

IFT 6085 - Lecture 7

Elements of statistical learning theory

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes

Winter 2020: Meraj Hashemizadeh

Winter 2019: Mingde (Harry) Zhao & Dylan Troop

Winter 2018: Brady Neal and Matthew Scicluna

Instructor: Ioannis Mitliagkas

1 Summary

In the previous lecture we dove into the details of Nesterov's accelerated gradient descent and made thorough comparisons with Polyak's momentum (the heavy ball method). In this lecture we will start discussing statistical learning theory, the goal of which is to determine how well a model performs on unseen data.

Lecture Topics:

- Define the generalization gap and illustrate why it is the focus of our study
- Introduce some concentration bounds, such as Markov's inequality, Chebyshev's inequality, Chernoff's bound and Hoeffding's Inequality
- Prove a bound on the generalization gap for countable, finite hypothesis classes using Hoeffding's Inequality and the Union Bound
- Introduce the uniform convergence framework for learning
- Introduce the VC dimension

2 Introduction and Notation

The goal in machine learning is not to perform well on training data, but to perform well on unseen data. We say that a model "generalizes well" if it performs roughly the same on test data as it does on training data. Statistical learning theory is largely concerned with theoretical bounds on this difference in performance, also known as the *generalization gap*. In this lecture, we focus specifically on binary classification, but these results can be easily extended to multi-class classification.

Notation:

- \mathcal{X} - domain set (input space)
- \mathcal{Y} - label set (output space)
- m - number of training examples
- $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ - training set where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- \mathcal{D} - distribution over the data. That is, $(x_i, y_i) \sim \mathcal{D}$. Note that in our setup we have a joint distribution rather than just x_i being random and y_i being a deterministic function of x_i
- \mathcal{H} - hypothesis class (class of possible models we can learn; examples below)

- \mathcal{H}_{SVM} : class of possible SVMs on a dataset
- \mathcal{H}_{LR} : class of possible logistic regression models on a dataset
- \mathcal{H}_{NN} : class of possible neural networks of a fixed architecture on a dataset
- $\mathcal{H} \subset \{h : \mathcal{X} \rightarrow \mathcal{Y}\}$: \mathcal{H} is a subset of all possible functions that map from input space to output space. Choosing this subset (hypothesis class), \mathcal{H} , introduces inductive bias.
- In the example of binary classification on a d dimensional real-valued dataset, we have $h(x_i) = \hat{y}_i$ where $h \in \mathcal{H}$, $x_i \in \mathbb{R}^d \equiv \mathcal{X}$, $\hat{y}_i \in \{0, 1\} \equiv \mathcal{Y}$
- $\ell(\hat{y}, y)$: loss, or error, function that measures the difference between the prediction, \hat{y} , and the true label, y (e.g. 0-1 loss, squared loss, etc.)
 - $\ell_{0-1}(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$
 - $\ell_{\text{squared}}(\hat{y}, y) = (\hat{y} - y)^2$

3 Empirical Risk Minimization and Generalization Gap

The goal is to identify the hypothesis $h \in \mathcal{H}$ that gives the best performance on \mathcal{D} . If we knew \mathcal{D} then we could evaluate h via the risk:

Definition 1 (True Risk).

$$R[h] \equiv \mathbb{E}_{(x,y) \sim \mathcal{D}}[l(h(x), y)]$$

Definition 2 (Empirical Risk).

$$\hat{R}_S[h] \equiv \frac{1}{m} \sum_{i=1}^m l(h(x_i), y_i)$$

The essential task of supervised learning is to maximize the performance on all of the possible data via the adjustment of h on a particularly drawn sample set S , which is often regarded as “generalization”. The difference between performance of h on S and on \mathcal{D} is the thing we want to minimize.

