2

Binary Classification

In this lecture we begin the study of statistical learning theory in the case of binary classification. Binary classifiers assign data to one of two classes. Even though we are often interested in more than two classes (for example, in digit identification), the insights gained from the binary case generalize easily.

Binary Classification

A **binary classifier** is a function

$$h\colon \mathcal{X} \to \mathcal{Y} = \{0, 1\},$$

where $\mathcal{X}$ is a space of features. The fact that we use $\{0, 1\}$ it not very important, and in many cases it will be convenient to consider classifiers taking values in $\{-1, 1\}$. Binary classifiers arise in a variety of applications. In medical diagnostics, for example, a classifier could take an image of a skin mole and determine if it is benign or if it is melanoma. A classifier can arise from a function $\mathcal{X} \to [0, 1]$ that assigns to every input $\boldsymbol{x}$ a probability p. If $p > 1/2$, then $\boldsymbol{x}$ is assigned to class 1 and otherwise to 0.

In the context of binary classification, we usually use the **unit loss function**

$$L(h(\boldsymbol{x}), y) = \mathbf{1}\{h(\boldsymbol{x}) \neq y\} = \begin{cases} 1 & h(\boldsymbol{x}) \neq y \\ 0 & h(\boldsymbol{x}) = y. \end{cases}$$

The unit loss does not distinguish between **false positives** and **false negatives**. A false positive is a pair $(\boldsymbol{x}, y)$ with $h(\boldsymbol{x}) = 1$ but $y = 0$, and a false negative is a pair for which $h(\boldsymbol{x}) = 0$ but $y = 1$. We would like to *learn* a classifier from observations

$$\{(\boldsymbol{x}^i, y^i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}. \tag{2.1}$$

The classifier should not only match the data, but **generalize** in order to be able to classify unseen data. For this, we assume that the points in (2.1) are drawn from a probability distribution on $\mathcal{X} \times \mathcal{Y}$, and replace each data point $(\boldsymbol{x}^i, y^i)$ in (2.1) with a

copy (X_i, Y_i) of a pair of random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$. We are after a classifier h such that the expected value of the **empirical risk**

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(X_i) \neq Y_i\} \tag{2.2}$$

is small. We can write this expectation as

$$\begin{aligned} \mathbb{E}[\hat{R}(h)] &\overset{(1)}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{1}\{h(X_i) \neq Y_i\}] \\ &\overset{(2)}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(h(X_i) \neq Y_i) \\ &\overset{(3)}{=} \mathbb{P}(h(X) \neq Y) =: R(h), \end{aligned}$$

where (1) uses the linearity of expectation, (2) expresses the expectation of an indicator function as probability, and (3) uses the fact that all the (X_i, Y_i) are identically distributed. The function $R(h)$ is called the **generalization risk** or simply **risk**: it is the probability that the classifier gets something wrong.

Example 2.1. Assume that the distribution is such that Y is completely determined by X, that is, $Y = f(X)$. Then

$$R(h) = \mathbb{P}(h(X) \neq f(X)),$$

and $R(h) = 0$ if $h = f$ almost everywhere. If $\mathcal{X}$ is a finite or compact set with the uniform distribution, then $R(h)$ simply measures the proportion of the input space on which h fails to classify inputs correctly.

While for certain tasks such as image classification there may be a unique label to each input, in general this need not be the case. In many applications, the input does not carry enough information to completely determine the output. Consider, for example, the case where $\mathcal{X}$ consists of whole genome sequences and the task is to predict hypertension (or any other condition) from it. The genome clearly does not carry enough information to make an accurate prediction, as other factors also play a role. To account for this lack of information, define the **regression function**

$$f(X) = \mathbb{E}[Y|X] = 1 \cdot \mathbb{P}(Y = 1|X) + 0 \cdot \mathbb{P}(Y = 0|X) = \mathbb{P}(Y = 1|X).$$

It is sometimes convenient to separate the dependency of Y on X into a deterministic part and random noise,

$$Y = f(X) + E.$$

In this cases we have $\mathbb{E}[E|X] = 0$, because

$$f(X) = \mathbb{E}[Y|X] = \underbrace{\mathbb{E}[f(X)|X]}_{=f(X)} + \mathbb{E}[E|X].$$

