errors?
  - Does the training process take longer in this case ?

Understanding Deep Learning Requires Rethinking Generalization - I

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Figure: Understanding Deep Learning Requires Rethinking Generalization

- Paper in ICLR 2017

Understanding Deep Learning Requires Rethinking Generalization - II

2.2 Implications

In light of our randomization experiments, we discuss how our findings pose a challenge for several traditional approaches for reasoning about generalization.

Rademacher complexity and VC-dimension. Rademacher complexity is commonly used and flexible complexity measure of a hypothesis class. The empirical Rademacher complexity of a hypothesis class $\mathcal{H}$ on a dataset $\{x_1, \ldots, x_n\}$ is defined as

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right] \tag{1}$$

where $\sigma_1, \ldots, \sigma_n \in \{\pm 1\}$ are i.i.d. uniform random variables. This definition closely resembles our randomization test. Specifically, $\hat{\mathfrak{R}}_n(\mathcal{H})$ measures ability of $\mathcal{H}$ to fit random ± 1 binary label assignments. While we consider multiclass problems, it is straightforward to consider related binary classification problems for which the same experimental observations hold. Since our randomization tests suggest that many neural networks fit the training set with random labels perfectly, we expect that $\hat{\mathfrak{R}}_n(\mathcal{H}) \approx 1$ for the corresponding model class $\mathcal{H}$. This is, of course, a trivial upper bound on the Rademacher complexity that does not lead to useful generalization bounds in realistic settings. A similar reasoning applies to VC-dimension and its continuous analog fat-shattering dimension, unless we further restrict the network. While Bartlett (1998) proves a bound on the fat-shattering

Figure: Discussion in the paper

Formal framework to study overfitting

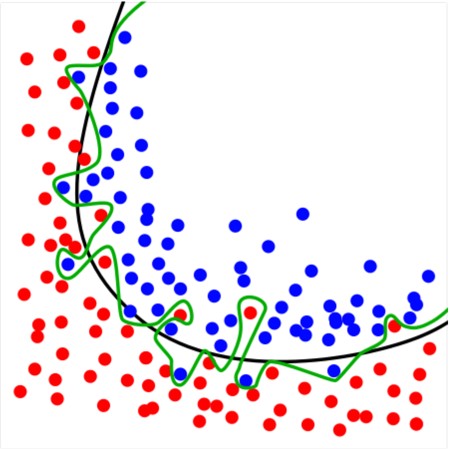
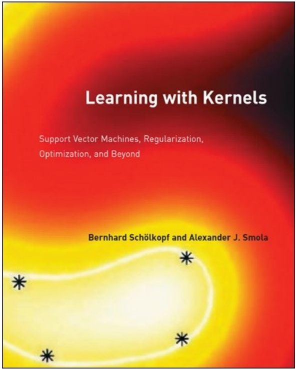
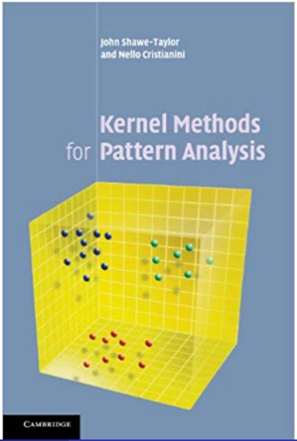


Figure: Overfitting example (picture from Wikipedia)

- Learning theory provides a formal framework to study over-fitting

- Material in this lecture is based on tutorial by von-Luxburg and Schoelkopf
- More details in
  - Chapter 5 - Learning with kernels - Schoelkopf and Smola
  - Chapter 4 - Kernel Methods for Pattern Analysis - Shawe-Taylor and Christianini





Terminology from last lecture

For a given classification problem :

- 0-1 loss ($\ell_{0,1}(.)$ denoted below by $\ell(.)$ for simplicity) of a classifier f on an input-output pair (x, y) is given by :

$$\ell(y, f(x)) = \begin{cases} 1 & \text{if } f(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

- Empirical error of f is $R_{emp}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$
- Expected error of f

$$R(f) := \mathbb{E}_{(x,y)}(\ell(y, f(x)))$$

The above expectation is w.r.t the joint distribution P over $\mathcal{X} \times \mathcal{Y}$

