Lecture 2, October 18, 2016

Machine Learning Theory

Ruth Urner

1 Examples

Threshold functions/initial segments A simple example of an (infinite) hypothesis class is the class of *thresholds* or *initial segments* over the real line. If, for example, we would like to classify people according to whether they'd make a good athlete or not, we may base this decision on their height. To model this, we let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \mathcal{H}_{\text{init}}$, defined as

$$\mathcal{H}_{\text{init}} = \{h_a \ : \ a \in \mathbb{R}\}$$

where

$$h_a(x) \ = \ \begin{cases} 1 \text{ if } x \leq a \\ 0 \text{ if } x > a \end{cases}$$

Linear classifiers/halfspaces To classify patients as to whether they have diabetes or not, we may represent each person as a vector whose entries as are the outcomes of some medical tests. That is, we set $\mathcal{X} = \mathbb{R}^d$, for dimension d that is the number of medical test, with elements

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_d \end{pmatrix} = \begin{pmatrix} \text{blood pressure} \\ \text{blood sugar} \\ . \\ . \\ \text{temperature} \end{pmatrix}$$

For classification, we use the class the of *linear classifiers* or *halfspaces*

$$\mathcal{H}_{\text{lin}} = \{h_{\boldsymbol{w}} \ : \ w \in \mathbb{R}^d\},$$

where

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \text{sign}\left(\langle \boldsymbol{x}, \boldsymbol{w} \rangle\right) = \text{sign}\left(\sum_{i=1}^{d} x_i w_i\right)$$

Boolean halfspaces In order to classify emails according to whether they are spam or not spam, we can represent each email by a vector whose entries indicate whether the email contains a certain word or not. For this, we set $\mathcal{X} = \{0,1\}^d$, where d is the number of words in a language (say English). An email is the represented as

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_d \end{pmatrix} = \begin{pmatrix} \text{life-insurance} \\ . \\ . \\ . \\ \text{pill-generics} \end{pmatrix}$$

For classification, we use the class the of *boolean halfspaces*

$$\mathcal{H}_{\text{lin}} = \{h_{\boldsymbol{w}} \ : \ w \in \{0,1\}^d\},$$

where

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \text{sign}\left(\langle \boldsymbol{x}, \boldsymbol{w}\rangle\right) = \text{sign}\left(\sum_{i=1}^{d} x_i w_i\right)$$

2 Learnability of finite classes in the agnostic case

We now show that, for finite function classes, we also get control over the error of an empirical risk minimizer without the realizability assumption. In this case, we can not expect that the error of the output of our learner will get arbitrarily small as we see more and more samples. There may simply not be a function in the class that has a very small error. There may not even be any hypothesis of very small error. In this case, the best we can expect from a learner that is using some hypothesis class $\mathcal{H}$, is that, as it sees more and more data, the error of its output will converge to the best error possible with $\mathcal{H}$. That is, we would like the error of the output to converge to $\inf_{h\in\mathcal{H}} L(h)$. The following theorem establishes this for finite classes and also quantifies the rate of convergence. For simplicity, we assume that there is a function $h^* \in \mathcal{H}$ that attains this error, that is $L(h^*) = \inf_{h\in\mathcal{H}} L(h)$.

Theorem 1. *Let $\mathcal{H} = \{h_1, \ldots, h_N\}$ and $\delta \in (0,1]$. We have with probability at least $(1-\delta)$ over the generation of the sample S*

$$L(\hat{h}_n) \leq L(h^*) + \sqrt{\frac{2(\log 2N + \log(1/\delta))}{n}}.$$

For the proof, we require the following concentration inequality:

Theorem 2 (Hoeffding's inequality)**.** *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random variables taking values in the interval $[0,1]$. Then*

$$\mathbb{P}\{|\frac{1}{n}\sum_{i=1}^{n} Z_i \ - \ \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} Z_i]| \ > \ \epsilon\} \leq 2e^{-2\epsilon^2}$$

Proof. We have

$$\begin{aligned} L(\hat{h}_n) - L(h^*) \ &\leq \ L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^*) - L(h^*) \\ &\leq \ 2\max_{h\in\mathcal{H}} |L(h) - L_n(h)|, \end{aligned}$$

where the first inequality holds since $-L_n(\hat{h}_n) + L_n(h^*) > 0$ since $\hat{h}_n$ is a function from $\mathcal{H}$ that minimizes the empirical risk. Let $h \in \mathcal{H}$ and define random variables Z_i by $Z_i = \mathbf{1}[h(X_i) \neq Y_i]$. Then Hoeffding's inequality yields

$$\mathbb{P}\{|L(h) - L_n(h)| \ > \ \epsilon\} \leq 2e^{-2\epsilon^2}.$$

This implies

$$\begin{aligned} & \mathbb{P}\{|L(\hat{h}_n) - L_n(h^*)| \ > \ \epsilon\} \\ \leq \ & \mathbb{P}\{2 \max_{h \in \mathcal{H}} |L(h) - L_n(h)| > \epsilon\} \\ \leq \ & \mathbb{P}\{\bigvee_{h \in \mathcal{H}} |L(h) - L_n(h)| > \epsilon/2\} \\ \leq \ & |\mathcal{H}| 2e^{-2\epsilon^2}, \end{aligned}$$