The Bayes classifier

While in Example 2.1 we could choose (at least in principle) $h(\boldsymbol{x}) = f(\boldsymbol{x})$ and get $R(h) = 0$, in the presence of noise this is not possible. However, we can define a classifier by setting

$$h^*(\boldsymbol{x}) = \begin{cases} 1 & f(\boldsymbol{x}) > \frac{1}{2} \\ 0 & f(\boldsymbol{x}) \leq \frac{1}{2}, \end{cases}$$

We call this the **Bayes classifier**. Note that the Bayes classifier can be interpreted as a maximum a posteriori (MAP) estimator:

$$h^*(\boldsymbol{x}) = \arg\max_y \ \mathbb{P}(Y = y | X = \boldsymbol{x}).$$

The following result shows that this is the best possible classifier.

Theorem 2.2. *Let $h \colon \mathcal{X} \to \mathcal{Y}$ be any binary classifier. Then*

$$R(h) - R(h^*) = \mathbb{E}[|2f(X) - 1| \cdot \mathbf{1}\{h(X) \neq h^*(X)\}].$$

In particular, the Bayes classifier h^* satisfies

$$R(h^*) = \inf_h R(h),$$

where the infimum is over all measurable h. Moreover, $R(h^*) \leq 1/2$.

The difference $\mathcal{E}(h) = R(h) - R(h^*) \geq 0$ is called the **excess risk** of h.

Proof. Let h be any classifier. To compute the risk $R(h)$, we first condition on X and then average over X:

$$R(h) = \mathbb{E}[\mathbf{1}\{h(X) \neq Y\}] = \mathbb{E}\left[\mathbb{E}[\mathbf{1}\{h(X) \neq Y\}|X]\right]. \tag{2.3}$$

For the inner expectation, we have

$$\begin{aligned} \mathbb{E}[\mathbf{1}\{h(X) \neq Y\}|X] &= \mathbb{E}[\mathbf{1}\{h(X) = 1, Y = 0\} + \mathbf{1}\{h(X) = 0, Y = 1\}|X] \\ &= \mathbb{E}[(1 - Y)\mathbf{1}\{h(X) = 1\}|X] + \mathbb{E}[Y\mathbf{1}\{h(X) = 0\}|X] \\ &\overset{(1)}{=} \mathbf{1}\{h(X) = 1\}\,\mathbb{E}[(1 - Y)|X] + \mathbf{1}\{h(X) = 0\}\,\mathbb{E}[Y|X] \\ &= \mathbf{1}\{h(X) = 1\}(1 - f(X)) + \mathbf{1}\{h(X) = 0\}f(X). \end{aligned}$$

To see why (1) holds, recall that the random variable $\mathbb{E}[Y\mathbf{1}\{h(X) = 0\}|X]$ takes values $\mathbb{E}[Y\mathbf{1}\{h(x) = 0\}|X = \boldsymbol{x}]$, and will therefore only be non-zero if $h(x) = 0$. We can therefore pull the indicator function out of the expectation. Hence, using (2.3),

$$\begin{aligned} R(h) &= \mathbb{E}[\mathbf{1}\{h(X) = 1\}(1 - f(X)) + \mathbf{1}\{h(X) = 0\}f(X)] \\ &= \mathbb{E}[1 - f(X) + (2f(X) - 1) \cdot \mathbf{1}\{h(X) = 0\}]. \end{aligned} \tag{2.4}$$

Therefore, for the difference $R(h) - R(h^*)$ we get

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

Going through all the possible combinations of values $(h(\boldsymbol{x}), h^*(\boldsymbol{x})) \in \{0, 1\}^2$, we arrive at the case distinction

$$\mathbf{1}\{h(\boldsymbol{x}) = 0\} - \mathbf{1}\{h^*(\boldsymbol{x}) = 0\} = \begin{cases} 1 & \text{if } h(\boldsymbol{x}) = 0, h^*(\boldsymbol{x}) = 1 \\ -1 & \text{if } h(\boldsymbol{x}) = 1, h^*(\boldsymbol{x}) = 0 \\ 0 & \text{if } h(\boldsymbol{x}) = h^*(\boldsymbol{x}) \end{cases}$$

