6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 9

1. BIAS-VARIANCE TRADEOFF

From previous lecture, we see that there is a systematic way to map data into a high-dimensional space where it is linearly separable. One can then achieve zero error on the training set. Yet, we feel that this is not a good idea, as one can simply memorize the dataset and have no "predictive power", a phenomenon called "overfitting". So, was minimizing empirical fit to data in the last 8 lectures the wrong approach to learning? This is the topic of today's lecture.

What is predictive power? Hopefully, one should be able to predict better on an example that was not in the training dataset! One way to formalize the question of predictive power is to assume that the world is *static*. What we observe are data drawn i.i.d. from an unknown distribution $P_{X\times Y}$, and the goal is to find a hyperplane $x \mapsto \langle w, \phi(x)\rangle$ (or, more generally, a function $f : \mathcal{X} \mapsto \mathbb{R}$) that does well on *another* random draw (X, Y) from this distribution, an example that we had not seen before. By "doing well on (X, Y)" we mean the ability to predict the value Y given only X.

Let $\boldsymbol{\ell}(\langle w, x\rangle, y)$ denote a cost function that compares the linear prediction $\langle w, x\rangle$ to y. We are not specifying whether Y is binary, real-valued, or even multiclass – for the analysis below it does not matter.

Expected loss of a hypothesis w is

$$\mathbb{E}\boldsymbol{\ell}(\langle w, X\rangle, Y),$$

where the expectation is with respect to (X, Y) drawn from $P_{X\times Y}$. When w is computed from the training sample, we denote it by $\widehat{w} = \widehat{w}(X_1, Y_1, \ldots, X_n, Y_n)$. In fact, $\widehat{w}$ is an algorithm: it is a hyperplane that is calculated based on the training data. Since data are random, $\widehat{w}$ is random too. $\widehat{w}$ can be a solution of SGD, or exact SVM, or whatever.

We now think of the ball $\{w : \|w\| \le B\}$ as a *model* whose complexity is parametrized by B. This model may or may not capture the best that one may do in terms of explaining the relationship between X and Y. What is this best explanation? It is the one that minimizes $\mathbb{E}\boldsymbol{\ell}(\eta(X), Y)$ over all measurable functions η. We will never be able to find η since $P_{X\times Y}$ is not known to us.

The *estimation error* is defined to be

$$\mathcal{E}_{\text{est}} = \mathbb{E}\boldsymbol{\ell}(\langle \widehat{w}, X\rangle, Y) - \inf_{\|w\|\le B} \mathbb{E}\boldsymbol{\ell}(\langle w, X\rangle, Y), \tag{1}$$

where the first expectation above is over $n+1$ i.i.d. tuples $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$. The *approximation error* of the model is

$$\mathcal{E}_{\text{approx}} = \inf_{\|w\|\le B} \mathbb{E}\boldsymbol{\ell}(\langle w, X\rangle, Y) - \inf_{\eta} \mathbb{E}\boldsymbol{\ell}(\eta(X), Y), \tag{2}$$

the difference between the constrained minimization and minimization over all measurable functions $\eta : \mathcal{X} \to \mathbb{R}$. The approximation error does not depend on the learning algorithm, but rather on the choice of the model. The choice of B may be viewed as a bias-variance tradeoff. The larger the radius B, the smaller is the bias of our model, but the larger is the variance of our estimate $\widehat{w}$. The larger variance comes from having to consider many possible models on the basis of a sample of size n. This difficulty can be even seen in optimization: the radius typically enters the number of steps one needs to make to optimize a function.

The bias-variance tradeoff is a fundamental issue when one posits a model and the true minimum η of the loss might live outside the model. Model selection and the method of sieves are just some of the keywords you may search for if you are interested in learning more.

While in practice we use a penalty $\lambda\|w\|^2$, it is essentially equivalent to a restriction on $\|w\|^2$ as we've seen in an earlier lecture.

Our goal is now to understand the estimation error (1), and we will not study the approximation error any further.

