Statistical Learning Theory

2. Statistical learning theory for binary classification

BINARY CLASSIFICATION

- In the previous section, we looked broadly at the problems that machine learning seeks to solve and the techniques we will cover in this course.
- Today, we will focus on one such problem, binary classification, and review some important notions that will be foundational for the rest of the course.

Bayes Classifier

- Recall the setup of binary classification: we observe a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent draws from a joint distribution $P_{X,Y}$.
- The variable Y (called the label) takes values in $\{0, 1\}$, and the variable X takes values in some space $\mathcal{X}$ representing "features" of the problem.

Bayes Classifier

- Since Y is supported on $\{0, 1\}$, the conditional random variable $Y \mid X$ is distributed according to a Bernoulli distribution.
- We write $Y \mid X \sim \text{Bernoulli}(\eta(X))$, where

$$\eta(X) = \mathbb{P}(Y = 1 \mid X) = \mathbb{E}[Y \mid X]$$

  (The function η is called the regression function.)
- We begin by defining an optimal classifier called the Bayes classifier. Intuitively, the Bayes classifier is the classifier that "knows" η -it is the classifier we would use if we had perfect access to the distribution $Y \mid X$.

(*) It will turn out that the Bayes classifier does not depend on the marginal distribution P_X of X. This is why we can focus on discriminative approaches without loss of generality.

Bayes Classifier

Definition
The Bayes classifier of X given Y, denoted h^*, is the function defined by the rule

$$h^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{if } \eta(x) \leq 1/2. \end{cases}$$

In other words, $h^*(X) = 1$ whenever $\mathbb{P}(Y = 1 \mid X) > \mathbb{P}(Y = 0 \mid X)$.

Bayes Classifier

- Our measure of performance for any classifier h (that is, any function mapping X to $\{0,1\}$) will be the classification error: $R(h)=\mathbb{P}(Y \neq h(X))$.
- The Bayes risk is the value $R^{*}=R\left(h^{*}\right)$ of the classification error associated with the Bayes classifier.
- The following theorem establishes that the Bayes classifier is optimal with respect to this metric.

Bayes Classifier

Theorem
For any classifier h, the following identity holds:

$$
\begin{aligned}
R(h)-R\left(h^*\right) &= \int_{h \neq h^*}|2 \eta(x)-1| P_x(d x) \\
&= \mathbb{E}_X\left[|2 \eta(X)-1| 1\left(h(X) \neq h^*(X)\right)\right]
\end{aligned} \tag{1}
$$

where $h=h^*$ is the (measurable) set $\{x \in \mathcal{X} \mid h(x) \neq h^*(x)\}$. In particular, since the integrand is nonnegative, the classification error R^* of the Bayes classifier is the minimizer of $R(h)$ over all classifiers h. Moreover,

$$
R\left(h^*\right)=\mathbb{E}[\min (\eta(X), 1-\eta(X))] \leq \frac{1}{2}
$$

Remark 1

- The quantity $R(h) - R(h^*)$ in the statement of the theorem above is called the excess risk of h and denoted $\mathcal{E}(h)$. ("Excess," that is, above the Bayes classifier.)
- The theorem implies that $\mathcal{E}(h) \geq 0$.

Bayes Classifier

Remark 2

- The risk of the Bayes classifier R^* equals $1/2$ if and only if $\eta(X) = 1/2$ almost surely.
- This maximal risk for the Bayes classifier occurs precisely when Y "contains no information" about the feature variable X.
- Equation (1) makes clear that the excess risk weighs the discrepancy between h and h^* according to how far η is from $1/2$.
- When η is close to $1/2$, no classifier can perform well and the excess risk is low.
- When η is far from $1/2$, the Bayes classifier performs well and we penalize classifiers that fail to do so more heavily.

Bayes Classifier

- Linear discriminant analysis attacks binary classification by putting some model on the data (i.e. generative model).
- One way to achieve this is to impose some distributional assumptions on the conditional distributions $X \mid Y = 0$ and $X \mid Y = 1$.
- We can reformulate the Bayes classifier in these terms by applying Bayes' rule:

$$\begin{aligned}\eta(x) &= \mathbb{P}(Y = 1 \mid X = x) \\ &= \frac{\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(X = x \mid Y = 0)\mathbb{P}(Y = 0)}\end{aligned}$$

(In general, when P_X is a continuous distribution, we should consider infinitesimal probabilities $\mathbb{P}(X \in dx)$.)

