Statistical Learning Theory

1. Introduction to statistical learning theory

Binary classification

- A large part of this class will be devoted to one of the simplest problem of statistical learning theory: binary classification.
- In this problem, we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are n independent random copies of $(X, Y) \in \mathcal{X} \times \{0, 1\}$.
- Denote by $P_{X,Y}$ the joint distribution of (X, Y).
- The so-called feature X lives in some abstract space $\mathcal{X}$ (think $\mathbb{R}^d$) and $Y \in \{0, 1\}$ is called label.
- For example, X can be a collection of gene expression levels measured on a patient and Y indicates if this person suffers from obesity.

Binary classification

- The goal of binary classification is to build a rule to predict Y given X using only the data at hand.
- Such a rule is a function $h : \mathcal{X} \to \{0, 1\}$ called a classifier.
- Some classifiers are better than others and we will favor ones that have low classification error $R(h) = \mathbb{P}(h(X) \neq Y)$.

Binary classification

- Let us make some important remarks.
- First of all, since $Y \in \{0, 1\}$ then Y has a Bernoulli distribution: so much for distribution free assumptions!
- However, we will not make assumptions on the marginal distribution of X or, what matters for prediction, the conditional distribution of Y given X.
- We write, $Y \mid X \sim \text{Ber}(\eta(X))$, where $\eta(X) = \mathbb{P}(Y = 1 \mid X) = \mathbb{E}[Y \mid X]$ is called the *regression function* of Y onto X.

Binary classification

- Next, note that we did not write $Y = \eta(X)$.
- Actually we have $Y = \eta(X) + \varepsilon$, where $\varepsilon = Y - \eta(X)$ is a "noise" random variable that satisfies $\mathbb{E}[\varepsilon \mid X] = 0$.
- In particular, this noise accounts for the fact that X may not contain enough information to predict Y perfectly.
- This is clearly the case in our genomic example above
- The noise vanishes if and only if $\eta(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$. Figure ?? illustrates the case where there is no noise and the the more realistic case where there is noise.

