Lecture 14

Sasha Rakhlin

Oct 24, 2018

Outline

Perceptron, Continued

Bias-Variance Tradeoff

Outline

Perceptron, Continued

Bias-Variance Tradeoff

Recall from last lecture: for any T and $(x_1, y_1), \ldots, (x_T, y_T)$,

$$\sum_{t=1}^{T} \mathbf{I}\{y_t \langle w_t, x_t \rangle \leq 0\} \leq \frac{D^2}{\gamma^2}$$

where $\gamma = \gamma(x_{1:T}, y_{1:T})$ is margin and $D = D(x_{1:T}, y_{1:T}) = \max_t \|x_t\|$.

Let w^* denote the max margin hyperplane, $\|w^*\| = 1$.

Consequence for i.i.d. data (I)

Do *one pass* on i.i.d. sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ (i.e. $T = n$).

Claim: expected indicator loss of randomly picked function $x \mapsto \langle w_\tau, x \rangle$ ($\tau \sim \text{unif}(1, \ldots, n)$) is at most

$$\frac{1}{n} \times \mathbb{E}\left[\frac{D^2}{\gamma^2}\right].$$

Proof: Take expectation on both sides:

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{Y_t\langle w_t, X_t\rangle \le 0\}\right] \le \mathbb{E}\left[\frac{D^2}{n\gamma^2}\right]$$

Left-hand side can be written as (recall notation $\mathscr{S} = (X_i, Y_i)_{i=1}^n$)

$$\mathbb{E}_\tau \mathbb{E}_{\mathscr{S}} \mathbf{I}\{Y_\tau \langle w_\tau, X_\tau\rangle \le 0\}$$

Since w_τ is a function of $X_{1:\tau-1}, Y_{1:\tau-1}$, above is

$$\mathbb{E}_{\mathscr{S}} \mathbb{E}_\tau \mathbb{E}_{X,Y} \mathbf{I}\{Y\langle w_\tau, X\rangle \le 0\} = \mathbb{E}_{\mathscr{S}} \mathbb{E}_\tau \mathbf{L}_{01}(w_\tau)$$

Claim follows.

NB: To ensure $\mathbb{E}[D^2/\gamma^2]$ is not infinite, we assume margin γ in P.

Consequence for i.i.d. data (II)

Now, rather than doing one pass, cycle through data

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

repeatedly until no more mistakes (i.e. $T \le n \times (D^2/\gamma^2)$).

Then **final** hyperplane of Perceptron separates the data perfectly, i.e. finds minimum $\widehat{\mathbf{L}}_{01}(w_T) = 0$ for

$$\widehat{\mathbf{L}}_{01}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\{Y_i \langle w, X_i \rangle \le 0\}.$$

Empirical Risk Minimization with finite-time convergence. Can we say anything about future performance of w_T?

Consequence for i.i.d. data (II)

Claim: expected indicator loss of w_T is at most

$$\frac{1}{n+1} \times \mathbb{E}\left[\frac{D^2}{\gamma^2}\right]$$

Proof: Shortcuts: $z = (x, y)$ and $\ell(w, z) = \mathbf{I}\{y \langle w, x \rangle \le 0\}$. Then

$$\mathbb{E}_{\mathscr{S}} \mathbb{E}_Z \ell(w_T, Z) = \mathbb{E}_{\mathscr{S}, Z_{n+1}} \left[\frac{1}{n+1} \sum_{t=1}^{n+1} \ell(w^{(-t)}, Z_t) \right]$$

where $w^{(-t)}$ is Perceptron's final hyperplane after cycling through data $Z_1, \ldots, Z_{t-1}, Z_{t+1}, \ldots, Z_{n+1}$.

That is, *leave-one-out* is unbiased estimate of expected loss.

Now consider cycling Perceptron on $Z_1, \ldots, Z_{n+1}$ until no more errors. Let $i_1, \ldots, i_m$ be indices on which Perceptron errs in *any* of the cycles. We know $m \le D^2/\gamma^2$. However, if index $t \notin \{i_1, \ldots, i_m\}$, then whether or not Z_t was included does not matter, and Z_t is correctly classified by $\mathbf{w}^{(-t)}$. Claim follows.

This last consequence is pretty nice. Note that Bayes error $\mathbf{L}_{01}(f^*) = 0$ since we assume margin in the distribution P (and, hence, in the data). Furthermore, we have $\mathbb{E}\mathbf{L}_{01}(w_T) = O(1/n)$.

The reason we were able to achieve a nice convergence rate for $\mathbb{E}\mathbf{L}_{01}(w_T) - \mathbf{L}_{01}(f^*)$ is because we made an assumption about the distribution (recall our statement last lecture that nothing can be said about this difference if we make no assumptions).

Important: our assumption on P is not about its parametric or nonparametric form (e.g. as in the Bayesian analysis), but rather on what happens at the boundary.