Bayes Classifier

- Assume that $X \mid Y = 0$ and $X \mid Y = 1$ have densities p_0 and p_1.
- Also let $\mathbb{P}(Y = 1) = \pi$ is some constant reflecting the underlying tendency of the label Y. (Typically, we imagine that π is close to $1/2$, but that need not be the case: in many applications, such as anomaly detection, $Y = 1$ is a rare event.)
- Then $h^*(X) = 1$ whenever $\eta(X) \geq 1/2$, or, equivalently, whenever
$$\frac{p_1(x)}{p_0(x)} \geq \frac{1-\pi}{\pi}$$

Bayes Classifier

- When $\pi = 1/2$, this rule amounts to reporting 1 or 0 by comparing the densities p_1 and p_0.
- For instance, in Figure 2, if $\pi = 1/2$ then the Bayes classifier reports 1 whenever $p_1 \geq p_0$, i.e., to the right of the dotted line, and 0 otherwise.
- On the other hand, when π is far from 1/2, the Bayes classifier is weighed towards the underlying bias of the label variable Y.

Bayes Classifier

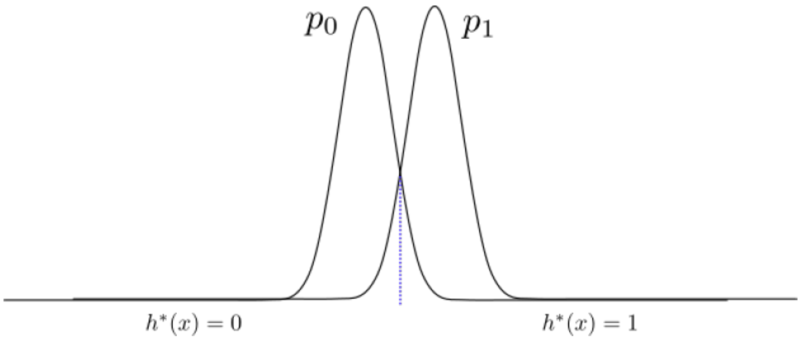

Figure 1: The Bayes classifier when $\pi = 1/2$

Empirical Risk Minimization

- The above considerations are all probabilistic, in the sense that they discuss properties of some underlying probability distribution.
- The statistician does not have access to the true probability distribution $P_{X,Y}$; she only has access to i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We consider now this statistical perspective.
- However, note that the underlying distribution $P_{X,Y}$ still appears explicitly in what follows, since that is how we measure our performance: we judge the classifiers we produced on future i.i.d. draws from $P_{X,Y}$.

- Given data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we build a classifier $\hat{h}_n(X)$, which is random in two senses: it is a function of a random variable X and also depends implicitly on the random data $\mathcal{D}_n$.
- As above, we judge a classifier according to the quantity $\mathcal{E}\left(\hat{h}_n\right)$. This is a random variable: though we have integrated out X, the excess risk still depends on the data $\mathcal{D}_n$.
- We therefore will consider bounds both on its expected value and bounds that hold in high probability.
- In any case, the bound $\mathcal{E}\left(\hat{h}_n\right) \geq 0$ always holds.

Empirical Risk Minimization

Definition
The empirical risk of a classifier h is given by

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

Empirical Risk Minimization

- Minimizing the empirical risk over the family of all classifiers is useless, since we can always minimize the empirical risk by mimicking the data and classifying arbitrarily otherwise.
- We therefore limit our attention to classifiers in a certain family $\mathcal{H}$.

Empirical Risk Minimization

Definition
The Empirical Risk Minimizer (*ERM*) over $\mathcal{H}$ is any element $\hat{h}^{\text{erm}}$ of the set $\operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_n(h)$.

(*) In fact, even an approximate solution will do: our bounds will still hold whenever we produce a classifier $\hat{h}$ satisfying $\hat{R}_n(\hat{h}) \leq \inf_{h \in \mathcal{H}} R_n(h) + \varepsilon$.

(*) ERM is one of many learning algorithms. We focus on ERM since there are well developed learning theories.

Empirical Risk Minimization

- In order for our results to be meaningful, the class $\mathcal{H}$ must be much smaller than the space of all classifiers.
- On the other hand, we also hope that the risk of $\hat{h}^{\text{erm}}$ will be close to the Bayes risk, but that is unlikely if $\mathcal{H}$ is too small.
- We will learn how to quantify this tradeoff.

Oracle Inequalities

- An oracle is a mythical classifier, one that is impossible to construct from data alone but whose performance we nevertheless hope to mimic.
- Specifically, given $\mathcal{H}$ we define $\bar{h}$ to be an element of $\operatorname{argmin}_{h \in \mathcal{H}} R(h)$ - a classifier in $\mathcal{H}$ that minimizes the true risk.
- Of course, we cannot determine $\bar{h}$, but we can hope to prove a bound of the form

$$R(\hat{h}) \leq R(\bar{h}) + \text{ something small.} \tag{2}$$

- Since $\bar{h}$ is the best minimizer in $\mathcal{H}$ given perfect knowledge of the distribution, a bound of the form given in Equation(2) would imply that $\hat{h}$ has performance that is almost best-inclass.