Binary classification

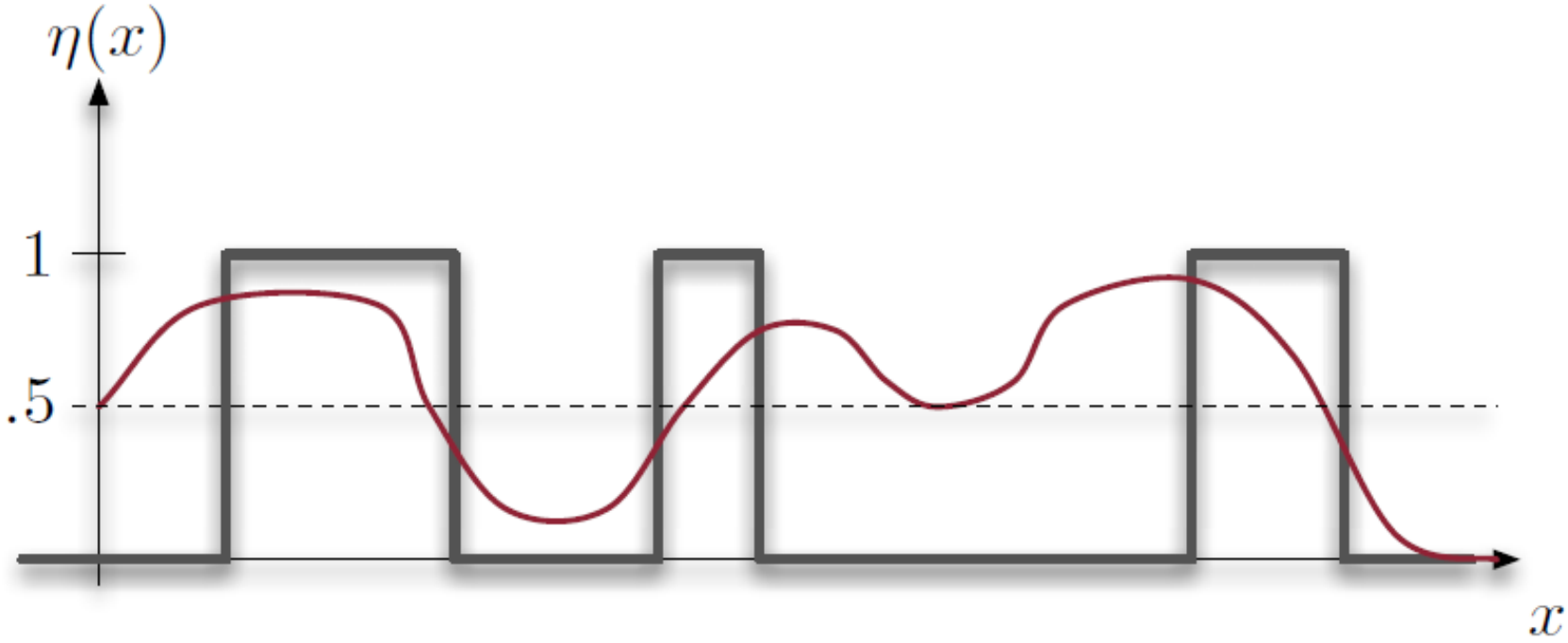

Figure 1: The thick black curve corresponds to the noiseless case where $Y = \eta(X) \in \{0, 1\}$ and the thin red curve corresponds to the more realistic case where $\eta \in [0, 1]$. In the latter case, even full knowledge of η does not guarantee a perfect prediction of Y.

Binary classification

- When $\eta(x)$ is close to .5, there is essentially no information about Y in X as the Y is determined essentially by a toss up.
- In this case, it is clear that even with an infinite amount of data to learn from, we cannot predict Y well since there is nothing to learn.

Binary classification

- In the presence of noise, since we cannot predict Y perfectly, and thus we cannot drive the classification error $R(h)$ to zero, regardless of what classifier h we use.
- What is the smallest value of $R(h)$ that can be achieved?
- As a thought experiment, assume to begin with that we have all the information that we may ever hope to get, namely we know the regression function $\eta(\cdot)$.

Binary classification

- For a given X to classify, if $\eta(X) = 1/2$ we may just toss a coin to decide our prediction and discard X since it contains no information about Y.
- However, if $\eta(X) \neq 1/2$, we have an edge over random guessing: if $\eta(X) > 1/2$, it means that $\mathbb{P}(Y = 1 \mid X) > \mathbb{P}(Y = 0 \mid X)$ or, in words, that 1 is more likely to be the correct label.
- We will see that the classifier $h^*(X) = \mathbb{I}(\eta(X) > 1/2)$ (called Bayes classifier) is actually the best possible classifier in the sense that

$$R(h^*) = \inf_{h(\cdot)} R(h)$$

where the infimum is taken over all classifiers, i.e. functions from $\mathcal{X}$ to $\{0, 1\}$.

Binary classification

- Note that unless $\eta(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$ (noiseless case), we have $R(h^*) \neq 0$.
- However, we can always look at the excess risk $\mathcal{E}(h)$ of a classifier h defined by

$$\mathcal{E}(h) = R(h) - R(h^*) \geq 0.$$

- In particular, we can hope to drive the excess risk to zero with enough observations by estimating h^* accurately.

Empirical risk

- The Bayes classifier h^*, while optimal, presents a major drawback: we cannot compute it because we do not know the regression function η.
- Instead, we have access to the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which contains some (but not all) information about η and thus h^*.
- In order to mimic the properties of h^* recall that it minimizes $R(h)$ over all h.

