18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Jonathan Weed

Lecture 2
Sep. 14, 2015

# Part I
# Statistical Learning Theory

## 1. BINARY CLASSIFICATION

In the last lecture, 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.

Our present focus on the problem of binary classification is justified because both binary classification encompasses much of what we want to accomplish in practice and because the response variables in the binary classification problem are bounded. (We will see a very important application of this fact below.) It also happens that there are some nasty surprises in non-binary classification, which we avoid by focusing on the binary case here.

### 1.1 Bayes Classifier

Recall the setup of binary classification: we observe a sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ 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. We can of course speak of the marginal distribution $P_X$ of $X$ alone; moreover, since $Y$ is supported on $\{0, 1\}$, the conditional random variable $Y|X$ is distributed according to a Bernoulli distribution. We write $Y|X \sim \text{Bernoulli}(\eta(X))$, where

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

(The function $\eta$ 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" $\eta$—it is the classifier we would use if we had perfect access to the distribution $Y|X$.

**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 $\mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X)$.

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 $R^* = R(h^*)$ of the classification error associated with the Bayes classifier. The following theorem establishes that the Bayes classifier is optimal with respect to this metric.

**Theorem:** For any classifier $h$, the following identity holds:

$$R(h) - R(h^*) = \int_{h \neq h^*} |2\eta(x) - 1| \, P_x(dx) = \mathbb{E}_X[|2\eta(X) - 1| \mathbf{1}(h(X) \neq h^*(X))] \quad (1.1)$$

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,

$$R(h^*) = \mathbb{E}[\min(\eta(X), 1 - \eta(X))] \leq \frac{1}{2}. \quad (1.2)$$

*Proof.* We begin by proving Equation (1.2). The definition of $R(h)$ implies

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y = 1, h(X) = 0) + \mathbb{P}(Y = 0, h(X) = 1),$$

where the second equality follows since the two events are disjoint. By conditioning on $X$ and using the tower law, this last quantity is equal to

$$\mathbb{E}[\mathbb{E}[\mathbf{1}(Y = 1, h(X) = 0)|X]] + \mathbb{E}[\mathbb{E}[\mathbf{1}(Y = 0, h(X) = 1)|X]]$$

Now, $h(X)$ is measurable with respect to $X$, so we can factor it out to yield

$$\mathbb{E}[\mathbf{1}(h(X) = 0)\eta(X) + \mathbf{1}(h(X) = 1)(1 - \eta(X))]], \quad (1.3)$$

where we have replaced $\mathbb{E}[Y|X]$ by $\eta(X)$.

In particular, if $h = h^*$, then Equation 1.3 becomes

$$\mathbb{E}[\mathbf{1}(\eta(X) \leq 1/2)\eta(X) + \mathbf{1}(\eta(x) > 1/2)(1 - \eta(X))].$$

But $\eta(X) \leq 1/2$ implies $\eta(X) \leq 1 - \eta(X)$ and conversely, so we finally obtain

$$\begin{aligned} R(h^*) &= \mathbb{E}[\mathbf{1}(\eta(X) \leq 1/2)\eta(X) + \mathbf{1}(\eta(x) > 1/2)(1 - \eta(X))] \\ &= \mathbb{E}[(\mathbf{1}(\eta(X) \leq 1/2) + \mathbf{1}(\eta(x) > 1/2)) \min(\eta(X), 1 - \eta(X))] \\ &= \mathbb{E}[\min(\eta(X), 1 - \eta(X))], \end{aligned}$$

as claimed. Since $\min(\eta(X), 1 - \eta(X)) \leq 1/2$, its expectation is also certainly at most $1/2$ as well.

