Lecture 4: Binary Classification (I)

Hao Helen Zhang

Binary Classification

- Basic problem set-up binary classifiers
- Optimal classifier: under equal costs
- Optimal classifier: under unequal costs
- Simple linear classifiers

General Setup

- input vector $\boldsymbol{X} \in \mathcal{X} \subset R^d$
- output $Y \in \{0, 1\}$
- the goal is to construct a function $f : \mathcal{X} \longrightarrow \{0, 1\}$

A classification rule is often characterized as

$$f(\boldsymbol{X}) = I(b(\boldsymbol{X}) > 0),$$

where $b(\boldsymbol{X})$ is the boundary between two classes.

If $b(\boldsymbol{X})$ is a linear in $\boldsymbol{X}$, then the classifier has a *linear* boundary. It is called a *linear classification rule*.

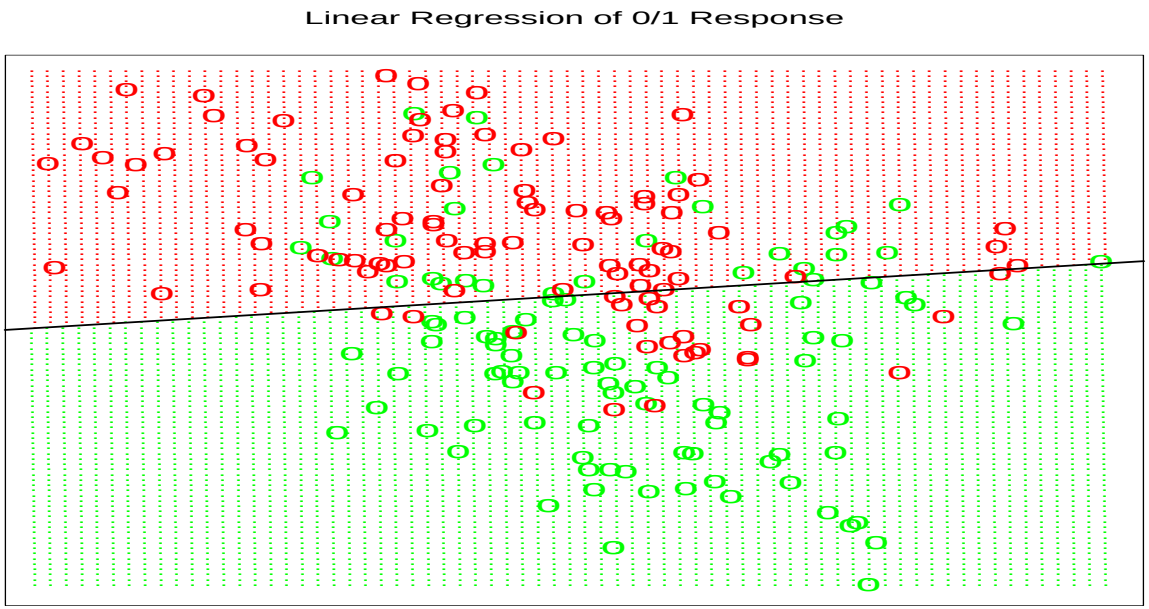


Figure 2.1: *A classification example in two dimensions. The classes are coded as a binary variable—* GREEN $= 0$, RED $= 1$ *—and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The red shaded region denotes that part of input space classified as* RED, *while the green region is classified as* GREEN.

Probability Framework

Both $\boldsymbol{X}$ and Y are random quantities.