Some important classifiers

- Bayes classifier $f_{Bayes} := \arg\min_f R(f)$ over unrestricted function class.
- For a function class $\mathcal{F}$, $f_{\mathcal{F}} := \arg\min_{f \in \mathcal{F}} R(f)$ - one which minimizes **expected error** in a pre-defined function class $\mathcal{F}$
- Based on training sample of size n, the classifier $f_n := \arg\min_{f \in \mathcal{F}} R_{emp}(f)$ is the one which minimizes **empirical error** in $\mathcal{F}$

Bayes Classifier

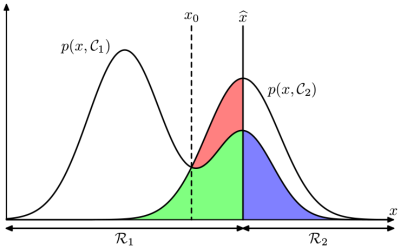


Figure: Depiction of noisy labels (picture from Chris Bishop's book, chapter 1)

$p(x, C_1)$ refers to the joint probability density of input x and class C_1

- Given a point, say $x = \hat{x}$, how do we compute $P(y = C_1 | X = \hat{x})$ or $P(y = C_2 | X = \hat{x})$? Using Bayes rule,

$$P(y = C_1 | X = \hat{x}) = \frac{P(y = C_1, X = \hat{x})}{P(X = \hat{x})} = \frac{p(y = C_1, X = \hat{x})dx}{p(y = C_1, X = \hat{x})dx + p(y = C_2, X = \hat{x})dx}$$

- At what point in the graph $P(y = C_1 | X = x) = 0.5$?
- What kind of errors are signified by the red, green and blue regions?

Large vs Small Function class

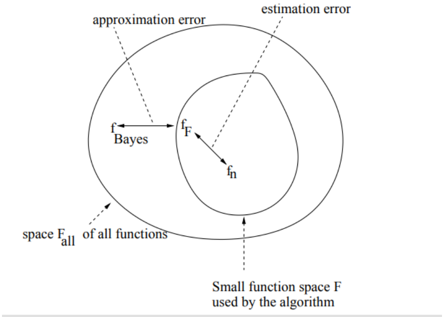


Figure: Pictorial depiction of the components of classification error

- The space F_{all} contains all possible functions that may be implmented using SVM, Deep nets, Random Forest and everything else
- **Estimation error** - $(R(f_n) - R(f_{\mathcal{F}}))$ - **finiteness of training data**
- **Approximation error** - $(R(f_{\mathcal{F}}) - R(f_{Bayes}))$ - **choice of function class**

Large vs Small Function class

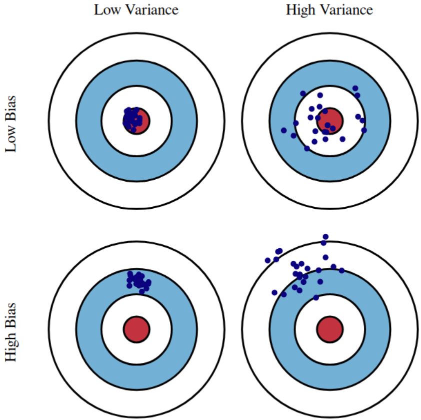


Figure: Pictorial depiction of the components of classification error

- **Estimation error** - $(R(f_n) - R(f_{\mathcal{F}}))$ - corresponds to **Variance**
- **Approximation error** - $(R(f_{\mathcal{F}}) - R(f_{Bayes}))$ - corresponds to **Bias**

Empirical Risk Minimization

In practice, learning algorithms (do not have access to the underlying data generating distribution P over $\mathcal{X} \times \mathcal{Y}$) are based on minimizing error on the training data. Formally, this is given as follows :

Principle of ERM

The idea behind the principle of Empirical Risk Minimization is to find a classifier in a pre-defined function class which minimizes the empirical risk. That is

$$f_n := \arg\min_{f \in \mathcal{F}} R_{emp}(f)$$

- We want to check if the classifier (function) f_n that we learn from **ERM is consistent or not**?
- The motivation for the consistency of the principle of ERM comes from the law of large numbers.