1.1 Stochastic Approximation vs Sample Average Approximation

Recall from Lecture 3 that we can guarantee, for any w with $\|w\| \le B$,

$$\mathbb{E}f(\bar{w}) - f(w) \le \frac{BG}{\sqrt{T}} \tag{3}$$

after T steps of stochastic subgradient descent with $\mathbb{E}\|\nabla_t\|^2 \le G^2$. Here $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$. The expectation on the first term is *with respect to the random choices of subgradients*. There is no other randomness in the general statement of (3).

We will now apply this result to two objective functions:

$$\widehat{f}(w) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}(\langle w, X_i\rangle, Y_i) \triangleq \widehat{\mathbb{E}}\boldsymbol{\ell}(\langle w, X\rangle, Y) \tag{4}$$

and

$$f(w) = \mathbb{E}\boldsymbol{\ell}(\langle w, X\rangle, Y). \tag{5}$$

The first objective will be called the *empirical objective*, and the second — the *expected objective*. The expectation $\mathbb{E}$ in the second expression is with respect to the unknown $P_{X\times Y}$, while $\widehat{\mathbb{E}}$ is the empirical expectation with respect to the known distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}$ on the data. Since data are random, $\widehat{f}$ is a random objective. Furthermore,

$$\mathbb{E}\widehat{f} = f,$$

when we take expectation over the data $\{(X_i, Y_i)\}_{i=1}^n$.

What does SGD look like for f and $\widehat{f}$? Well, for $\widehat{f}$ we already know the answer — this is precisely what we did in the last few lectures. What was the reasoning for taking a subgradient by sampling index i uniformly at random from $\{1, \dots, n\}$? We exchange differentiation and expectation:

$$\nabla\left[\widehat{\mathbb{E}}\boldsymbol{\ell}(\langle w, X\rangle, Y)\right] = \nabla\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\ell}(\langle w, X_i\rangle, Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\nabla\boldsymbol{\ell}(\langle w, X_i\rangle, Y_i) = \widehat{\mathbb{E}}\nabla\boldsymbol{\ell}(\langle w, X\rangle, Y) \tag{6}$$

and thus *a subgradient of a randomly chosen loss function is, on average, the exact subgradient of the empirical objective.*

The general result (3) applied to $\widehat{f}$ tells us that if we sample T times $i_1, \ldots, i_T$ with replacement from the indices $\{1, \ldots, n\}$, the optimization bound is

$$\mathbb{E}_{i_1,\ldots,i_T}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\langle \bar{w}, X_i\rangle, Y_i)\right] - \inf_{\|w\|\le B}\frac{1}{n}\sum_{i=1}^{n}\ell(\langle w, X_i\rangle, Y_i) \le \frac{BG}{\sqrt{T}} \tag{7}$$

where the expectation is with respect to the uniform-with-replacement random choices of points from the dataset (*not* with respect to the data). The whole expression in (7) holds conditionally: the dataset is held fixed. Over the T steps of the optimization method, we might have encountered each example multiple times. This is ok: the argument that shows unbiasedness of the subgradients works fine for sampling with replacement.

Suppose now we perform stochastic subgradient descent for f. What should the subgradients be? Just as in (6), we may exchange expectation with differentiation

$$\nabla\left[\mathbb{E}\ell(\langle w, X\rangle, Y)\right] = \mathbb{E}\left[\nabla\ell(\langle w, X\rangle, Y)\right]. \tag{8}$$

Hence, a subgradient with respect to any (X, Y) drawn from $P_{X\times Y}$ will give a subgradient of the expected objective. Nice! Suppose we have i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then we have n subgradients to use. However, we may only pass through the data once. Prove that passing more than once breaks the above argument.

If we pass through the data once, (3) gives a guarantee for the expected loss f:

$$\mathbb{E}\ell(\langle w', X\rangle, Y) - \inf_{\|w\|\le B}\mathbb{E}\ell(\langle w, X\rangle, Y) \le \frac{BG}{\sqrt{n}} \tag{9}$$

where G is the Lipschitz constant of the loss function with respect to w, and $w' = \frac{1}{n}\sum_{t=1}^{n} w_t$ is the average of the trajectory (recall that $T = n$ here).