Definition 3 (Generalization Gap). *Given an sample set $S = (x_i, y_i)$, $i \in \{1, \dots, m\}$, drawn i.i.d. from \mathcal{D} , a hypothesis h_S learnt on S , and a specific definition of loss l , the generalization gap is defined as*

$$\epsilon_{\text{gen}}(h_S) = |R[h_S] - \hat{R}_S[h_S]|$$

One of the most featured results of statistical learning theory is upper bounding this generalization gap, *i.e.* to find $R(h_S) \leq \hat{R}_S(h_S) + \epsilon$. Modern results bound the generalization gap tighter with the help of the properties of the specific hypotheses, while the earlier results are more general which did not take into account the properties of the hypotheses. In this lecture, we will discuss the latter.

4 Generalization Bound for Finite Hypothesis Classes

We will first introduce some tail inequalities from probability theory and use them to bound the generalization gap for a **fixed** $h \in \mathcal{H}$. Then we will use the union bound to show that the generalization gap for **any** $h \in \mathcal{H}$ can be bounded.

Lemma 4 (Markov’s Inequality). *Let Z be a non-negative random variable. Then for $\forall a > 0$,*

$$\mathbb{P}\{Z \geq a\} \leq \frac{\mathbb{E}[Z]}{a}$$

This may not be a tight bound, but it is useful to arrive at other results. Chebyshev’s inequality is one of the most famous corollaries of Markov’s inequality.

Lemma 5 (Chebyshev's Inequality). *Let X be an integrable random variable with finite expectation and finite non-zero variance. Then for $\forall a > 0$,*

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq a\} \leq \frac{\text{Var}[X]}{a^2}$$

Lemma 6 (Generic Chernoff's Bound). *Let X be a random variable. Then for $\forall t > 0$ and a constant a ,*

$$\mathbb{P}\{X \geq a\} = \mathbb{P}\{e^{tX} \geq e^{ta}\} \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

The generic Chernoff's bound is a family of upper-bounds obtained by using the monotonicity of the exponential function and Markov's inequality. Note that though each t gives a different bound, we can minimize the bound with respect to t to get the tightest upper-bound, *i.e.* $\mathbb{P}\{X \geq a\} \leq \inf_{t \geq 0} e^{-ta} \mathbb{E}[e^{tX}]$. Such probabilistic bounds that show some random variable is close to its mean with high probability are called *concentration bounds*.

Lemma 7 (Hoeffding's Lemma). *Let X be a random variable taking values in the interval $[a, b]$, with the expectation value of 0. Then for $\forall \lambda > 0$,*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

$e^{\lambda X}$ is often regarded as the moment generating function of the random variable X .

Theorem 8 (Hoeffding's Inequality). *Let Z_1, \dots, Z_m be independent random variables such that $\mathbb{P}\{a \leq Z_i \leq b\} = 1$ for $i = 1, \dots, m$. Let $\bar{Z} = \frac{1}{m} \sum_{i=1}^m Z_i$. Then, for any $\epsilon > 0$:*

$$P(|\bar{Z} - \mathbb{E}[\bar{Z}]| > \epsilon) \leq 2 \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

Proof. (See Understanding Machine Learning [1] Appendix B.4)

First, we will try to shift Z_i by its mean. Define $X_i \equiv Z_i - \mathbb{E}[Z] = Z_i - \mu$, $i \in \{1, \dots, m\}$. Denote $\bar{X} \equiv \frac{1}{m} \sum_{i=1}^m X_i$.

With this we use the Chernoff bounds and get that for $\forall \lambda > 0$,

$$\mathbb{P}\{\bar{X} \geq \epsilon\} = \mathbb{P}\{e^{\lambda \bar{X}} \geq e^{\lambda \epsilon}\} \leq \frac{\mathbb{E}[e^{\lambda \bar{X}}]}{e^{\lambda \epsilon}} = e^{-\lambda \epsilon} \cdot \mathbb{E}[e^{\lambda (\frac{1}{m} \sum_{i=1}^m X_i)}] = e^{-\lambda \epsilon} \prod_{i=1}^m \mathbb{E}[e^{\lambda X_i/m}]$$

Here X_i/m is a zero mean random variable that lives in the interval $[\frac{a-\mu}{m}, \frac{b-\mu}{m}]$. Thus we can use Hoeffding's lemma and get

$$\mathbb{P}\{\bar{X} \geq \epsilon\} \leq e^{-\lambda \epsilon + \frac{\lambda^2(b-a)^2}{8m}}$$