Moreover, $h^*(\boldsymbol{x}) = 1$ if and only if $2f(\boldsymbol{x}) - 1 > 0$, and $h^*(\boldsymbol{x}) = 0$ if and only if $2f(\boldsymbol{x}) - 1 \leq 0$. Hence,

$$(2f(\boldsymbol{x}) - 1) \cdot (\mathbf{1}\{h(\boldsymbol{x}) = 0\} - \mathbf{1}\{h^*(\boldsymbol{x}) = 0\}) = |2f(\boldsymbol{x}) - 1| \cdot \mathbf{1}\{h(\boldsymbol{x}) \neq h^*(\boldsymbol{x})\},$$

and taking the expectation gives the desired characterization. From (2.4) we also get

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

which completes the proof. □

We have seen in Example 2.1 that the Bayes risk is 0 if Y is completely determined by X. At the other extreme, if the response Y does not depend on X at all, then the Bayes risk is $1/2$. This means that for every input, the best possible classifier consists of "guessing" without any prior information!

Example 2.3 (Gaussian Mixture)**.** Assume we have a list of house prices from two different neighbourhoods, let's call them 0 and 1 (or Coventry and Leamington Spa), and we would like to devise a method to "guess" the neighbourhood from the price. We assume that the prices in each neighbourhood are, after normalizing to account for intrinsic factors such as size, approximately normally distributed. There may also be more available houses in one neighbourhood than in the other, leading to an intrinsically higher probability of picking a house from one neighbourhood. Under these assumptions, we can model the distribution house prices as a distribution on $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0, 1\}$ as follows. Let $Y \sim \mathrm{Ber}(p)$ be a Bernoulli random variable with parameter p, that is, $\mathbb{P}(Y = 1) = p$ and $\mathbb{P}(Y = 0) = 1 - p$. Let X be a real-valued random variable with density ρ, such that the conditional density of X given $Y = 1$ is a Gaussian density $\rho_1(x) := \rho_{Y=1}(x)$, and the conditional density of X given $Y = 0$ is a Gaussian density $\rho_0(x) := \rho_{Y=0}(x)$, see Figure 2.1 for an illustration.

We can think of the data generating process as first sampling a value $y \in \{0, 1\}$ from Y, and then sampling x from a Gaussian distribution with density ρ_y. The goal is to find a binary classifier $h \colon \mathbb{R} \to \{0, 1\}$ that assigns x to one of the two distributions

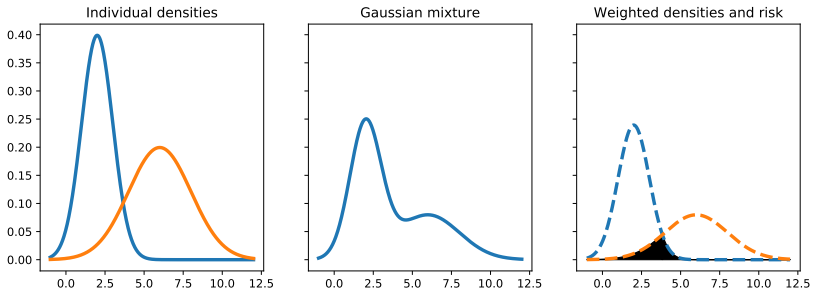

Figure 2.1: A mixture of Gaussians. If $p = 1/2$, the Bayes classifier assigns each sample to the density that is larger at this sample. The intersection of the area below the curves is the risk $R(h^*)$. If $p \neq 1/2$, then the densities are weighted.

.

that generated the data. To get the Bayes classifier, we use Bayes' rule to compute the regression function:

$$f(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{\rho_1(x)\mathbb{P}(Y = 1)}{\rho(x)} = \frac{p \cdot \rho_1(x)}{(1-p)\rho_0(x) + p\rho_1(x)}.$$

We can rearrange:

$$\frac{p \cdot \rho_1(x)}{(1-p)\rho_0(x) + p\rho_1(x)} > \frac{1}{2} \quad \Leftrightarrow \quad \frac{\rho_1(x)}{\rho_0(x)} > \frac{1-p}{p}.$$

Therefore, the Bayes classifier for the problem above is given by

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{\rho_1(x)}{\rho_0(x)} > \frac{1-p}{p} \\ 0 & \text{else} \end{cases}$$

If $p = 1/2$, then this amounts to assigning x to the class y for which $\rho_y(x)$ is larger. The risk of this classifier is clearly non-zero.

