Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures — Week 3

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 Overfitting Phenomenon

Goal. Find a good prediction rule $\hat{g}$ such that $\mathbb{P}\left(Y \neq \hat{g}(X)\right)$ is small.

Recall that we are given the "training data," $(X_1, Y_1), \dots, (X_n, Y_n)$, which are i.i.d. with joint distribution P. We want to find g with small binary loss

$$\Pr\left(Y \neq g(X)\right) = \mathbb{E}\left[\mathbb{I}\left(Y \neq g(X)\right)\right] \simeq \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq g(X_j)\right),$$

where the $\simeq$ means "approximate equality" due to the Law of Large Numbers (LLN). Previously, we tried to estimate the regression function itself. Another method is to try to minimize

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq \hat{g}(X_j)\right)$$

directly. If you try to do this over all measurable functions, you find that for any $\tilde{g} : S \to \{\pm 1\}$ such that $\tilde{g}(X_j) = Y_j$

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq \tilde{g}(X_j)\right) = 0.$$

So we find that we have zero loss. To further the point, consider

$$\tilde{g}(x) = \begin{cases} Y_j & x \in \{X_1, \dots, X_n\} \\ 0 & \text{otherwise} \end{cases}$$

Thus, $\mathbb{P}\left(Y_j \neq \hat{g}(X_j)\right) = 1$ for any "non-trivial" distribution (e.g. Π has a density in $\mathbb{R}^d$). This is the "overfitting phenomenon." This occurred because we initially tried to minimize over a class of functions that was too large, so we want to restrict the class. Thus, instead of minimizing the risk over all measurable functions, choose some "base class" $\mathcal{G}$ (based on the problem at hand), and consider

$$\hat{g}_n = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq g(X_j)\right).$$

1.2 Overcoming the Overfitting Phenomenon

Recall that in the previous lectures we had the training data $(X_i, Y_i), \ldots, (X_n, Y_n)$ i.i.d from a joint distribution P where every $(X, Y) \in \mathbb{S} \times \{\pm 1\}$. Note, if you wish you may think about the abstract space $\mathbb{S}$ as $\mathbb{R}^d$. We set out with the goal of finding a function $g : S \to \{\pm 1\}$ such that the loss of g, $L(g) = \mathbb{P}(Y \neq g(X))$ is "small", where "small" corresponds to having an excess risk that is sufficiently small. Recall that excess risk is

$$\mathcal{E}(g) = L(g) - L(g_*),$$

where $g_* = \operatorname{sign}(\eta)$ is the Bayes Classifier.

Last time we looked at the empirical risk minimization, where we let $\mathcal{G}$ be the base class, the collection of all functions $g : S \to \{\pm 1\}$. Now, instead define

$$\hat{g}_n = \operatorname*{argmin}_{g \in \mathcal{G}} P_n \, \mathbb{I}\left(y \neq g(x)\right) := \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq g(X_j)\right) \tag{1.1}$$

Remark 1. So what do we mean when writing $P_n g$? We treat P_n as a linear functional that acts on bounded measurable functions. To see this, let $\mathbb{Q}$ be a probability measure (a distribution). We have that

$$\mathbb{Q}f := \int f d\mathbb{Q} = \mathbb{E}\left[f(\xi)\right]$$

where $\xi \sim \mathbb{Q}$. In a more familiar setting, if $\mathbb{F}$ is a distribution function of a random variable ξ, then

$$\mathbb{E}\left[g(\xi)\right] = \int g(t) d\mathbb{F}(t) = \mathbb{F}g.$$

In what follows, this notation will help us to avoid cumbersome expressions.

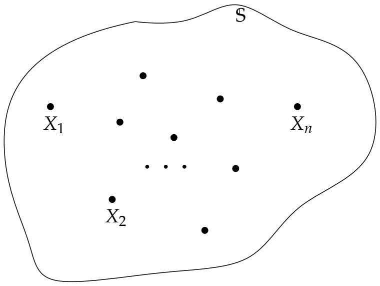

Figure 1.1. Points in our space $\mathbb{S}$.

If we have a collection of points, say $X_1, \ldots, X_n$ as shown in figure 1.1, we can assign mass to the points X_i by

$$P_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}$$

where δ_{X_j} is the Dirac delta measure concentrated at point X_j. We find that for the function f,

$$P_n f = \frac{1}{n} \sum_{j=1}^{n} f(X_j).$$

Thus, we may arrive at (1.1) by setting $f = \mathbb{I}\left(Y \neq g(X)\right)$ and taking the argmin over all $g \in \mathcal{G}$.

At the end of the last lecture we saw the effect of overfitting, where we took our base class $\mathcal{G}$ to be too large. This leads us to some empirical observations.

Empirical observations:

(a) $\mathcal{G}$ cannot be too large,

(b) $\hat{g}_n$ is difficult to calculate due to its non-linear and non-convex nature.

While $\hat{g}_n$ may be difficult to calculate, there are cases when it is possible. One of such cases is the class of so-called "decision stumps" as illustrated in the next example.