Consistency of Learning Algorithm

Definition

Let $(x_i, y_i)_{i\in\mathbb{N}}$ be a sequence of training input-output pairs drawn according to some data distribution P. For each $n \in \mathbb{N}$, let f_n be the classifier that is learnt by some learning algorithm by seeing the first n training points, Then

- The learning algorithm (such as SVM and k-Nearest Neighbor) is called consistent w.r.t the function class $\mathcal{F}$ and the distribution P if the risk $R(f_n)$ converges in probability to the risk of the best possible classifier in $\mathcal{F}$

$$\mathbb{P}(R(f_n) - R(f_{\mathcal{F}}) > \epsilon) \to \text{ as } n \to \infty$$

Law of Large numbers

Let ξ_i be independent random variables drawn identically from a distribution P. Then the mean of the random variables converges to the mean of the distribution P when the sample size goes to infinity :

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \to \mathbb{E}(\xi) \text{ as } n \to \infty$$

Law of Large numbers

Let ξ_i be independent random variables drawn identically from a distribution P. Then the mean of the random variables converges to the mean of the distribution P when the sample size goes to infinity :

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \rightarrow \mathbb{E}(\xi) \text{ as } n \rightarrow \infty$$

- For a classifier f, let $\xi_i = \ell(f(x_i), y_i)$ be the loss on training sample (x_i, y_i), then the law of large numbers gives the following :

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(x_i)) \rightarrow \mathbb{E}_{(x,y)}(\ell(y, f(x))) \text{ as } n \rightarrow \infty$$

- The above implies that the true risk (unknown due to the unknown probability distribution P) can be approximated by the empirical risk (which can be computed from the training data)

Chernoff Bound

Non-asymptotic result

Chernoff Bound

Let $\xi_i \in [0,1]$ be independent random variables drawn identically from a distribution P. Then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbb{E}(\xi)\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

- The above inequality says that the probability that sample mean deviates from its expectation by ϵ goes down exponentially fast w.r.t sample size n

Chernoff Bound

Non-asymptotic result

Chernoff Bound

Let $\xi_i \in [0,1]$ be independent random variables drawn identically from a distribution P. Then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbb{E}(\xi)\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

- The above inequality says that the probability that sample mean deviates from its expectation by ϵ goes down exponentially fast w.r.t sample size n
- The same bound can be applied to empirical error and expected error of a classifier f. That is, for a **fixed function** f

$$\mathbb{P}(|R_{emp}(f) - R(f)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2)$$

- The above statement is a probabilistic argument, which means that it may not hold every time, and in fact, be violated in some cases (but with low probability)

Pictorial representation

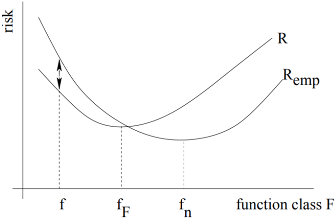


Figure: Depiction of training error and test error over a function class

- The above picture shows the dependence of test error (R) and training error (R_{emp},for a particular training set) on various functions in the function class F

Pictorial representation

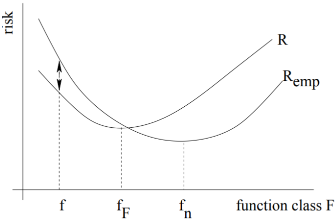


Figure: Depiction of training error and test error over a function class

- The above picture shows the dependence of test error (R) and training error (R_{emp},for a particular training set) on various functions in the function class F
- As we choose more and more training sets, by Chernoff's bound, for **every fixed function**, $R_{emp}(f)$ converges to $R(f)$,

Pictorial representation

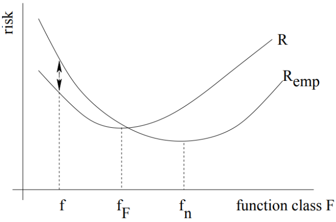


Figure: Depiction of training error and test error over a function class

- The above picture shows the dependence of test error (R) and training error (R_{emp},for a particular training set) on various functions in the function class F
- As we choose more and more training sets, by Chernoff's bound, for **every fixed function**, $R_{emp}(f)$ converges to $R(f)$,
- However, the above bound holds for a fixed function, which is not the case for ERM, which returns a different function **depending on training data**

