22
Classification

22.1 Introduction

The problem of predicting a discrete random variable Y from another random variable X is called **classification**, **supervised learning**, **discrimination**, or **pattern recognition.**

Consider IID data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$X_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X} \subset \mathbb{R}^d$$

is a d-dimensional vector and Y_i takes values in some finite set $\mathcal{Y}$. A **classification rule** is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new X, we predict Y to be $h(X)$.

22.1 Example. Here is a an example with fake data. Figure 22.1 shows 100 data points. The covariate $X = (X_1, X_2)$ is 2-dimensional and the outcome $Y \in \mathcal{Y} = \{0, 1\}$. The Y values are indicated on the plot with the triangles representing $Y = 1$ and the squares representing $Y = 0$. Also shown is a linear classification rule represented by the solid line. This is a rule of the form

$$h(x) = \begin{cases} 1 & \text{if } a + b_1 x_1 + b_2 x_2 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Everything above the line is classified as a 0 and everything below the line is classified as a 1. ■

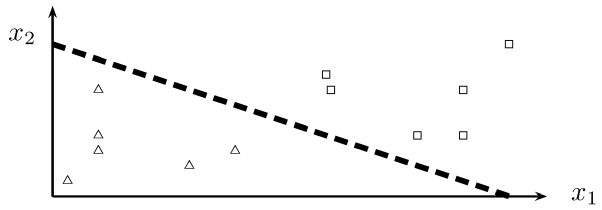

FIGURE 22.1. Two covariates and a linear decision boundary. △ means $Y = 1$. □ means $Y = 0$. These two groups are perfectly separated by the linear decision boundary; you probably won't see real data like this.

22.2 Example. Recall the the Coronary Risk-Factor Study (CORIS) data from Example 13.17. There are 462 males between the ages of 15 and 64 from three rural areas in South Africa. The outcome Y is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease and there are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. I computed a linear decision boundary using the LDA method based on two of the covariates, systolic blood pressure and tobacco consumption. The LDA method will be explained shortly. In this example, the groups are hard to tell apart. In fact, 141 of the 462 subjects are misclassified using this classification rule. ■

At this point, it is worth revisiting the Statistics/Data Mining dictionary:

Statistics	Computer Science	Meaning
classification	supervised learning	predicting a discrete Y from X
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	map $h : \mathcal{X} \to \mathcal{Y}$
estimation	learning	finding a good classifier

22.2 Error Rates and the Bayes Classifier

Our goal is to find a classification rule h that makes accurate predictions. We start with the following definitions:

22.3 Definition. *The* **true error rate**[1] *of a classifier h is*

$$L(h) = \mathbb{P}(\{h(X) \neq Y\}) \tag{22.1}$$

and the **empirical error rate** *or* **training error rate** *is*

$$\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(X_i) \neq Y_i). \tag{22.2}$$

First we consider the special case where $\mathcal{Y} = \{0, 1\}$. Let

$$r(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$$

denote the **regression function**. From Bayes' theorem we have that

$$\begin{aligned} r(x) &= \mathbb{P}(Y = 1|X = x) \\ &= \frac{f(x|Y = 1)\mathbb{P}(Y = 1)}{f(x|Y = 1)\mathbb{P}(Y = 1) + f(x|Y = 0)\mathbb{P}(Y = 0)} \\ &= \frac{\pi f_1(x)}{\pi f_1(x) + (1 - \pi)f_0(x)} \end{aligned} \tag{22.3}$$

where

$$\begin{aligned} f_0(x) &= f(x|Y = 0) \\ f_1(x) &= f(x|Y = 1) \\ \pi &= \mathbb{P}(Y = 1). \end{aligned}$$

22.4 Definition. *The* **Bayes classification rule** h^* *is*

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{22.4}$$

The set $\mathcal{D}(h) = \{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 0|X = x)\}$ is called the **decision boundary.**

Warning! The Bayes rule has nothing to do with Bayesian inference. We could estimate the Bayes rule using either frequentist or Bayesian methods.

The Bayes rule may be written in several equivalent forms:

[1] One can use other loss functions. For simplicity we will use the error rate as our loss function.

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 & \text{otherwise} \end{cases} \tag{22.5}$$

and

$$h^*(x) = \begin{cases} 1 & \text{if } \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{otherwise.} \end{cases} \tag{22.6}$$

22.5 Theorem. *The Bayes rule is optimal, that is, if h is any other classification rule then $L(h^*) \leq L(h)$.*

The Bayes rule depends on unknown quantities so we need to use the data to find some approximation to the Bayes rule. At the risk of oversimplifying, there are three main approaches:

1. Empirical Risk Minimization. Choose a set of classifiers $\mathcal{H}$ and find $\widehat{h} \in \mathcal{H}$ that minimizes some estimate of $L(h)$.