No Free Lunch

In practice, a classifier $\hat{h}_n$ is constructed from **training data** $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^n$. This can be formalized by setting

$$\hat{h}_n(\boldsymbol{x}) = \hat{g}_n(\boldsymbol{x}; (\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^n, y^n))$$

for a function $\hat{g}_n \colon \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$. For example, we could construct $\hat{h}_n$ by minimizing the empirical risk $\hat{R}(h)$ over $\mathcal{H}$, that is

$$\hat{h}_n = \operatorname*{arg\,min}_{h \in \mathcal{H}} \hat{R}(h) = \operatorname*{arg\,min}_{h \in \mathcal{H}} \sum_{i=1}^{n} \mathbf{1}\{h(\boldsymbol{x}^i) \neq y^i\}.$$

Intuitively, we would expect that more samples allow us to construct classifiers that better approximate the Bayes classifier. To study this formally, we consider the classifier $\hat{h}_n$ as constructed from random samples, i.e.,

$$\hat{h}_n(\boldsymbol{x}) = \hat{g}_n(\boldsymbol{x}; (X_1, Y_1), \ldots, (X_n, Y_n)), \tag{2.5}$$

where the (X_i, Y_i) are i.i.d. samples from the given distribution on $\mathcal{X} \times \mathcal{Y}$. A sequence of classifiers $\{\hat{h}_n\}_{n \geq 0}$ as in (2.5) is called **consistent**, if $\mathcal{E}(\hat{h}_n) \to 0$ in probability and **strongly consistent** if $\mathcal{E}(\hat{h}_n) \to 0$ almost surely. A naive approach to construct an estimator is to construct an unbiased estimator of the regression function $f(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$, and then define $\hat{h}_n$ from this estimator just like we defined the Bayes classifier using f. It can be shown that such an estimator is **universally strongly consistent**, which means that it is strongly consistent for any data distribution.

Example 2.4. In Example 2.3 we studied the Bayes classifier for the problem of assigning data to one of two components of a Gaussian mixtures. We now look at how to arrive at an estimator based solely on the observation of training data $\{\boldsymbol{x}^i, y^i\}$. If we assume that a Gaussian mixture is a good model for the data, then we may use maximum likelihood or MAP to produce an estimate of the relevant parameters of the mixture density (in practice, one would employ standard algorithms such as Expectation-Maximization (EM) to estimate the parameters of the distribution). This would then provide us with an estimate of the regression function $f(\boldsymbol{x})$ as in Example 2.3, which in turn leads to a classifier $\hat{h}_n$ that allows us to classify new, unseen data (see Figure 2.2).

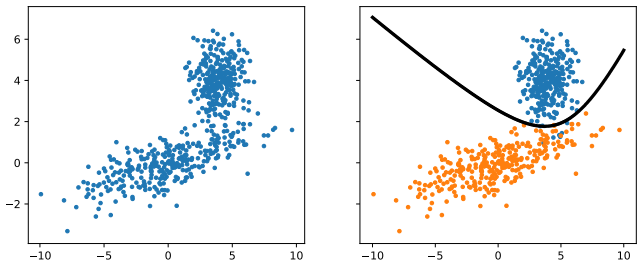

Figure 2.2: A Gaussian mixture model in two dimensions.

A subtle variation of the problem of consistency is whether it is possible to construct a series of classifiers $\hat{h}_n$ such that the excess risk $\mathcal{E}(\hat{h}_n)$ converges to 0 *uniformly* for all distributions. The following theorem shows that this is not the case. We denote by $\mathcal{D}^n$ the product distribution on $(\mathcal{X} \times \mathcal{Y})^n$ induced by a distribution $\mathbb{P}$ on the data space $\mathcal{X} \times \mathcal{Y}$.