Pictorial representation

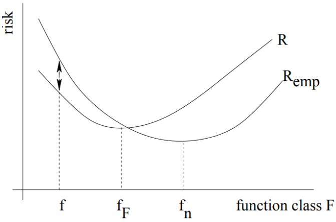


Figure: Depiction of training error and test error over a function class

- The above picture shows the dependence of test error (R) and training error (R_{emp},for a particular training set) on various functions in the function class F
- As we choose more and more training sets, by Chernoff's bound, for **every fixed function**, $R_{emp}(f)$ converges to $R(f)$,
- However, the above bound holds for a fixed function, which is not the case for ERM, which returns a different function **depending on training data**
- However, it is **not guaranteed** that $R(f_n)$ converges to $R(f_F)$, which is what we are interested in (from the definition of consistency).

When can ERM be inconsistent?

An Empirical Risk Minimization Example

- Assume that the data lies in $[0,1]$, i.e., $x \in \mathcal{X} = [0,1]$
  - Input x is chosen uniformly at random on $\mathcal{X}$,
  - the label y is chosen in a deterministic way as follows :

$$y = \begin{cases} -1 & \text{if } x < 0.5 \\ +1 & \text{otherwise} \end{cases}$$

- Consider, a potential classifier based on n training samples given as follows :

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1 \dots n \\ +1 & \text{otherwise} \end{cases}$$

- What is it error on the training set?
  - training error $= 0$
  - Has it learnt anything?

When can ERM be incosistent?

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?

When can ERM be incosistent?

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?
- Therefore, the Empirical risk minimizer is not converging to the best function in the class (What is the Bayes error (classifier) in this case?)
- Why does it happen?
  - Because we allow any function (could be highly non-smooth) in our function class
- In order to generalize, we need to **restrict our function class** by imposing some condition, which we study next.

Uniform Convergence - I

- **Uniform Convergence** is a condition over a function class which ensures consistency of ERM, and is given by $|R_{emp}(f) - R(f)| < \epsilon, \forall f \in \mathcal{F}$ for some finite sample size n
- Alternatively, the condition of Uniform Convergence can be stated $\sup_{f \in \mathcal{F}} |R_{emp}(f) - R(f)| < \epsilon$
- That is, for all functions $f \in \mathcal{F}$, the difference $|R_{emp}(f) - R(f)|$ becomes small **simultaneously**

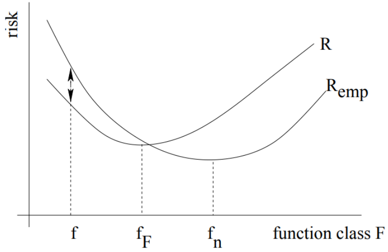


Figure: Under Uniform Convergence, the difference between the two curves becomes arbitrarily small for some large but finite sample size n

Uniform Convergence - II

We will show that if Uniform Convergence holds for a function class $\mathcal{F}$, then the Empirical Risk Minimizer is guaranteed to be consistent, i.e., $R(f_n) \to R(f_{\mathcal{F}})$ as $n \to \infty$. (Proof on the next slide)
First, couple of inequalities for the proof :

- The following holds (by definition of supremum/maximum), for any function $f \in \mathcal{F}$

$$|R(f) - R_{emp}(f)| \leq \sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)|$$

- Therefore, it also holds for the Empirical Risk Minimizer f_n which is chosen based on finite number (n) of samples

$$\mathbb{P}(|R(f_n) - R_{emp}(f_n)| \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)}
\end{aligned}
$$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&= R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_{\mathcal{F}}) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}})
\end{aligned}
$$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&= R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_{\mathcal{F}}) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\
&\leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \text{ (Why?)}
\end{aligned}
$$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&= R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_{\mathcal{F}}) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\
&\leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&\leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)|
\end{aligned}
$$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&= R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_{\mathcal{F}}) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\
&\leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&\leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)|
\end{aligned}
$$

Therefore, $\mathbb{P}(|R(f_n) - R(f_{\mathcal{F}})| \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon/2)$

Uniform Convergence - III

Sufficiency of Uniform Convergence for consistency of ERM

$$
\begin{aligned}
&|R(f_n) - R(f_{\mathcal{F}})| \\
&= R(f_n) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&= R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_{\mathcal{F}}) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\
&\leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \text{ (Why?)} \\
&\leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)|
\end{aligned}
$$

Therefore, $\mathbb{P}(|R(f_n) - R(f_{\mathcal{F}})| \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon/2)$

- Since uniform law of large numbers holds for a function class $\mathcal{F}$ by uniform convergence, the RHS tends to 0
- Since the LHS is upper bounded by RHS, this implies the consistency of ERM over the function class.