Important conclusion: the two procedures (one samples with replacement from the dataset and the other passes through the data once) aim to minimize different objectives. The first procedure (called *sample average approximation*, SAA) goes after the empirical minimum

$$\inf_{\|w\|\le B}\frac{1}{n}\sum_{i=1}^{n}\ell(\langle w, X_i\rangle, Y_i) \tag{10}$$

while the second (called *stochastic approximation*, SA) goes after the expected value

$$\inf_{\|w\|\le B}\mathbb{E}\ell(\langle w, X\rangle, Y). \tag{11}$$

The SAA objective is also called Empirical Risk Minimization (ERM) objective.

What happens if one passes through the data several times in a cyclic manner? What is the optimization target? Well, it is no longer SA and the method is now aiming at the empirical objective (this is not immediate and requires some work).

The next question is how far the two objectives are from each other. At a given vector w, the difference between the two curves (see Figure 1) is $f(w) - \widehat{f}(w)$. The worst such difference over the ball $\{w : \|w\| \le B\}$ is

$$\sup_{\|w\|\le B}\left\{f(w) - \widehat{f}(w)\right\}.$$

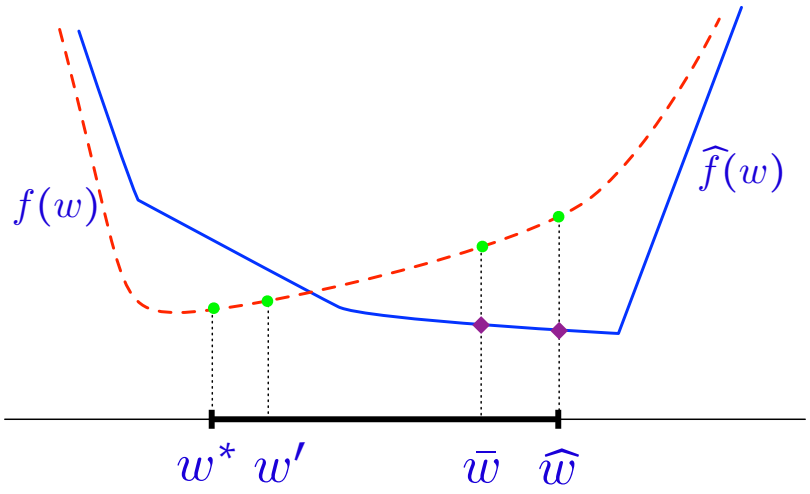

Figure 1: Expected vs empirical minimizers. Black solid region is the ball of radius B. Red dashed line is the expected objective $f(w)$. Blue solid line is the empirical objective $\widehat{f}(w)$. w^* is the minimizer of f over the ball; $\widehat{w}$ is the minimizer of $\widehat{f}$ over the ball; $\bar{w}$ is the output of an optimization procedure that aims to minimize $\widehat{f}(w)$; w' is the output of an optimization procedure that minimizes $f(w)$.

We write it as a one-sided difference, although we can also consider the two-sided version

$$\sup_{\|w\| \le B} \left| f(w) - \widehat{f}(w) \right|.$$

The supremum of the difference is a random quantity, since $(X_1, Y_1), \ldots, (X_n, Y_n)$ are random and $\widehat{f}$ is a random function. We need to decide how to measure the size of a random quantity. The most basic is through expectation:

$$\mathbb{E} \sup_{\|w\| \le B} \left\{ f(w) - \widehat{f}(w) \right\} \tag{12}$$

We shall call this expression the *uniform deviations* of empirical and expected objectives. The word *uniform* refers to the supremum over the ball. (12) is a fundamental quantity in statistics and learning theory. In fact, we know the value of (12) for the ball, up to a constant, as we show later. Further, there exist tools to study more general expressions of the form

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ \mathbb{E} g(Z) - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right\} \tag{13}$$

for some collection of functions $\mathcal{G}$.

Now, we see that the closeness of the two curves, which is given by uniform deviations, gives us a handle on how "wrong" we might be when we minimize the empirical objective instead of the desired expected objective. In fact, this is the central question of statistics. We phrase an objective in terms of an unknown distribution, yet minimize some empirical objective based on the observed data. Statistics studies how good our solution is in terms of the expected objective.