2. Regression. Find an estimate $\widehat{r}$ of the regression function r and define

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

3. Density Estimation. Estimate f_0 from the X_i's for which $Y_i = 0$, estimate f_1 from the X_i's for which $Y_i = 1$ and let $\widehat{\pi} = n^{-1} \sum_{i=1}^n Y_i$. Define

$$\widehat{r}(x) = \widehat{\mathbb{P}}(Y=1|X=x) = \frac{\widehat{\pi}\widehat{f}_1(x)}{\widehat{\pi}\widehat{f}_1(x) + (1-\widehat{\pi})\widehat{f}_0(x)}$$

and

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Now let us generalize to the case where Y takes on more than two values as follows.

22.6 Theorem. *Suppose that $Y \in \mathcal{Y} = \{1, \ldots, K\}$. The optimal rule is*

$$h(x) = \text{argmax}_k \, \mathbb{P}(Y=k|X=x) \tag{22.7}$$

$$= \text{argmax}_k \pi_k \, f_k(x) \tag{22.8}$$

where

$$\mathbb{P}(Y=k|X=x) = \frac{f_k(x)\pi_k}{\sum_r f_r(x)\pi_r}, \tag{22.9}$$

$\pi_r = P(Y=r)$, $f_r(x) = f(x|Y=r)$ and argmax_k *means "the value of k that maximizes that expression."*

22.3 Gaussian and Linear Classifiers

Perhaps the simplest approach to classification is to use the density estimation strategy and assume a parametric model for the densities. Suppose that $\mathcal{Y} = \{0, 1\}$ and that $f_0(x) = f(x|Y = 0)$ and $f_1(x) = f(x|Y = 1)$ are both multivariate Gaussians:

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\}, \quad k = 0, 1.$$

Thus, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$.

22.7 Theorem. *If $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$, then the Bayes rule is*

$$h^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 + 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & \text{otherwise} \end{cases} \tag{22.10}$$

where

$$r_i^2 = (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i), \quad i = 1, 2 \tag{22.11}$$

is the **Manalahobis distance.** *An equivalent way of expressing the Bayes' rule is*

$$h^*(x) = \text{argmax}_k \delta_k(x)$$

where

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k \tag{22.12}$$

and $|A|$ denotes the determinant of a matrix A.

The decision boundary of the above classifier is quadratic so this procedure is called **quadratic discriminant analysis (QDA)**. In practice, we use sample estimates of $\pi, \mu_1, \mu_2, \Sigma_0, \Sigma_1$ in place of the true value, namely:

$$\begin{aligned}
\widehat{\pi}_0 &= \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i), \quad \widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i \\
\widehat{\mu}_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0} X_i, \quad \widehat{\mu}_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1} X_i \\
S_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0}(X_i - \widehat{\mu}_0)(X_i - \widehat{\mu}_0)^T, \quad S_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1}(X_i - \widehat{\mu}_1)(X_i - \widehat{\mu}_1)^T
\end{aligned}$$

where $n_0 = \sum_i (1 - Y_i)$ and $n_1 = \sum_i Y_i$.

A simplification occurs if we assume that $\Sigma_0 = \Sigma_0 = \Sigma$. In that case, the Bayes rule is

$$h^*(x) = \text{argmax}_k \delta_k(x) \tag{22.13}$$

where now

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} + \log \pi_k. \tag{22.14}$$

The parameters are estimated as before, except that the MLE of Σ is

$$S = \frac{n_0 S_0 + n_1 S_1}{n_0 + n_1}.$$

The classification rule is

$$h^*(x) = \begin{cases} 1 & \text{if } \delta_1(x) > \delta_0(x) \\ 0 & \text{otherwise} \end{cases} \tag{22.15}$$

where

$$\delta_j(x) = x^T S^{-1} \widehat{\mu}_j - \frac{1}{2} \widehat{\mu}_j^T S^{-1} \widehat{\mu}_j + \log \widehat{\pi}_j$$

is called the **discriminant function**. The decision boundary $\{x : \ \delta_0(x) = \delta_1(x)\}$ is linear so this method is called **linear discrimination analysis (LDA).**

22.8 Example. Let us return to the South African heart disease data. The decision rule in in Example 22.2 was obtained by linear discrimination. The outcome was

	classified as 0	classified as 1
$y = 0$	277	25
$y = 1$	116	44

The observed misclassification rate is $141/462 = .31$. Including all the covariates reduces the error rate to .27. The results from quadratic discrimination are

	classified as 0	classified as 1
$y = 0$	272	30
$y = 1$	113	47

which has about the same error rate $143/462 = .31$. Including all the covariates reduces the error rate to .26. In this example, there is little advantage to QDA over LDA. ■

Now we generalize to the case where Y takes on more than two values.