NASC for consistency of ERM - I

- The previous proof shows that uniform convergence is a sufficient condition for the consistency of ERM

NASC for consistency of ERM - I

- The previous proof shows that uniform convergence is a sufficient condition for the consistency of ERM
- Is it also necessary?

NASC for consistency of ERM - I

- The previous proof shows that uniform convergence is a sufficient condition for the consistency of ERM
- Is it also necessary?

Theorem by Vapnik and Chervonenkis

Uniform convergence, i.e.,

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon) \to 0 \text{ as } n \to \infty$$

$\forall \epsilon > 0$ is a necessary and sufficient condition for consistency of ERM with respect to the function class $\mathcal{F}$.

NASC for consistency of ERM - II

Theorem by Vapnik and Chervonenkis

Uniform convergence, i.e.,

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon) \to 0 \text{ as } n \to \infty$$

$\forall \epsilon > 0$ is a necessary and sufficient condition for consistency of ERM with respect to the function class $\mathcal{F}$.

- Learning with all possible functions
  - Larger the function class $\mathcal{F}$, so is $|R(f) - R_{emp}(f)|$, and hence difficult to achieve consistency
- Learning with a restricted function class
  - On the contrary, small $\mathcal{F}$ means easier to learn consistent classifiers
- However, unfortunately, it is not easy to find out if the *uniform convergence holds for a function class or not*

Capacity of Function Class

The main quantity of interest from the previous theorem is the following :

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon)$$

- Can we study the above quantity in the non-asymptotic regime, i.e. when the sample size n is finite
  - Practically, this also matters more since we normally have finite data size

Capacity of Function Class

The main quantity of interest from the previous theorem is the following :

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon)$$

- Can we study the above quantity in the non-asymptotic regime, i.e. when the sample size n is finite
  - Practically, this also matters more since we normally have finite data size
- In bounding the quantity $\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$, there are two challenges :
  - Infinitely many functions, due to **continuous nature of the function class**
  - The expected risk $R(f)$, which depends on the underlying probability distribution, and **cannot be computed from training data**

Capacity of Function Class

The main quantity of interest from the previous theorem is the following :

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon)$$

- Can we study the above quantity in the non-asymptotic regime, i.e. when the sample size n is finite
  - Practically, this also matters more since we normally have finite data size
- In bounding the quantity $\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$, there are two challenges :
  - Infinitely many functions, due to **continuous nature of the function class**
  - The expected risk $R(f)$, which depends on the underlying probability distribution, and **cannot be computed from training data**
- To get a handle on this, we need the following three concepts :
  - Union bound
  - Symmetrization
  - Shattering

Union Bound

For the sake of simplicity to start with

Suppose there are m functions in the function class $\mathcal{F}$

$$\begin{aligned}
&\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \\
&= \mathbb{P}(|R(f_1) - R_{emp}(f_1)| \geq \epsilon) \text{ or } \dots \text{ or } \mathbb{P}(|R(f_m) - R_{emp}(f_m)| \geq \epsilon) \\
&\leq \sum_{i=1}^{m} \mathbb{P}(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \text{ (result on the probability of union of events)} \\
&\leq 2m \exp(-2n\epsilon^2) \text{ (application of Chernoff bound)}
\end{aligned}$$

- Thus, if a function class $\mathcal{F}$ has finite number of functions, then we can bound the probability that for some $f \in \mathcal{F}$, the difference between empirical error and expected error is greater than ϵ.