We then minimize the RHS *w.r.t.* λ and get

$$\min_{\lambda} e^{-\lambda \epsilon + \frac{\lambda^2(b-a)^2}{8m}} = \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

By the same argument we can show that $\mathbb{P}\{-\bar{X} \geq \epsilon\} \leq e^{-\lambda \epsilon + \frac{\lambda^2(b-a)^2}{8m}}$, and thus we get the desired result from the union of two disjoint probabilities:

$$P(|\bar{X}| > \epsilon) \leq 2 \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

□

The expression above says that for any positive ϵ , our sample mean will be at least ϵ away from its expected value with a probability that decays exponentially with the number of training examples we have.

Definition 9 (PAC Learning). A hypothesis class \mathcal{H} is (Agnostic) PAC Learnable if given some arbitrary $\epsilon, \delta > 0$ there exist an $m_{\mathcal{H}}(\epsilon, \delta)$ such that for any S with $|S| > m_{\mathcal{H}}(\epsilon, \delta)$ we have $\epsilon_{\text{gen}}(h_S) \leq \epsilon$ with probability at least $1 - \delta$. The “probably” (P) part of PAC corresponds to $1 - \delta$ while the “approximately correct” (AC) part corresponds to ϵ . $m_{\mathcal{H}}(\epsilon, \delta)$ is known as the Sample Complexity of the hypothesis class.

We will now show that any finite hypothesis class \mathcal{H} is agnostic PAC learnable. We first derive a probabilistic bound on the following distance that holds for any $h \in \mathcal{H}$:

$$\left| R[h] - \hat{R}_S[h] \right|$$

We notice that this is just the absolute distance between the empirical average $\hat{R}_S[h]$ and its mean since:

$$\mathbb{E}[\hat{R}_S[h]] = \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m \ell(h(x_i), y_i) \right] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[\ell(h(x_i), y_i)] = R[h]$$

We can use Hoeffding’s inequality to decide how many samples we would need to take to guarantee that

$$P \left(\left| \hat{R}_S[h] - R[h] \right| < \epsilon \right) > 1 - \delta$$

If we set $\delta = \exp(-2m\epsilon^2)$ we can solve to get $m = O\left(\frac{-\log(\delta)}{\epsilon^2}\right)$. Note: this is a lower bound on the sample size which guarantees the statement above, (see *Sample Complexity* in the PAC Learning definition above).

Now, suppose we consider an arbitrary $h \in \mathcal{H}$ and a loss function ℓ with range $[0, 1]$. For the random variable $\hat{R}_S[h]$ we can use Hoeffding’s inequality to get:

$$P \left(\left| \hat{R}_S[h] - R[h] \right| \geq \epsilon \right) \leq 2 \exp(-2m\epsilon^2)$$

Note: We cannot simply replace h with h_S in this bound because the loss on the data points will not be independent any more, and therefore Hoeffding inequality will not hold.

So to extend this bound for $\epsilon_{\text{gen}}(h_S)$ we will use union bound:

$$\begin{aligned} P \left(\left| \hat{R}_S[h_S] - R[h_S] \right| \geq \epsilon \right) &\leq P \left(\max_{h \in \mathcal{H}} \left| \hat{R}_S[h] - R[h] \right| > \epsilon \right) \\ &= P \left(\bigcup_{h \in \mathcal{H}} \left\{ \left| \hat{R}_S[h] - R[h] \right| > \epsilon \right\} \right) \\ &\stackrel{(a)}{\leq} \sum_{h \in \mathcal{H}} P \left(\left| \hat{R}_S[h] - R[h] \right| > \epsilon \right) \\ &= 2|\mathcal{H}| \exp(-2m\epsilon^2) \end{aligned}$$

Where (a) follows using a union bound argument. We can prove this in the case of 2 events and then use induction. In this case $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$.