Theorem 2.5 (No Free Lunch)**.** *Assume $|\mathcal{X}|$ is infinite. For a fixed n, consider a binary classifier $\hat{h}_n$ constructed from random data as in* (2.5)*. Let $\epsilon > 0$. Then there*

exists a probability distribution on $\mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{E}_{\mathcal{D}^n}[\mathcal{E}(\hat{h}_n)] \geq \frac{1}{2} - \epsilon.$$

Proof. Choose m elements $S = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m\} \subset \mathcal{X}$ and consider a random variable X on $\mathcal{X}$ defined as

$$\mathbb{P}(X = \boldsymbol{x}) = \frac{1}{m}\mathbf{1}\{\boldsymbol{x} \in S\}.$$

Fix a binary (0-1) sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$, and consider the random variable Y on $\mathcal{Y} = \{0, 1\}$ with conditional distribution

$$\mathbb{P}(Y = y \mid X = \boldsymbol{x}_j) = \mathbf{1}\{y = \sigma_j\}$$

and $\mathbb{P}(Y = y \mid X = \boldsymbol{x}) = 0$ if $\boldsymbol{x} \notin S$. The random variable Y is thus completely determined by X, and the regression function is

$$f_{\boldsymbol{\sigma}}(\boldsymbol{x}_j) := \mathbb{E}[Y|X = \boldsymbol{x}_j] = \sigma_j$$

on S. In particular, the Bayes risk is $R(h^*) = 0$ and the excess risk is

$$\mathcal{E}(\hat{h}_n) = R(\hat{h}_n) = \mathbb{E}_X[\mathbf{1}\{\hat{h}_n(X) \neq f_{\boldsymbol{\sigma}}(X)\}].$$

We now consider all possible such distribution, as $\boldsymbol{\sigma}$ ranges over all 2^m possible sign vectors in $\{0, 1\}^m$. Specifically, we consider $\boldsymbol{\sigma}$ itself to be uniformly distributed over $\{0, 1\}^m$. Given X, the corresponding random variable $Y = f_{\boldsymbol{\sigma}}(X)$ thus depends on random X *and* random $\boldsymbol{\sigma}$. Taking the expectation with respect to this random sign vector and with respect to random data, we get:

$$\mathbb{E}_{\boldsymbol{\sigma}}\,\mathbb{E}_{\mathcal{D}^n}[\mathcal{E}(\hat{h}_n)] = \mathbb{E}[\mathbb{P}_{\boldsymbol{\sigma}}(\hat{h}_n(X; (X_1, f_{\boldsymbol{\sigma}}(X_1)), \ldots, (X_n, f_{\boldsymbol{\sigma}}(X_n))) \neq f_{\boldsymbol{\sigma}}(X))],$$

where we exchanged the order of the expectations, replaced the expectation of an indicator function with a probability, and the outer expectation is over all X and $X_1, \ldots, X_n$. If X is different from $X_1, \ldots, X_n$, then $f_{\boldsymbol{\sigma}}(X)$ is independent of the signs $f_{\boldsymbol{\sigma}}(X_i)$ for $i \in \{1, \ldots, n\}$, and takes the values 0 and 1 with equal probability. Conditioning on $X \neq X_j$ for $j \in \{1, \ldots, n\}$, we therefore get the bound

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\,\mathbb{E}_{\mathcal{D}^n}[\mathcal{E}(\hat{h}_n)] &\geq \frac{1}{2}\mathbb{P}(X \neq X_j \colon j \in \{1, \ldots, n\}) \\ &= \frac{1}{2}\mathbb{P}(X \neq X_1)^n = \frac{1}{2}\left(1 - \frac{1}{m}\right)^n\end{aligned}$$

Given $\varepsilon > 0$, we can therefore find m such that the lower bound is great than $1/2 - \varepsilon$. Now since the *average* expected value $\mathbb{E}_{\mathcal{D}^n}[\mathcal{E}(\hat{h}_n)]$ over all distributions defined by a sign vector $\boldsymbol{\sigma} \in \{0, 1\}^m$ is greater than $1/2 - \varepsilon$, there surely has to be at least one sign vector for which which this is the case. This was to be shown. □

Notes

The classification setting described in this chapter is classic, see for example [2]. The binary classification problem may appear to be excessively simple, but once the problem of binary classification is understood, the underlying ideas generalize easily to more general situations. Our treatment of the Bayes classifier is based on [1, Chapter 2], and the version of the No Free Lunch Theorem presented here is based on [1, Section 7.1].

[1] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[2] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2013.