SGD vs Multi-Pass

Perceptron update can be seen as gradient descent step with respect to loss

$$\max\left\{Y_t \langle w, X_t\rangle, 0\right\}$$

and step size 1.

Hence, the two consequences presented earlier can be viewed, respectively, as SGD on

$$\mathbb{E}\max\left\{Y \langle w, X\rangle, 0\right\}$$

and multi-pass cycling “SGD” on

$$\frac{1}{n}\sum_{t=1}^{n}\max\left\{Y_t \langle w, X_t\rangle, 0\right\}.$$

The first optimizes expected loss, the second optimizes empirical loss.

An aside..

Much hype in ML community is about surprising performance of deep networks *despite* overparametrization and despite fitting data perfectly.

Remark: dimension d never appears in the mistake bound of Perceptron. Hence, we can even take $\mathsf{d} = \infty$.

Note: d is number of neurons in the 1-layer neural network.

In high enough dimension, any n points in general position can be separated (zero empirical error).

Conclusions:

- More parameters than data does not necessarily contradict good prediction performance ("generalization").
- Perfectly fitting the data does not necessarily contradict good generalization.
- Complexity is a subtle notion (e.g. number of parameters vs margin)

Rates vs sample complexity

We will be making statements such as: in expectation or with high probability, expected loss is upper bounded by

$$n^{-\alpha}$$

where, typically, $\alpha = 1/2$ or $\alpha = 1$, but we will also see examples with $\alpha \in (0, 1]$.

In the previous example, $\alpha = 1$. We say that last output of Perceptron (run to convergence) has expected error rate of $O(1/n)$.

Alternatively, number of samples required to achieve error ϵ is $\epsilon^{-1/\alpha}$.

Outline

Perceptron, Continued

Bias-Variance Tradeoff

Recall that we obtained

$$\mathbb{E}\mathbf{L}(w_T) - \mathbf{L}(f^*) \leq O(1/n)$$

under the assumption that P is linearly separable with margin γ.

The assumption on the probability distribution implies that the Bayes classifier is a linear separator. Such a setup is sometimes called “realizable” because the Bayes classifier belongs to the class of functions with which we are working.

We may relax the margin assumption, yet still hope to do well with linear separators. That is, we would aim to minimize

$$\mathbb{E}\mathbf{L}_{01}(w_T) - \mathbf{L}_{01}(w^*)$$

hoping that

$$\mathbf{L}_{01}(w^*) - \mathbf{L}_{01}(f^*)$$

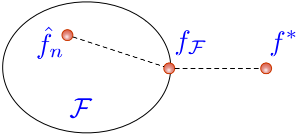
is small, where w^* is the best hyperplane classifier with respect to $\mathbf{L}_{01}$.

More generally, we will work with some class of functions $\mathcal{F}$ that we hope captures well the relationship between X and Y. The choice of $\mathcal{F}$ gives rise to a bias-variance decomposition.

Let $\mathsf{f}_{\mathcal{F}} \in \operatorname*{argmin}_{\mathsf{f} \in \mathcal{F}} \mathbf{L}(\mathsf{f})$ be the function in $\mathcal{F}$ with smallest expected loss.

Bias-Variance Tradeoff

$$\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f^*) = \underbrace{\mathbf{L}(\widehat{f}_n) - \inf_{f \in \mathcal{F}} \mathbf{L}(f)}_{\text{Estimation Error}} + \underbrace{\inf_{f \in \mathcal{F}} \mathbf{L}(f) - \mathbf{L}(f^*)}_{\text{Approximation Error}}$$



Clearly, the two terms are at odds with each other:

- Making $\mathcal{F}$ larger means smaller approximation error but (as we will see) larger estimation error
- Taking a larger sample n means smaller estimation error and has no effect on the approximation error.
- Thus, it makes sense to trade off size of $\mathcal{F}$ and n. This is called *Structural Risk Minimization*, or *Method of Sieves*, or *Model Selection*.

Bias-Variance Tradeoff

We will only focus on the estimation error, yet the ideas we develop will make it possible to read about model selection on your own.

Once again, if we guessed correctly and $f^* \in \mathcal{F}$, then

$$\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f^*) = \mathbf{L}(\widehat{f}_n) - \inf_{f \in \mathcal{F}} \mathbf{L}(f)$$

This was the case in Perceptron under margin assumption.

For a particular problem, one hopes that prior knowledge about the problem can ensure that the approximation error $\inf_{f \in \mathcal{F}} \mathbf{L}(f) - \mathbf{L}(f^*)$ is small.

Bias-Variance Tradeoff

In simple problems (e.g. linearly separable data) we might get away with having no bias-variance decomposition (e.g. as was done in the Perceptron case).

However, when we start considering more complex problems, our assumptions on P often imply that $\mathcal{F}$ is too large and the estimation error cannot be controlled. In such cases, we need to do the bias-variance decomposition.

Finally, the bias-variance tradeoff need not be in the form we just presented. For instance, we will consider a different decomposition for local rules later in the course.