Oracle Inequalities

- There is a natural tradeoff between the two terms on the right-hand side of Equation (**??**).
- When $\mathcal{H}$ is small, we expect the performance of the oracle $\bar{h}$ to suffer, but we may hope to approximate $\bar{h}$ quite closely.

(*) Indeed, at the limit where $\mathcal{H}$ is a single function, the "something small" in Equation (2) is equal to zero.

Oracle Inequalities

- On the other hand, as $\mathcal{H}$ grows the oracle will become more powerful but approximating it becomes more statistically difficult.
- In other words, we need a larger sample size to achieve the same measure of performance.
- Since $R(\hat{h})$ is a random variable, we ultimately want to prove a bound in expectation or tail bound of the form

$$\mathbb{P}\left(R(\hat{h}) \leq R(\bar{h}) + \Delta_{n,\delta}(\mathcal{H})\right) \geq 1 - \delta$$

  where $\Delta_{n,\delta}(\mathcal{H})$ is some explicit term depending on our sample size and our desired level of confidence.

Oracle Inequalities

- In the end, we should recall that

$$\mathcal{E}(\hat{h}) = R(\hat{h}) - R\left(h^*\right) = (R(\hat{h}) - R(\bar{h})) + \left(R(\bar{h}) - R\left(h^*\right)\right).$$

- The second term in the above equation is the approximation error, which is unavoidable once we fix the class $\mathcal{H}$.
- Oracle inequalities give a means of bounding the first term, the stochastic error.

Hoeffding's Theorem

Theorem (Hoeffding's Theorem)
Let $X_1, \ldots, X_n$ be n independent random variables such that $X_i \in [0,1]$ almost surely. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i\right| > t\right) \leq 2e^{-2nt^2}$$

- In other words, deviations from the mean decay exponentially fast in n and t.

Maximal inequality

- Hoeffding's Theorem implies that, for any classifier h, the bound
$$\left|\hat{R}_n(h) - R(h)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$
holds with probability $1 - \delta$.
- If $\mathcal{H}$ is a finite family, i.e., $\mathcal{H} = \{h_1, \ldots, h_M\}$, then with probability $1 - \delta/M$ the bound
$$\left|\hat{R}_n\left(h_j\right) - R\left(h_j\right)\right| \leq \sqrt{\frac{\log(2M/\delta)}{2n}}$$
holds.

Maximal inequality

- The event that $\max_j \left|\hat{R}_n(h_j) - R(h_j)\right| > t$ is the union of the events $\left|\hat{R}_n(h_j) - R(h_j)\right| > t$ for $j = 1, \ldots, M$, so the union bound immediately implies that

$$\max_j \left|\hat{R}_n(h_j) - R(h_j)\right| \leq \sqrt{\frac{\log(2M/\delta)}{2n}}$$

  with probability $1 - \delta$.
- The logarithmic dependence on M implies that we can increase the size of the family $\mathcal{H}$ exponentially quickly with n and maintain the same guarantees on our estimate.

Learning with a finite dictionary

- Assume $|\mathcal{H}| = M$.
- Let $\hat{h}$ be

$$\hat{h} \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_n(h)$$

- Let $\bar{h}$ be

$$\bar{h} \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} R(h).$$

Learning with a finite dictionary

Theorem
The estimator $\hat{h}$ *satisfies*

$$R(\hat{h}) \leq R(\bar{h}) + \sqrt{\frac{2\log(2M/\delta)}{n}}$$

with probability at least $1 - \delta$.

(*) It can be shown that

$$\mathbb{E}[R(\hat{h})] \leq R(\bar{h}) + \sqrt{\frac{2\log(2M)}{n}}$$

Proof

From the definition of $\hat{h}$, we have $\hat{R}_n(\hat{h}) \leq \hat{R}_n(\bar{h})$, which gives

$$R(\hat{h}) \leq R(\bar{h}) + \left[\hat{R}_n(\bar{h}) - R(\bar{h})\right] + \left[R(\hat{h}) - \hat{R}_n(\hat{h})\right]$$

The only term here that we need to control is the second one, but since we don't have any real information about $\bar{h}$, we will bound it by a maximum over $\mathcal{H}$ and then apply Hoeffding:

$$\left[\hat{R}_n(\bar{h}) - R(\bar{h})\right] + \left[R(\hat{h}) - \hat{R}_n(\hat{h})\right]$$

$$\leq 2 \max_j \left|\hat{R}_n(h_j) - R(h_j)\right| \leq 2\sqrt{\frac{\log(2M/\delta)}{2n}}$$

with probability at least $1 - \delta$, which completes the proof.