So, how does the minimizer of the empirical objective do in terms of its expected value? For the given draw of data,

$$f(\widehat{w}) - f(w^*) = f(\widehat{w}) - \widehat{f}(\widehat{w}) + \widehat{f}(\widehat{w}) - \widehat{f}(w^*) + \widehat{f}(w^*) - f(w^*). \tag{14}$$

The middle difference is non-positive since $\widehat{w}$ is a minimizer of $\widehat{f}$, and the last difference is zero in expectation over the data (remember that $\mathbb{E}\widehat{f} = f$). However, expectation of the first difference is not zero (why?). Taking expectation over data,

$$\mathbb{E}f(\widehat{w}) - f(w^*) \le \mathbb{E}\left[f(\widehat{w}) - \widehat{f}(\widehat{w})\right] \le \mathbb{E}\sup_{\|w\|\le B}\left\{f(w) - \widehat{f}(w)\right\}. \tag{15}$$

The last step uses the fact that $\widehat{w}$ is in the ball of radius B.

Conclusion: on average, the suboptimality of empirical minimizer $\widehat{w}$ in terms of its expected performance is upper bounded by the uniform deviations. Since so many quantities appear to depend on these uniform deviations, we should try to get a sense of the size of these expected suprema.

Take the particular case of hinge loss

$$\ell(\langle w, x\rangle, y) = \max\{0, 1 - y\langle w, x\rangle\}.$$

It is possible to show (we will do it in the following lectures) that uniform deviations in (12) are upper bounded as

$$\mathbb{E}\sup_{\|w\|\le B}\left\{f(w) - \widehat{f}(w)\right\} \le 2 \times \frac{BG}{\sqrt{n}} \tag{16}$$

where $G^2 \ge \mathbb{E}\|X\|^2$.

Putting everything together, the difference of the two objectives (SA vs SAA minima) in (**??**) and the bound on suboptimality of $\widehat{w}$ with respect to the expected loss f is at most

$$2 \times \frac{BG}{\sqrt{n}}. \tag{17}$$

SURPRISE: up to a factor 2, this uniform deviations upper bound is the same as the SGD guarantee for either the empirical objective (7) (with $T = n$) or the expected objective in (9). A priori, there is no reason this should have happened: one quantity is the convergence rate of SGD (which is an online procedure), and the other is uniform deviations (which is purely a question about the supremum of a certain stochastic process). Yet, the equivalence of upper bounds is not a coincidence, but rather a fundamental connection between online and offline procedures. To understand the underlying mechanisms we first need to build an arsenal of tools. We will come back to the problem at the end of the semester.

1.2 Conclusions

The reasoning below is based on comparing upper bounds we derived. In general, such a reasoning can lead to erroneous conclusions (we can point to a few examples in the literature if interested). However, for the case of hinge loss, the upper bounds are "tight" up to constants, so our conclusions are valid (unless one makes some further specific assumptions on the problem).

First, we have a guarantee that minimization of the empirical objective gives us a good predictor in the sense of the expected objective. This can be seen as a vindication of the SAA (or, ERM) principle for this problem.

Second, recall from (15) that the guarantee of the exact ERM solution in terms of its expected performance is given by (17). However, since we do not have the exact ERM

solution, but only an approximate one (obtained, e.g., as SGD solution $\bar{w}$ in (7)), one incurs an additional optimization error

$$\mathcal{E}_{\text{opt}} = f(\bar{w}) - f(\widehat{w}),$$

where $\widehat{w}$ is an exact empirical minimizer, and $\bar{w}$ is an approximate one (in Figure 1 this difference is negative, but it need not be). We see that there is no need to bring this optimization error below the level of the estimation error, which is anyway $BG/\sqrt{n}$. Hence, there is no need to choose T to be larger than $4n$.

However, why would we choose $T = 4n$ and sample *with replacement* (as dictated by SGD on the empirical objective) if the SA procedure that passes through the data once and outputs w' (see Figure 1) achieves the same error bound $BG/\sqrt{n}$. Given the bounds we derived, there is no way we can beat this! The argument speaks in favor of SA, as opposed to SAA, for the particular problem with hinge loss (the argument should not be taken as a general statement in all situations).

But the story is not over, as practitioners often make several passes through the data, which is neither SA nor SAA. An interesting research question is to quantify the performance of this method.

References