In practice, useful Function classes have infinitely many functions

- Now, how do we bound the original quantity $\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$ when there are infinitely many functions
- We have infinitely many functions when learning with SVMs or Deep networks :
    - Each of the many possible orientations of an SVM hyper-plane
    - Each of the many possible settings of weights of hidden units in a deep net
- The above challenge of infinite number of functions in bounding $\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$ is handled using two concepts
    - Symmetrization
    - Shattering

Symmetrization

Symmetrization Lemma

For $n\epsilon^2 \geq 2$

$$\mathbb{P}(\sup_{f\in\mathcal{F}} |R(f) - R_{emp}(f)| > \epsilon) \leq 2\mathbb{P}(\sup_{f\in\mathcal{F}} |R_{emp}(f) - R'_{emp}(f)| > \epsilon/2)$$

where $R_{emp}(f) - R'_{emp}(f)$ refers to the difference between the empirical errors of two samples of size n, where first sample is the training sample, and the second one is called *ghost sample!* (since it does not exist in practice)

- The above lemma bounds $\mathbb{P}(\sup_{f\in\mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$ with **something that completely depends on training data** (we don't need to compute the expected error $R(f)$)
- **In simple words**, symmetrization lemma says - if the difference between empirical error of two independent samples is small (RHS of the above inequality), then the difference between empirical (training) error and expected error is also small (LHS of the above inequality).

Shattering Coefficient - I

Now we only need to focus on the following quantity :

$$2\mathbb{P}(\sup_{f\in\mathcal{F}} |R_{emp}(f) - R'_{emp}(f)| > \epsilon/2)$$

- An interesting thing to note is that even though the number of functions $f \in \mathcal{F}$ are infinite, for a training set of size n, since each instance can have label from $\{+1, -1\}$, there are only 2^n different possible labelings
- Therefore, the number of effective functions on a sample of size $2n$ over which we have to find the supremum $2\mathbb{P}(\sup_{f\in\mathcal{F}} |R_{emp}(f) - R'_{emp}(f)| > \epsilon/2)$ is **atmost** 2^{2n}.

Shattering Coefficient

For the training data, $Z_n := \{(x_i, y_i)\}_{i=1}^n$, lets define a quantity called shattering co-efficient of a function class which counts the maximum number of possible labelings it can exhibit on **any sample** of size n. It is defined as follows :

$$\mathcal{N}(\mathcal{F}, n) := \max\{|\mathcal{F}_{Z_n}| \text{ such that } x_1, \ldots, x_n \in \mathcal{X}\}$$

- Clearly, $\mathcal{N}(\mathcal{F}, n) \leq 2^n$

Shattering Coefficient - 1-D Pictorial representation

When $\mathcal{N}(\mathcal{F}, n) = 2^n$, then the function class $\mathcal{F}$ is said to shatter n points, that **there exists** a sample of n points, on which it can achieve all possible labelings. If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many can be shattered by linear classifiers in 1-dimension?

Shattering Coefficient - 1-D Pictorial representation

When $\mathcal{N}(\mathcal{F}, n) = 2^n$, then the function class $\mathcal{F}$ is said to shatter n points, that **there exists** a sample of n points, on which it can achieve all possible labelings. If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many can be shattered by linear classifiers in 1-dimension?

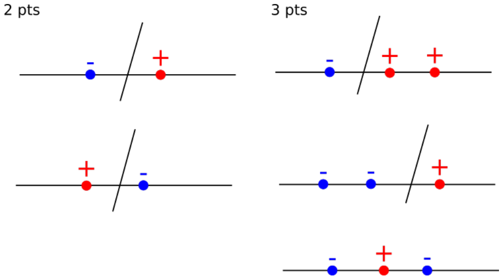


Figure: Any labeling is possible on 2 points, but not on 3

Shattering Coefficient - 2-D Pictorial representation

If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many can be shattered by linear classifiers in 2-dimension?

Shattering Coefficient - 2-D Pictorial representation

If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many can be shattered by linear classifiers in 2-dimension?

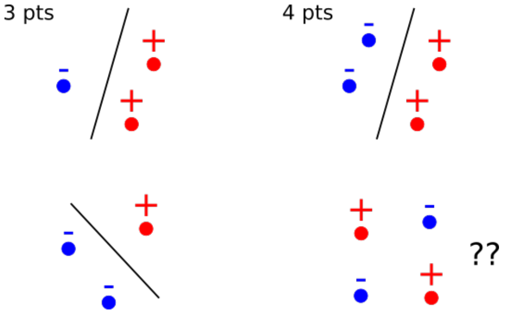


Figure: Any labeling is possible on 3 points, but not on 4

Shattering Coefficient - 3-D Pictorial representation

If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many points can be shattered by linear classifiers in 1-dimension?