22.9 Theorem. *Suppose that* $Y \in \{1, \ldots, K\}$. *If* $f_k(x) = f(x|Y = k)$ *is Gaussian, the Bayes rule is*

$$h(x) = \text{argmax}_k \delta_k(x)$$

where

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k. \tag{22.16}$$

If the variances of the Gaussians are equal, then

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} + \log \pi_k. \tag{22.17}$$

We estimate $\delta_k(x)$ by by inserting estimates of μ_k, Σ_k and π_k. There is another version of linear discriminant analysis due to Fisher. The idea is to first reduce the dimension of covariates to one dimension by projecting the data onto a line. Algebraically, this means replacing the covariate $X = (X_1, \ldots, X_d)$ with a linear combination $U = w^T X = \sum_{j=1}^d w_j X_j$. The goal is to choose the vector $w = (w_1, \ldots, w_d)$ that "best separates the data." Then we perform classification with the one-dimensional covariate Z instead of X.

We need define what we mean by separation of the groups. We would like the two groups to have means that are far apart relative to their spread. Let μ_j denote the mean of X for Y_j and let Σ be the variance matrix of X. Then $\mathbb{E}(U|Y = j) = \mathbb{E}(w^T X|Y = j) = w^T \mu_j$ and $\mathbb{V}(U) = w^T \Sigma w$. [2] Define the separation by

$$\begin{aligned} J(w) &= \frac{(\mathbb{E}(U|Y = 0) - \mathbb{E}(U|Y = 1))^2}{w^T \Sigma w} \\ &= \frac{(w^T \mu_0 - w^T \mu_1)^2}{w^T \Sigma w} \\ &= \frac{w^T (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T \Sigma w}. \end{aligned}$$

We estimate J as follows. Let $n_j = \sum_{i=1}^n I(Y_i = j)$ be the number of observations in group j, let $\overline{X}_j$ be the sample mean vector of the X's for group j, and let S_j be the sample covariance matrix in group j. Define

$$\widehat{J}(w) = \frac{w^T S_B w}{w^T S_W w} \tag{22.18}$$

[2] The quantity J arises in physics, where it is called the Rayleigh coefficient.

where

$$\begin{aligned} S_B &= (\overline{X}_0 - \overline{X}_1)(\overline{X}_0 - \overline{X}_1)^T \\ S_W &= \frac{(n_0 - 1)S_0 + (n_1 - 1)S_1}{(n_0 - 1) + (n_1 - 1)}. \end{aligned}$$

22.10 Theorem. *The vector*

$$w = S_W^{-1}(\overline{X}_0 - \overline{X}_1) \tag{22.19}$$

is a minimizer of $\widehat{J}(w)$. *We call*

$$U = w^T X = (\overline{X}_0 - \overline{X}_1)^T S_W^{-1} X \tag{22.20}$$

the **Fisher linear discriminant function**. *The midpoint* m *between* $\overline{X}_0$ *and* $\overline{X}_1$ *is*

$$m = \frac{1}{2}(\overline{X}_0 + \overline{X}_1) = \frac{1}{2}(\overline{X}_0 - \overline{X}_1)^T S_B^{-1}(\overline{X}_0 + \overline{X}_1) \tag{22.21}$$

Fisher's classification rule is

$$h(x) = \begin{cases} 0 & \text{if } w^T X \geq m \\ 1 & \text{if } w^T X < m. \end{cases}$$

Fisher's rule is the same as the Bayes linear classifier in equation (22.14) when $\widehat{\pi} = 1/2$.

22.4 Linear Regression and Logistic Regression

A more direct approach to classification is to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$ without bothering to estimate the densities f_k. For the rest of this section, we will only consider the case where $\mathcal{Y} = \{0, 1\}$. Thus, $r(x) = \mathbb{P}(Y = 1|X = x)$ and once we have an estimate $\widehat{r}$, we will use the classification rule

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{22.22}$$

The simplest regression model is the linear regression model

$$Y = r(x) + \epsilon = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon \tag{22.23}$$

where $\mathbb{E}(\epsilon) = 0$. This model can't be correct since it does not force $Y = 0$ or 1. Nonetheless, it can sometimes lead to a decent classifier.