Now, given an arbitrary $h$, applying Equation 1.3 to both $h$ and $h^*$ yields

$$\begin{aligned} R(h) - R(h^*) = \mathbb{E}[&\mathbf{1}(h(X) = 0)\eta(X) + \mathbf{1}(h(X) = 1)(1 - \eta(X))] \\ &-\mathbf{1}(h^*(X) = 0)\eta(X) + \mathbf{1}(h^*(X) = 1)(1 - \eta(X))]], \end{aligned}$$

which is equal to

$$\mathbb{E}[(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))\eta(X) + (\mathbf{1}(h(X) = 1) - \mathbf{1}(h^*(X) = 1))(1 - \eta(X))].$$

Since $h(X)$ takes only the values 0 and 1, the second term can be rewritten as $-(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))$. Factoring yields

$$\mathbb{E}[(2\eta(X) - 1)(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))].$$

The term $\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0)$ is equal to $-1$, 0, or 1 depending on whether $h$ and $h^*$ agree. When $h(X) = h^*(X)$, it is zero. When $h(X) \neq h^*(X)$, it equals 1 whenever $h^*(X) = 0$ and $-1$ otherwise. Applying the definition of the Bayes classifier, we obtain

$$\mathbb{E}[(2\eta(X) - 1)\mathbf{1}(h(X) \neq h^*(X)) \operatorname{sign}(\eta - 1/2)] = \mathbb{E}[|2\eta(X) - 1|\mathbf{1}(h(X) \neq h^*(X))],$$

as desired. □

We make several remarks. First, 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 $\mathcal{E}(h) \geq 0$.

Second, 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 (1.1) makes clear that the excess risk weighs the discrepancy between $h$ and $h^*$ according to how far $\eta$ is from $1/2$. When $\eta$ is close to $1/2$, no classifier can perform well and the excess risk is low. When $\eta$ is far from $1/2$, the Bayes classifier performs well and we penalize classifiers that fail to do so more heavily.

As noted last time, linear discriminant analysis attacks binary classification by putting some model on the data. One way to achieve this is to impose some distributional assumptions on the conditional distributions $X|Y = 0$ and $X|Y = 1$.

We can reformulate the Bayes classifier in these terms by applying Bayes' rule:

$$\eta(x) = \mathbb{P}(Y = 1|X = x) = \frac{\mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0)}.$$

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

Assume that $X|Y = 0$ and $X|Y = 1$ have densities $p_0$ and $p_1$, and $\mathbb{P}(Y = 1) = \pi$ is some constant reflecting the underlying tendency of the label $Y$. (Typically, we imagine that $\pi$ 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

$$\frac{p_1(x)}{p_0(x)} \geq \frac{1 - \pi}{\pi}.$$

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 1, 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 $\pi$ is far from $1/2$, the Bayes classifier is weighed towards the underlying bias of the label variable $Y$.

### 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. 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 $P_{X,Y}$.

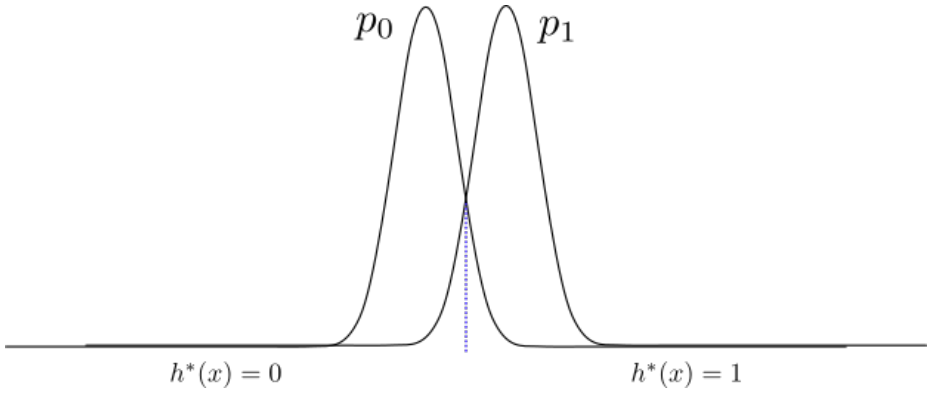


Figure 1: The Bayes classifier when $\pi = 1/2$.