where the last inequality holds by the union bound. Now, with $\epsilon = \sqrt{\frac{2(\log 2N + \log(1/\delta))}{n}}$, we get

$$\mathbb{P}\{|L(\hat{h}_n) - L_n(h^*)| \ > \ \sqrt{\frac{2(\log 2N + \log(1/\delta))}{n}}\} \leq \delta$$

which is equivalent to the statement of the theorem. □

3 (PAC) Learnability

With Theorem 1, we have seen that for finite classes, we can bound the true risk of the output of an ERM learner. This risk converges to the best possible risk in $\mathcal{H}$ as we see more an more data, that is for growing sample sizes n. The theorem also provides a concrete rate of convergence.

It is worth noting that the statement of the theorem holds without any conditions on the data generating distribution. Many times, we do not know which process generated the data. Thus, it is important to derive risk guarantees that hold regardless of what the process was. Likewise, it is important that such guarantees are quantifiable and yield meaningful bounds for finite sample sizes (rather than, for example, just holding in the limit of infinite data).

We now abstract out these requirements into the following definition of learnability.

Definition 1 (Learnability)**.** *A hypothesis class $\mathcal{H}$ is* (PAC)-learnable *if there is a learner $\mathcal{A}$ such that for all $\epsilon > 0$, for all $\delta > 0$ there is a sample size $n(\epsilon, \delta) \in \mathbb{N}$ such that for all distribution P over $\mathcal{X} \times \{0, 1\}$*

$$\mathbb{P}_{S \sim P^{n(\epsilon,\delta)}}[L(\mathcal{A}(S)) \leq \inf_{h \in \mathcal{H}} L(h) + \epsilon] \ \geq \ 1 - \delta.$$

We employ (a variant) of the notion of PAC learnability, which was introduced by Leslie Valiant [1]. PAC stands for Probably Approximately Correct. We require that the learner's output is Probably (with probability $1 - \delta$ over the sample S) Approximately (at most ϵ away) Correct (has low risk). This guarantee in terms of a sample size n needs to hold uniformly over all data generating distributions.

Theorem 1 shows that finite classes are learnable. This is easy to verify by setting $n(\epsilon, \delta) = \frac{2(\log 2N + \log(1/\delta))}{\epsilon^2}$.

4 The VC-dimension

We have seen that finite classes are learnable by any ERM learner. However, most practically relevant hypothesis classes (such as halfspaces in $\mathbb{R}^d$) are not finite classes. The example of the stubborn learner whose hypothesis class were all possible functions from the domain $\mathcal{X}$ to $\{0,1\}$ showed that not all infinite classes are learnable by any ERM. In fact, we will see that this class is not learnable at all.

We now introduce a parameter, the VC-dimension, that measures the complexity of a hypothesis class. It will turn out that finiteness of this parameter characterizes exactly whether a class is learnable or not. VC stands for Vladimir Vapnik and Alexey Chervonenkis, who introduced the notion and proved that characterization [2].

To define the VC-dimension, we need the notion of a restriction of a function class to a subset of the domain.

Definition 2. *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and let $C \subseteq \mathcal{X}$ be a finite subset of the domain. We let $\mathcal{H}_C$ denote the* restriction of $\mathcal{H}$ to C*, that is*

$$\mathcal{H}_C = \{h \in \{0,1\}^C \ : \ \text{there is } h' \in \mathcal{H} \text{ with } h'(c) = h(c) \text{ for all } c \in C\}$$

The VC-dimension measures how flexible a class is to mimic various labels in a given data samples. Recall how the stubborn learner was able to set the values of its output to 1 on points that were in the sample and to 0 on other points. The VC-dimension measure this capability. This is done via the following notion of shattering.

Definition 3 (Shattering)**.** *A class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$* shatters *a set $C \subseteq \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is equal to the set of all function from C to $\{0,1\}$, that is if $|\mathcal{H}_C| = 2^{|C|}$.*

Example 1. *It is easy to see that the class of thresholds can shatter one point $x_0 \in \mathbb{R}$. To see this, we need a function in $\mathcal{H}_{\text{init}}$ that assigns x_0 the label 0 and another function in $\mathcal{H}$ that assigns x_0 the label 1. We can take any function h_a and h_b with $a < x_0$ and $b > x_0$ to certify this.*

Note that the class of thresholds can not shatter any two distinct points $x_0 < x_1$. There is no function in $\mathcal{H}_{\text{init}}$ that assigns label 1 to x_0 and label 0 to x_1.

Definition 4. *The VC-dimension of a hypothesis class $\mathcal{H}$ is the maximal size of a set C that can be shattered by $\mathcal{H}$.*

The above examples shows that the VC-dimension of thresholds is 1.

References

[1] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[2] Vladimir N. Vapnik and Alexey J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.