Empirical risk

- But the function $R(\cdot)$ is unknown since it depends on the unknown distribution $P_{X,Y}$ of (X, Y).
- We estimate it by the empirical classification error, or simply empirical risk $\hat{R}_n(\cdot)$ defined for any classifier h by

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(h\left(X_i\right) \neq Y_i\right).$$

Empirical risk

- Since $\mathbb{E}\left[\mathbb{I}\left(h\left(X_i\right) \neq Y_i\right)\right] = \mathbb{P}\left(h\left(X_i\right) \neq Y_i\right) = R(h)$, we have $\mathbb{E}\left[\hat{R}_n(h)\right] = R(h)$ so $\hat{R}_n(h)$ is an unbiased estimator of $R(h)$.
- Moreover, for any h, by the law of large numbers, we have $\hat{R}_n(h) \to R(h)$ as $n \to \infty$, almost surely.
- This indicates that if n is large enough, $\hat{R}_n(h)$ should be close to $R(h)$.

Empirical risk

- As a result, in order to mimic the performance of h^*, let us use the empirical risk minimizer $(ERM)\hat{h}$ defined to minimize $\hat{R}_n(h)$ over all classifiers h.
- This is an easy enough task: define $\hat{h}$ such $\hat{h}(X_i) = Y_i$ for all $i = 1, \ldots, n$ and $h(x) = 0$ if $x \notin \{X_1, \ldots, X_n\}$. We have $\hat{R}_n(\hat{h}) = 0$, which is clearly minimal.

Empirical risk

- The problem with this classifier is obvious: it does not generalize outside the data.
- Rather, it predicts the label 0 for any x that is not in the data.
- Similarly, we could have predicted 1 or any combination of 0 and 1 and still get $\hat{R}_n(\hat{h}) = 0$.
- Thus, it is unlikely that $\mathbb{E}[R(\hat{h})]$ will be small.

Empirical risk

Remark

- Recall that $R(h) = \mathbb{P}(h(X) \neq Y)$.
- If $\hat{h}(\cdot) = \hat{h}\left(\{(X_1, Y_1), \ldots, (X_n, Y_n)\}; \cdot\right)$ is constructed from the data, $R(\hat{h})$ denotes the conditional probability

$$R(\hat{h}) = \mathbb{P}\left(\hat{h}(X) \neq Y \mid (X_1, Y_1), \ldots, (X_n, Y_n)\right)$$

  rather than $\mathbb{P}(\hat{h}(X) \neq Y)$.
- As a result $R(\hat{h})$ is a random variable since it depends on the randomness of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- One way to view this is to observe that we compute the deterministic function $R(\cdot)$ and then plug in the random classifier $\hat{h}$.

Generative vs. Discriminative approaches

- To study the behavior of $R(\hat{h})$ (in particular when $n \to \infty$), we need certain restrictions on the distribution $P_{X,Y}$ of (X, Y).
- Unless, there exists a $P_{X,Y}$ such that $\hat{h}$ does not behave well.
- There are essentially two schools: generative and discriminative approaches.

Generative vs. Discriminative approaches

GENERATIVE

- This approach consists in restricting the set of candidate distributions $P_{X,Y}$.
- This is what is done in discriminant analysis where it is assumed that the condition distributions of X given Y (there are only two of them: one for $Y = 0$ and one for $Y = 1$) are Gaussians on $\mathcal{X} = \mathbb{R}^d$ (see for example [HTFo9] for an overview of this approach).

Generative vs. Discriminative approaches

DISCRIMINATIVE

- Rather than making assumptions directly on the distribution, one makes assumptions on what classifiers are likely to perform well (e.g. smooth decison boundaries).
- In turn, this allows to eliminate classifiers such as the one described above and that does not generalize well.
- We may make assumptions on $\eta(X)$ rather than on classifiers.

Generative vs. Discriminative approaches

- While it is important to understand both, we will focus on the discriminative approach in this class. Specifically we are going to assume that we are given a class $\mathcal{H}$ of classifiers such that $R(h)$ is small for some $h \in \mathcal{H}$.

Estimation vs. approximation

- Assume that we are given a class $\mathcal{H}$ in which we expect to find a classifier that performs well.
- This class may be constructed from domain knowledge or simply computational convenience.
- We will see some examples in the class.