Shattering Coefficient - 3-D Pictorial representation

If the function class $\mathcal{F}$ consists of linear classifiers only, then :

- How many points can be shattered by linear classifiers in 1-dimension?

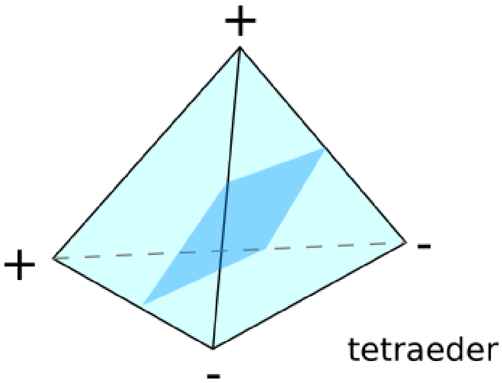


Figure: Any labeling is possible on 4 points, but not on 5

Uniform Convergence Bounds - I

We can now finally bound $\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon)$ in the following way :

$$
\begin{aligned}
&\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \\
&\leq 2\mathbb{P}(\sup_{f \in \mathcal{F}} |R_{emp}(f) - R'_{emp}(f)| > \epsilon/2) \text{ (By symmetrization)} \\
&= 2\mathbb{P}(\sup_{f \in \mathcal{F}_{Z_{2n}}} |R_{emp}(f) - R'_{emp}(f)| > \epsilon/2) \text{ (By Shattering argument)} \\
&\leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4) \text{ (Union and Chernoff's bound)}
\end{aligned}
$$

Uniform Convergence Bounds - II

Key inequality

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

We consider two following important cases :

- When $\mathcal{N}(\mathcal{F}, 2n)$ grows polynomially with n, i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ for some constant k

Uniform Convergence Bounds - II

Key inequality

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

We consider two following important cases :

- When $\mathcal{N}(\mathcal{F}, 2n)$ grows polynomially with n, i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ for some constant k
  - $2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4) \leq 2 \times (2n)^k \times \exp(-n\epsilon^2/4) = 2 \exp(k \log(2n) - n\epsilon^2/4)$

Uniform Convergence Bounds - II

Key inequality

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

We consider two following important cases :

- When $\mathcal{N}(\mathcal{F}, 2n)$ grows polynomially with n, i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ for some constant k
  - $2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4) \leq 2 \times (2n)^k \times \exp(-n\epsilon^2/4) = 2 \exp(k \log(2n) - n\epsilon^2/4)$
  - $2 \exp(k \log(2n) - n\epsilon^2/4)$ goes to 0 as $n \to \infty$ satisfying the condition for Uniform Convergence, and hence implying consistency of ERM

Uniform Convergence Bounds - II

Key inequality

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

We consider two following important cases :

- When $\mathcal{N}(\mathcal{F}, 2n)$ grows polynomially with n, i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ for some constant k
  - $2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4) \leq 2 \times (2n)^k \times \exp(-n\epsilon^2/4) = 2\exp(k \log(2n) - n\epsilon^2/4)$
  - $2\exp(k \log(2n) - n\epsilon^2/4)$ goes to 0 as $n \to \infty$ satisfying the condition for Uniform Convergence, and hence implying consistency of ERM
- Use the function class $\mathcal{F}_{all}$ (even consisting of highly non-smooth over-fitting functions), a class that can classify each sample in any way desired, $\mathcal{N}(\mathcal{F}, 2n) = 2^{2n}$
  - $2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4) = 2 \times 2^{2n} \times \exp(-n\epsilon^2/4) = 2\exp(n(2\log(2) - \epsilon^2/4))$
  - Does not go to 0 as $n \to \infty$ implying inconsistency of ERM when the function class contains all functions

Other conditions for consistency of ERM

Empirical Risk minimization if consistent with respect to $\mathcal{F}$ if and only if VC-dimension of $\mathcal{F}$ is finite.

Other concepts

- VC-Dimension
- Rademacher Complexity

Summary

- Types of error
  - Empirical Error, Expected Error, Generalization gap
  - Estimation and Approximation error
- Consistency
  - Example of Inconsistency of ERM
  - Uniform convergence as a necessary and sufficient condition for consistency of ERM
  - Symmetrization and Schattering