Their distributions can be specified by

- marginal distribution of Y *prior* class probabilities:

$$\pi_1 = P(Y = 1), \quad \pi_0 = P(Y = 0)$$

- conditional density of $\boldsymbol{X}$ given Y (*class* distributions):

$$\boldsymbol{X}|Y = 1 \sim g_1(\boldsymbol{x}), \quad \boldsymbol{X}|Y = 0 \sim g_0(\boldsymbol{x})$$

Probability Framework

From the above, we can derive

- marginal density of $\boldsymbol{X}$ (a mixture):

$$g(\boldsymbol{x}) = \pi_1 g_1(\boldsymbol{x}) + \pi_0 g_0(\boldsymbol{x})$$

- conditional dist of Y given X (*posterior* class probability)

$$\begin{aligned} P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) &= \frac{\pi_1 g_1(\boldsymbol{x})}{g(\boldsymbol{x})}, \\ P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) &= 1 - P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_0 g_0(\boldsymbol{x})}{g(\boldsymbol{x})}. \end{aligned}$$

Scenario 1

- Generate 100 observations from $BVN(\boldsymbol{\mu}_1, \Sigma_1)$ with $\Sigma_1 = \mathbf{I}$ (the identity matrix), and label them as **Green** (Class 1);
- Generate 100 observations from $BVN(\boldsymbol{\mu}_0, \Sigma_0)$ with $\Sigma_0 = \mathbf{I}$, and label them as **Red** (Class 0).

Two classes have the same prior probabilities.

Write down their distributions.

Scenario 1

- Generate 100 observations from $BVN(\boldsymbol{\mu}_1, \Sigma_1)$ with $\Sigma_1 = \mathbf{I}$ (the identity matrix), and label them as **Green** (Class 1);
- Generate 100 observations from $BVN(\boldsymbol{\mu}_0, \Sigma_0)$ with $\Sigma_0 = \mathbf{I}$, and label them as **Red** (Class 0).

Two classes have the same prior probabilities.
Write down their distributions.

$$\pi_0 = \pi_1 = 0.5, \quad \boldsymbol{X}|Y = 1 \sim g_1(\boldsymbol{x}), \quad \boldsymbol{X}|Y = 0 \sim g_0(\boldsymbol{x}),$$

$$g_1(\boldsymbol{x}) = (2\pi)^{-1} \exp\{-\frac{1}{2}[(x_1 - \mu_{11})^2 + (x_2 - \mu_{12})^2]\}$$

$$g_0(\boldsymbol{x}) = (2\pi)^{-1} \exp\{-\frac{1}{2}[(x_1 - \mu_{01})^2 + (x_2 - \mu_{02})^2]\}.$$

The marginal pdf of $\boldsymbol{X}$ is

$$g(\boldsymbol{x}) = 0.5 g_0(\boldsymbol{x}) + 0.5 g_1(\boldsymbol{x}).$$

Bayes (Optimal) Rule under 0-1 Cost

For any f, we measure its performance under 0-1 loss by its risk

$$R(f) = EPE(f) = E_{X,Y} I(Y \neq f(\boldsymbol{X})) = P(Y \neq \boldsymbol{X}).$$

The Bayes rule is f^*, defined as $f^* = \arg\min_f R(f)$ and given by

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad P(Y=1|\boldsymbol{X}=\boldsymbol{x}) > P(Y=0|\boldsymbol{X}=\boldsymbol{x}) \\ 0 & \text{if} \quad P(Y=1|\boldsymbol{X}=\boldsymbol{x}) < P(Y=0|\boldsymbol{X}=\boldsymbol{x}). \end{cases}$$

- We sometimes denote the Bayes rule as ϕ_B.
- ϕ_B minimizes the probability of making an error.

The *classification boundary* of the Bayes rule is

$$\begin{aligned} & \{\boldsymbol{x} : P(Y=1|\boldsymbol{X}=\boldsymbol{x}) = P(Y=0|\boldsymbol{X}=\boldsymbol{x})\} \\ = \quad & \{\boldsymbol{x} : P(Y=1|\boldsymbol{X}=\boldsymbol{x}) - 0.5 = 0\}. \end{aligned}$$

Prior Class Probabilities (π)

- reflect prior knowledge of the proportion of each class
- can be used to make a decision without any extra knowledge

If there is no information regarding $\boldsymbol{x}$ and only the prior probabilities are available, a natural classifier would be

$$f = 1, \quad \text{if} \quad \pi_1 > \pi_0; \qquad f = 0, \quad \text{otherwise;}$$

If $\pi_1 = \pi_0$, we assign the sample randomly to one class.