Estimation vs. approximation

- For any candidate classifier $\hat{h}_n$ built from the data, we can decompose its excess risk as follows:

$$\begin{aligned}\mathcal{E}\left(\hat{h}_n\right) &= R\left(\hat{h}_n\right) - R\left(h^*\right) \\ &= \underbrace{R\left(\hat{h}_n\right) - \inf_{h \in \mathcal{H}} R(h)}_{\text{estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} R(h) - R\left(h^*\right)}_{\text{approximation error}}.\end{aligned}$$

Estimation vs. approximation

- On the one hand, estimation error accounts for the fact that we only have a finite amount of observations and thus a partial knowledge of the distribution $P_{X,Y}$.
- Hopefully we can drive this error to zero as $n \to \infty$.
- But this would not happen when $\mathcal{H}$ is too large (e.g. overfitting).
- Therefore, we need to take $\mathcal{H}$ small enough.

Estimation vs. approximation

- On the other hand, if $\mathcal{H}$ is too small, it is unlikely that we will find classifier with performance close to that of h^*.
- A tradeoff between estimation and approximation can be made by letting $\mathcal{H} = \mathcal{H}_n$ grow (but not too fast) with n.
- For now, assume that $\mathcal{H}$ is fixed. The goal of statistical learning theory is to understand how the estimation error drops to zero as a function not only of n but also of $\mathcal{H}$.

Estimation vs. approximation

- For the first argument, we will use concentration inequalities such as Hoeffding's and Bernstein's inequalities that allow us to control how close the empirical risk is to the classification error by bounding the random variable

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(h\left(X_i\right) \neq Y_i\right) - \mathbb{P}(h(X) \neq Y) \right|$$

  with high probability.

Estimation vs. approximation

- More generally we will be interested in results that allow to quantify how close the average of independent and identically distributed (i.i.d) random variables is to their common expected value.
- This can be controlled as follows.
- Define $\bar{h} \in \mathcal{H}$ to be any classifier that minimizes $R(\cdot)$ over $\mathcal{H}$ (assuming that such a classifier exist).

Estimation vs. approximation

- Then, we have

$$
\begin{aligned}
R\left(\hat{h}_n\right) - \inf_{h \in \mathcal{H}} R(h) &= R\left(\hat{h}_n\right) - R(\bar{h}) \\
&= \underbrace{\hat{R}_n\left(\hat{h}_n\right) - \hat{R}_n(\bar{h})}_{\leq 0} + R\left(\hat{h}_n\right) - \hat{R}_n\left(\hat{h}_n\right) + \hat{R}_n(\bar{h}) - R(\bar{h}) \\
&\leq \left|\hat{R}_n\left(\hat{h}_n\right) - R\left(\hat{h}_n\right)\right| + \left|\hat{R}_n(\bar{h}) - R(\bar{h})\right|.
\end{aligned}
$$

Estimation vs. approximation

- Since $\bar{h}$ is deterministic, we can use a concentration inequality to control $\left|\hat{R}_n(\bar{h}) - R(\bar{h})\right|$.
- However,

$$\hat{R}_n\left(\hat{h}_n\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\hat{h}_n\left(X_i\right) \neq Y_i\right)$$

  is not the average of independent random variables since $\hat{h}_n$ depends in a complicated manner on all of the pairs $\left(X_i, Y_i\right), i = 1, \ldots, n$.

Estimation vs. approximation

- To overcome this limitation, we often use a blunt, but surprisingly accurate tool: we "sup out" $\hat{h}_n$,

$$\left|\hat{R}_n\left(\hat{h}_n\right) - R\left(\hat{h}_n\right)\right| \leq \sup_{h \in \mathcal{H}} \left|\hat{R}_n\left(\hat{h}_n\right) - R\left(\hat{h}_n\right)\right|.$$

Estimation vs. approximation

- Controlling this supremum falls in the scope of suprema of empirical processes that we will study in quite a bit of detail.
- Clearly the supremum is smaller as $\mathcal{H}$ is smaller but $\mathcal{H}$ should be kept large enough to have good approximation properties.
- This is the tradeoff between approximation and estimation. It is also know in statistics as the bias-variance tradeoff.