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 $\mathcal{E}(\hat{h}_n)$. 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 $\mathcal{E}(\hat{h}_n) \geq 0$ always holds. (This inequality does not merely hold "almost surely," since we proved that $R(h) \geq R(h^*)$ uniformly over all choices of classifier $h$.)

Last time, we proposed two different philosophical approaches to this problem. In particular, generative approaches make distributional assumptions about the data, attempt to learn parameters of these distributions, and then plug the resulting values into the model. The discriminative approach—the one taken in machine learning—will be described in great detail over the course of this semester. However, there is some middle ground, which is worth mentioning briefly. This middle ground avoids making explicit distributional assumptions about $X$ while maintaining some of the flavor of the generative model.

The central insight of this middle approach is the following: since by definition $h^*(x) = \mathbf{1}(\eta(X) > 1/2)$, we estimate $\eta$ by some $\hat{\eta}_n$ and thereby produce the estimator $\hat{h}_n = \mathbf{1}(\hat{\eta}_n(X) > 1/2)$. The result is called a *plug-in estimator.*

Of course, achieving good performance with a plug-in estimator requires some assumptions. (No-free-lunch theorems imply that we can't avoid making an assumption somewhere!) One possible assumption is that $\eta(X)$ is smooth; in that case, there are many nonparamteric regression techniques available (Nadaraya-Watson kernel regression, wavelet bases, etc.).

We could also assume that $\eta(X)$ is a function of a particular form. Since $\eta(X)$ is only supported on $[0, 1]$, standard linear models are generally inapplicable; rather, by applying the logit transform we obtain *logistic regression*, which assumes that $\eta$ satisfies an identity of the form

$$\log\left(\frac{\eta(X)}{1-\eta(X)}\right) = \theta^T X.$$

Plug-in estimators are called "semi-paramteric" since they avoid making any assumptions about the distribution of $X$. These estimators are widely used because they perform fairly well in practice and are very easy to compute. Nevertheless, they will not be our focus here.

In what follows, we focus here on the discriminative framework and empirical risk minimization. Our benchmark continues to be the risk function $R(h) = \mathbb{E}\mathbf{1}(Y \neq h(X))$, which

is clearly not computable based on the data alone; however, we can attempt to use a naïve statistical "hammer" and replace the expectation with an average.

**Definition:** The *empirical risk* of a classifier $h$ is given by

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

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 $\mathcal{H}$.

**Definition:** The *Empirical Risk Minimizer (ERM)* over $\mathcal{H}$ is any element[1] $\hat{h}^{\mathrm{erm}}$ of the set $\operatorname{argmin}_{h\in\mathcal{H}} \hat{R}_n(h)$.

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 $\hat{h}^{\mathrm{erm}}$ will be close to the Bayes risk, but that is unlikely if $\mathcal{H}$ is too small. The next section will give us tools for quantifying this tradeoff.

### 1.3 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

$$R(\hat{h}) \leq R(\bar{h}) + \text{something small}. \tag{1.4}$$

Since $\bar{h}$ is the best minimizer in $\mathcal{H}$ given perfect knowledge of the distribution, a bound of the form given in Equation 1.4 would imply that $\hat{h}$ has performance that is almost best-in-class. We can also apply such an inequality in the so-called *improper learning* framework, where we allow $\hat{h}$ to lie in a slightly larger class $\mathcal{H}' \supset \mathcal{H}$; in that case, we still get nontrivial guarantees on the performance of $\hat{h}$ if we know how to control $R(\bar{h})$

There is a natural tradeoff between the two terms on the right-hand side of Equation 1.4. 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 1.4 is equal to zero.) 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

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

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

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

In the end, we should recall that

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

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.

### 1.4 Hoeffding's Theorem

Our primary building block is the following important result, which allows us to understand how closely the average of random variables matches their expectation.