Recall that the least squares estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^T$ minimizes the residual sums of squares

$$\text{RSS}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j \right)^2.$$

Let $\mathbf{X}$ denote the $N \times (d+1)$ matrix of the form

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1d} \\ 1 & X_{21} & \ldots & X_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{nd} \end{bmatrix}.$$

Also let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Then,

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

and the model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. From Theorem 13.13,

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

The predicted values are

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}.$$

Now we use (22.22) to classify, where $\widehat{r}(x) = \widehat{\beta}_0 + \sum_j \widehat{\beta}_j x_j$.

An alternative is to use logistic regression which was also discussed in Chapter 13. The model is

$$r(x) = \mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}} \tag{22.24}$$

and the MLE $\widehat{\beta}$ is obtained numerically.

22.11 Example. Let us return to the heart disease data. The MLE is given in Example 13.17. The error rate, using this model for classification, is .27. The error rate from a linear regression is .26.

We can get a better classifier by fitting a richer model. For example, we could fit

$$\text{logit } \mathbb{P}(Y = 1|X = x) = \beta_0 + \sum_j \beta_j x_j + \sum_{j,k} \beta_{jk} x_j x_k. \tag{22.25}$$

More generally, we could add terms of up to order r for some integer r. Large values of r give a more complicated model which should fit the data better. But there is a bias–variance tradeoff which we'll discuss later.

22.12 Example. If we use model (22.25) for the heart disease data with $r = 2$, the error rate is reduced to .22. ■

22.5 Relationship Between Logistic Regression and LDA

LDA and logistic regression are almost the same thing. If we assume that each group is Gaussian with the same covariance matrix, then we saw earlier that

$$\begin{aligned}\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) &= \log\left(\frac{\pi_0}{\pi_1}\right) - \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_1-\mu_0) \\ &\quad + x^T\Sigma^{-1}(\mu_1-\mu_0) \\ &\equiv \alpha_0 + \alpha^T x.\end{aligned}$$

On the other hand, the logistic model is, by assumption,

$$\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) = \beta_0 + \beta^T x.$$

These are the same model since they both lead to classification rules that are linear in x. The difference is in how we estimate the parameters.

The joint density of a single observation is $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$. In LDA we estimated the whole joint distribution by maximizing the likelihood

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(x_i|y_i)}_{\text{Gaussian}} \underbrace{\prod_i f(y_i)}_{\text{Bernoulli}}. \tag{22.26}$$

In logistic regression we maximized the conditional likelihood $\prod_i f(y_i|x_i)$ but we ignored the second term $f(x_i)$:

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(y_i|x_i)}_{\text{logistic}} \underbrace{\prod_i f(x_i)}_{\text{ignored}}. \tag{22.27}$$

Since classification only requires knowing $f(y|x)$, we don't really need to estimate the whole joint distribution. Logistic regression leaves the marginal

distribution $f(x)$ unspecified so it is more nonparametric than LDA. This is an advantage of the logistic regression approach over LDA.

To summarize: LDA and logistic regression both lead to a linear classification rule. In LDA we estimate the entire joint distribution $f(x, y) = f(x|y)f(y)$. In logistic regression we only estimate $f(y|x)$ and we don't bother estimating $f(x)$.

22.6 Density Estimation and Naive Bayes

The Bayes rule is $h(x) = \text{argmax}_k \ \pi_k f_k(x)$. If we can estimate π_k and f_k then we can estimate the Bayes classification rule. Estimating π_k is easy but what about f_k? We did this previously by assuming f_k was Gaussian. Another strategy is to estimate f_k with some nonparametric density estimator $\widehat{f}_k$ such as a kernel estimator. But if $x = (x_1, \ldots, x_d)$ is high-dimensional, nonparametric density estimation is not very reliable. This problem is ameliorated if we assume that $X_1, \ldots, X_d$ are independent, for then, $f_k(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_{kj}(x_j)$. This reduces the problem to d one-dimensional density estimation problems, within each of the k groups. The resulting classifier is called **the naive Bayes classifier.** The assumption that the components of X are independent is usually wrong yet the resulting classifier might still be accurate. Here is a summary of the steps in the naive Bayes classifier:

The Naive Bayes Classifier

1. For each group k, compute an estimate $\widehat{f}_{kj}$ of the density f_{kj} for X_j, using the data for which $Y_i = k$.
2. Let
$$\widehat{f}_k(x) = \widehat{f}_k(x_1, \ldots, x_d) = \prod_{j=1}^{d} \widehat{f}_{kj}(x_j).$$
3. Let
$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} I(Y_i = k)$$
where $I(Y_i = k) = 1$ if $Y_i = k$ and $I(Y_i = k) = 0$ if $Y_i \neq k$.
4. Let
$$h(x) = \text{argmax}_k \ \widehat{\pi}_k \, \widehat{f}_k(x).$$