Example 1. "Decision Stumps"

Let $\mathbb{S} = \mathbb{R}$, and take the class $\mathcal{G}$ to be the class that consists of functions of the form

$$\begin{aligned} g_t^{+}(x) &= \mathbb{I}\left(x \geq t\right) - \mathbb{I}\left(x < t\right) \\ \text{and} \quad g_t^{-}(x) &= \mathbb{I}\left(x \leq t\right) - \mathbb{I}\left(x > t\right), \end{aligned}$$

which are shown in figures 1.2 and 1.3. Thus, we have $\mathcal{G} = \{g_t^{+}(\cdot), g_t^{-}(\cdot) | t \in \mathbb{R}\}$.
The question is to minimize

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left(Y_j \neq g(X_j)\right)$$

over all $g \in \mathcal{G}$.

As shown in figure 1.4, if we are to take the "order statistics" (the ordering of the X_i from the smallest $X_{(1)}$ to the largest $X_{(n)}$), we would not differentiate between any two points $X_{(i)}$ and $X_{(i+1)}$ for $i = 1, \ldots, n-1$. Thus, we only need to consider n points, and further, we may consider the n points that are exactly $X_{(1)}, \ldots, X_{(n)}$. Therefore, we have $g^{\pm}_{(X_1)}, \ldots, g^{\pm}_{(X_n)}$. Note, we would need to add an ϵ buffer to $g^{\pm}_{(X_n)}$ if we were to want to be tidy and consider boundary effects.

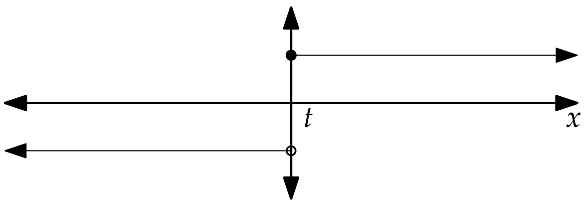

Figure 1.2. Graph of g_t^{+}.

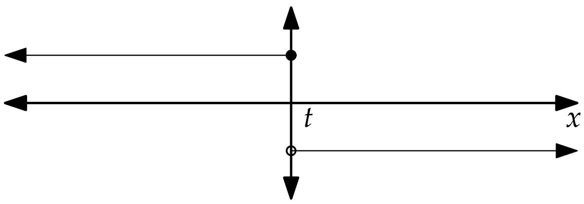

Figure 1.3. Graph of g_t^{-}.

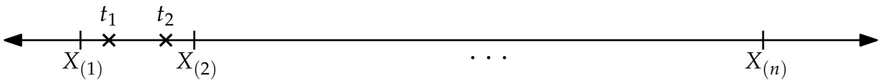

Figure 1.4. The order statistics with points marked.

Exercise 1. (Generalization) If $\mathbb{S} = \mathbb{R}^d$, consider decision stumps for each coordinate. For example, if $d = 2$ and $x = (x_1, x_2)$, we have

$$\begin{aligned} g^+_{t,1}(x) &= \mathbb{I}(x_1 \geq t) - \mathbb{I}(x_1 < t) \\ \text{and} \quad g^-_{t,2}(x) &= \mathbb{I}(x_2 \leq t) - \mathbb{I}(x_2 > t). \end{aligned}$$

and similar for $g^-_{t,1}(x)$ and $g^+_{t,2}(x)$, where $g^+_{t,i}(x)$ only depends on the i^{th} coordinate.

- In $\mathbb{R}^d$, the decision stumps cut the plane into pieces parallel to the coordinate plane.
- You can always generate decision stumps that do better than a random guess (e.g. flip of a fair coin).

Remember that we cannot minimize (1.1) since it is non-linear and non-convex. Due to this we cannot use gradient descent, or other methods to find the minimum. So instead we want to replace $\mathbb{I}(Y \neq g(X))$ in (1.1) with something "nicer."

Definition 1. We will say that a class G of binary classifiers satisfies the **weak learnability** condition if for any collection of data $(X_j, Y_j)_{j=1}^n$, $n \geq 1$ and any nonnegative weights $w_1, \ldots, w_n$, $\sum_{j=1}^n w_j = 1$, there exists $g \in G$ such that $\sum_{j=1}^n w_j \mathbb{I}(Y_j \neq g(X_j)) \leq 1/2$ (in simple terms, g "does better than a random guess").

Remark 2.

1. Weak learnability condition holds for any symmetric class G of binary classifiers, meaning that $g \in G \implies -g \in G$.
2. The Adaboost algorithm (due to R. Schapire and Y. Freund) was originally motivated by the following question: given a class G that satisfies a weak learnability condition, can one find $\hat{g}$ such that
$$P_n \mathbb{I}(y \neq \hat{g}(x)) \leq \varepsilon$$
for any $\varepsilon > 0$? For instance, such a $\hat{g}$ can be found by "combining" the elements of G.