**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 $t$.

*Proof.* Define centered random variables $Z_i = X_i - \mathbb{E}X_i$. It suffices to show that

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-2nt^2},$$

since the lower tail bound follows analogously. (Exercise!)

We apply Chernoff bounds. Since the exponential function is an order-preserving bijection, we have for any $s > 0$

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) = \mathbb{P}\left(\exp\left(s\sum Z_i\right) > e^{stn}\right) &\leq e^{-stn}\mathbb{E}[e^{s\sum Z_i}] \quad \text{(Markov)} \\ &= e^{-stn}\prod \mathbb{E}[e^{sZ_i}], \qquad (1.5)\end{aligned}$$

where in the last equality we have used the independence of the $Z_i$.

We therefore need to control the term $\mathbb{E}[e^{sZ_i}]$, known as the *moment-generating function* of $Z_i$. If the $Z_i$ were normally distributed, we could compute the moment-generating function analytically. The following lemma establishes that we can do something similar when the $Z_i$ are bounded.

**Lemma (Hoeffding's Lemma):** If $Z \in [a, b]$ almost surely and $\mathbb{E}Z = 0$, then

$$\mathbb{E}e^{sZ} \leq e^{\frac{s^2(b-a)^2}{8}}.$$

*Proof of Lemma.* Consider the log-moment generating function $\psi(s) = \log \mathbb{E}[e^{sZ}]$, and note that it suffices to show that $\psi(s) \leq s^2(b-a)^2/8$. We will investigate $\psi$ by computing the

first several terms of its Taylor expansion. Standard regularity conditions imply that we can interchange the order of differentiation and integration to obtain

$$\psi'(s) = \frac{\mathbb{E}[Ze^{sZ}]}{\mathbb{E}[e^{sZ}]},$$
$$\psi''(s) = \frac{\mathbb{E}[Z^2e^{sZ}]\mathbb{E}[e^{sZ}] - \mathbb{E}[Ze^{sZ}]^2}{\mathbb{E}[e^{sZ}]^2} = \mathbb{E}\left[Z^2\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}\right] - \left(\mathbb{E}\left[Z\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}\right]\right)^2.$$

Since $\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}$ integrates to 1, we can interpret $\psi''(s)$ as the variance of $Z$ under the probability measure $d\mathbb{F} = \frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}d\mathbb{E}$. We obtain

$$\psi''(s) = \mathrm{var}_{\mathbb{F}}(Z) = \mathrm{var}_{\mathbb{F}}\left(Z - \frac{a+b}{2}\right),$$

since the variance is unaffected under shifts. But $|Z - \frac{a+b}{2}| \leq \frac{b-a}{2}$ almost surely since $Z \in [a,b]$ almost surely, so

$$\mathrm{var}_{\mathbb{F}}\left(Z - \frac{a+b}{2}\right) \leq \mathbb{F}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}.$$

Finally, the fundamental theorem of calculus yields

$$\psi(s) = \int_0^s \int_0^u \psi''(u)\,du \leq \frac{s^2(b-a)^2}{8}.$$

This concludes the proof of the Lemma. □

Applying Hoeffding's Lemma to Equation (1.5), we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-stn}\prod e^{s^2/8} = e^{ns^2/8 - stn},$$

for any $s > 0$. Plugging in $s = 4t > 0$ yields

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-2nt^2},$$

as desired. □

Hoeffding's Theorem implies that, for any classifier $h$, the bound

$$|\hat{R}_n(h) - R(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

holds with probability $1 - \delta$. We can immediately apply this formula to yield a maximal inequality: if $\mathcal{H}$ is a finite family, i.e., $\mathcal{H} = \{h_1, \ldots, h_M\}$, then with probability $1 - \delta/M$ the bound

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

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

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

with probability $1-\delta$. In other words, for such a family, we can be assured that the empirical risk and the true risk are close. Moreover, 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.

18.657 Mathematics of Machine Learning
Fall 2015