If we take

$$m = O \left(\frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon^2} \right)$$

We get the desired result:

$$P \left(\left| \hat{R}_S[h_S] - R[h_S] \right| \geq \epsilon \right) \leq \delta$$

5 Uniform convergence

For ERM to work, it suffices to ensure that the empirical risk of all hypothesis in \mathcal{H} are good approximations of their true risk. In other words, we need that uniformly over all hypothesis in \mathcal{H} , the empirical risk is close to the true risk. In this section we will formalize this.

Definition 10 (ϵ -representative sample). A training set S is called ϵ -representative if:

$$\forall h \in \mathcal{H}, \quad \left| \hat{R}_S[h] - R[h] \right| \leq \epsilon$$

The following lemma shows that when we have a $\frac{\epsilon}{2}$ -representative sample the ERM learning rule (h_S) is guaranteed to be a good hypothesis.

Lemma 11. Assume that a training set S is $\frac{\epsilon}{2}$ -representative; then, any $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_S[h]$, satisfies:

$$R[h_S] \leq \min_{h \in \mathcal{H}} R[h] + \epsilon$$

Proof.

$$R[h_S] \stackrel{(a)}{\leq} \hat{R}_S[h_S] + \frac{\epsilon}{2} \stackrel{(b)}{\leq} \hat{R}_S[h] + \frac{\epsilon}{2} \stackrel{(a)}{\leq} R[h] + \frac{\epsilon}{2} + \frac{\epsilon}{2} = R[h] + \epsilon$$

where (a) comes from the ϵ -representative definition and (b) comes from the fact that h_S is a ERM learning rule. \square

This lemma shows that to be PAC learnable it suffices to prove the training set is ϵ -representative with high probability.

6 VC dimension

In the previous parts we considered finite hypothesis classes. What happens when the number of hypotheses is infinite? We cannot just apply a union bound any more. To solve this issue, we need to have more a suitable way of measuring the “complexity” of a hypothesis class other than cardinality.

Definition 12 (Shattering). A set of points S is shattered by a hypothesis class \mathcal{H} if there are hypotheses in \mathcal{H} that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in \mathcal{H} .

Definition 13 (VC dimension of \mathcal{H}). The VC dimension of a hypothesis space \mathcal{H} is the cardinality of the largest set S that can be shattered by \mathcal{H} . If arbitrarily large finite sets can be shattered by \mathcal{H} , then $\operatorname{VCdim}(\mathcal{H}) = \infty$.

Informally VC dimension is the maximum number of distinct points that a hypothesis in \mathcal{H} can correctly classify every possible labeling with zero error.

Example: (one dimensional threshold functions). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{H} = \{h_a(x) = \mathbb{1}[x \leq a] : a \in \mathbb{R}\}$. We will try to calculate the VC-dimension of this hypothesis class. To do so we need to see how many points we can label all of its configurations using threshold functions.

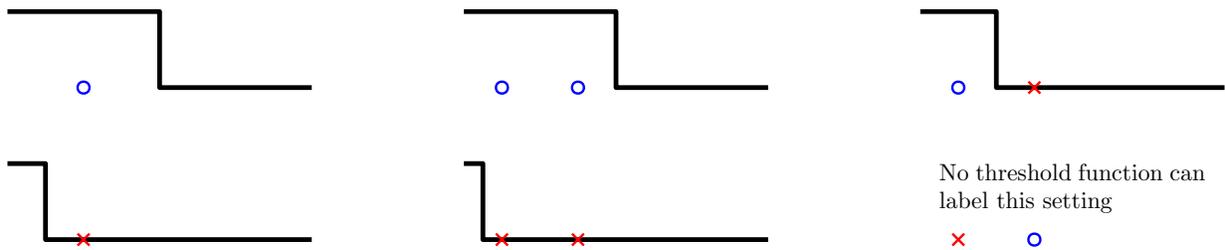


Figure 1: All the possible labelings for set of one and two points

As seen in Figure 1 our hypothesis class can label 1 points, but no set of 2 points can be labeled. Therefore the VC-dimension of this hypothesis class is 1.

References

- [1] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.