Machine Learning Theory
Lecture 4

Nicholas Harvey

September 25, 2018

1 Basic Probability

Fact 1.1 (The Law of Total Expectation). For any random variable X and any event $\mathcal{E}$, we have

$$\mathrm{E}[X] \;=\; \mathrm{E}[X \mid \mathcal{E}]\Pr[\mathcal{E}] + \mathrm{E}\left[X \mid \bar{\mathcal{E}}\right]\Pr\left[\bar{\mathcal{E}}\right].$$

References. Wikipedia.

Fact 1.2 (The "Probabilistic Method"). Let Ω be a probability space A a random variable A on that space. If $\mathrm{E}[A] \geq a$, then there exists an outcome $\omega \in \Omega$ with $\Pr[\omega] > 0$ such that $A(\omega) \geq a$.

One of the first concentration bounds that you learn in probability theory is Markov's inequality. It bounds the right-tail of a random variable, using very few assumptions.

Theorem 1.3 (Markov's Inequality). Let Y be a real-valued random variable that assumes only nonnegative values. Then, for all $a > 0$,

$$\Pr[Y \geq a] \;\leq\; \frac{\mathrm{E}[Y]}{a}.$$

References. Wikipedia, [1, Equation B.3], Grimmett-Stirzaker Lemma 7.2.7, Durrett Theorem 1.6.4.

In some scenarios, we would also like to bound the left tail of a random variable. Markov's inequality can be used for this purpose if we know an upper-bound on Y. The following result is an immediate corollary of Theorem 1.3.

Corollary 1.4 (Reverse Markov Inequality). Let Y be a random variable that is never larger than b. Then, for all $a < b$,

$$\Pr[Y \leq a] \;\leq\; \frac{\mathrm{E}[b - Y]}{b - a}.$$

References. Grimmett-Stirzaker Theorem 7.3.5.

2 The No-Free-Lunch Theorem

Main point. We saw in Chapter 1 that the user's "prior knowledge" about the learning problem is reflect in the choice of hypothesis class. With a trivial hypothesis class (i.e., the set of all

functions), the learning algorithm can achieve zero training error by memorizing the training data, and intuitively, this says nothing about the true error. In this section, we make the intuitive part precise via the "No-Free-Lunch Theorem".

In this section we will assume that there is a true labeling function f, and we will use the 0-1 loss function. (So the distribution $\mathcal{D}$ is on $\mathcal{X}$, not on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.) Obviously if the learner's training data has an example for every point in $\mathcal{X}$, then it has learned f perfectly. So we can only say anything about learners that see fewer examples.

Roughly, the message of the theorem is:

- Consider *any* learning algorithm that sees a "restricted" number of examples.
- To be a good PAC-learner it is supposed to do well with respect to *all* distributions $\mathcal{D}$ on $\mathcal{X}$ and all labeling functions f.
- It is supposed to output a hypothesis h_S with

$$L_{\mathcal{D}}(h_S) \;\leq\; \underbrace{\min_{\text{any function } h} L_{\mathcal{D}}(h)}_{=L_{\mathcal{D}}(f)=0} + \epsilon$$

  for an arbitrarily small ϵ.
- But the theorem says that this does not hold: there exist $\mathcal{D}$ and f such that

$$L_{\mathcal{D}}(h_S) \;\geq\; 1/8.$$

Formally,

Theorem 2.1. Let A be any learning algorithm that receives m training samples and whose output is h_S. Suppose that $m \leq |\mathcal{X}|/2$. Then there exists a distribution $\mathcal{D}$ and $f : \mathcal{X} \to \{0,1\}$ such that $\Pr\left[\, L_{\mathcal{D}}(h_S) \geq 1/8 \,\right] \geq 1/7$.

There is nothing mysterious about $\mathcal{D}$: we will take it to be the uniform distribution on $\mathcal{X}$. In fact, we will also pick f to be a uniformly random function from $\mathcal{X}$ to $\{0,1\}$.

Proof. Let $f : \mathcal{X} \to \{0,1\}$ be a uniformly random function, so that $\{\, f(x) \,:\, x \in \mathcal{X} \,\}$ are mutually independent random variables. Recall that the training sequence is $S = \big((x_1, y_1), ..., (x_m, y_m)\big)$, where $y_i = f(x_i)$. For convenience we will define $S_{|\mathcal{X}} = \{x_1, ..., x_m\}$. Let h_S be the hypothesis output by the algorithm on input S. Then

$$\mathrm{E}_{f,S}\left[\, L_{\mathcal{D},f}(h_S) \,\right] \;=\; \mathrm{E}_{f,\,S,\,x\sim\mathcal{D}}\left[\, \mathbf{1}_{f(x)\neq h_S(x)} \,\right] \quad \text{(definition of loss)}$$

The law of total expectation (Fact 1.1) says that this equals

$$\begin{aligned} &= \mathrm{E}_{f,\,S,\,x\sim\mathcal{D}}\left[\, \mathbf{1}_{f(x)\neq h_S(x)} \;\middle|\; x \notin S_{|\mathcal{X}} \,\right] \cdot \underbrace{\Pr\left[\, x \notin S_{|\mathcal{X}} \,\right]}_{\geq 1/2} \\ &\quad + \underbrace{\mathrm{E}_{f,\,S,\,x\sim\mathcal{D}}\left[\, \mathbf{1}_{f(x)\neq h_S(x)} \;\middle|\; x \in S_{|\mathcal{X}} \,\right] \cdot \Pr\left[\, x \in S_{|\mathcal{X}} \,\right]}_{\geq 0} \end{aligned}$$

Since $|S_{|\mathcal{X}}| \leq m \leq |\mathcal{X}|/2$, we have $\Pr\left[\, x \notin S_{|\mathcal{X}} \,\right] \geq 1/2$, so

$$\begin{aligned} &\geq \underbrace{\mathrm{E}_{f,\,S,\,x\sim\mathcal{D}}\left[\, \mathbf{1}_{f(x)\neq h_S(x)} \;\middle|\; x \notin S_{|\mathcal{X}} \,\right]}_{=1/2} \cdot (1/2) \\ &= 1/4. \end{aligned}$$

Why does this last equality hold? Assuming that x is not in the training data, the algorithm can have learned nothing about $f(x)$, since the values of f are mutually independent. Thus $f(x)$ is independent of $h_S(x)$, conditioned on $x \notin S_{|\mathcal{X}}$, and the probability of disagreement is $1/2$.

To conclude, we have shown that $\mathrm{E}_f\left[\,\mathrm{E}_S\left[\, L_{\mathcal{D},f}(h_S) \,\right]\,\right] \geq 1/4$. By the probabilistic method (Fact 1.2), there exists a particular function f such that $\mathrm{E}_S\left[\, L_{\mathcal{D},f}(h_S) \,\right] \geq 1/4$. Apply the Reverse Markov Inequality (Corollary 1.4) with $Y = L_{\mathcal{D},f}(h_S)$, $b = 1$ and $a = 1/8$ to obtain

$$\Pr_S\left[\, L_{\mathcal{D},f}(h_S) \leq 1/8 \,\right] \;\leq\; \frac{\mathrm{E}\left[\, 1 - L_{\mathcal{D},f}(h_S) \,\right]}{1 - 1/8} \;\leq\; \frac{3/4}{7/8} \;=\; \frac{6}{7}.$$

Thus, $\Pr_S\left[\, L_{\mathcal{D},f}(h_S) \geq 1/8 \,\right] \geq 1/7$. ■

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.