- This f classifies all the data points to one class
- This f minimizes the probability of making an error

Rare Disease Example

Example: Assume a certain rare disease occurs among 1% of the population. Now a person comes and we do not have any extra information about him/her. What is your prediction rule?

Define Class 1 = "disease", 0= "disease-free". Since

$$1\% = \pi_1 < \pi_0 = 99\%,$$

our classification rule is

$$f \equiv 0.$$

Compute its risk under 0-1 loss.

Rare Disease Example

Example: Assume a certain rare disease occurs among 1% of the population. Now a person comes and we do not have any extra information about him/her. What is your prediction rule?

Define Class 1 = "disease", 0= "disease-free". Since

$$1\% = \pi_1 < \pi_0 = 99\%,$$

our classification rule is

$$f \equiv 0.$$

Compute its risk under 0-1 loss.

Answer:

$$EPE(f) = P(Y \neq f) = P(Y \neq 0) = P(Y = 1) = 1\%.$$

We have used the *Prior Class* probabilities to make a prediction.

Posterior Class Probabilities

Recall the Bayes rule is

$$f(\boldsymbol{x}) = I(P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > P(Y = 0|\boldsymbol{X} = \boldsymbol{x})).$$

- posterior class probabilities $P(Y = j|\boldsymbol{X} = \boldsymbol{x})$ provide updated class probabilities after observing $\boldsymbol{x}$.
- If $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = 0.5$, we randomly assign data to one class.

Rare Disease Example (cont.)

Example: Assume a certain rare disease occurs among 1% of the population. There is a test for this disease that is 99% accurate in the sense: 99.5% of the disease will test positive, and only 0.5% of the disease-free group will test positive. (We assume the false positive and false negative rate are both 0.005.) Now a person comes with a positive test result. What is your prediction rule?

Prior Probabilities and Class Densities

Class 1 = "disease"
Class 0 = "disease-free".

The covariate x takes only two values,

$$\{\text{tested as "+"}, \text{ tested as "-"}\}$$

We have

$$\pi_1 = 1\%, \quad \pi_0 = 99\%.$$

And

$$P(X = +|Y = 1) = 0.995, \quad P(X = -|Y = 1) = 0.005,$$

$$P(X = +|Y = 0) = 0.005, \quad P(X = -|Y = 0) = 0.995.$$

Compute Posterior Probabilities

Using Bayes' Theorem,

$$
\begin{aligned}
& P(Y=1|X=+) \\
= & \frac{P(X=+|Y=1)P(Y=1)}{P(X=+|Y=1)P(Y=1)+P(X=+|Y=0)P(Y=0)} \\
= & \frac{0.995*0.01}{0.995*0.01+0.005*0.99} = 0.668.
\end{aligned}
$$

Since $P(Y=0|X=+) = 0.332 < 0.668$, the Bayes rule for a person with "+" is $f(x=+)=1$.

Similarly, $P(Y=0|X=-) = 0.9999, P(Y=1|X=-) = 0.0001$, so $f(x=-)=0$.

Partition Induced by f

Any learner f divides the whole input space as

$$\mathcal{X} = \Omega_1 \cup \Omega_0,$$

where each Ω_j represents the "territory" of class j specified by f.

$$\Omega_j = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = j\}, \quad j = 0, 1.$$

Therefore, f can be characterized as the indicator function

$$f(\boldsymbol{x}) = I(\boldsymbol{x} \in \Omega_1).$$

Likelihood Ratio Connection

Recall that $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi_0 g_0(\boldsymbol{x})/(\pi_0 g_0(\boldsymbol{x}) + \pi g_1(\boldsymbol{x}))$.
The Bayes rule can be expressed as

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} > \frac{\pi_0}{\pi_1}, \\ 0 & \text{if } \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} < \frac{\pi_0}{\pi_1}. \end{cases}$$

The Bayes decision boundary is given by

$$\{\boldsymbol{x} : \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} = \frac{\pi_0}{\pi_1}\}.$$

Here $l(\boldsymbol{x}) = \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}$ is known as the *likelihood ratio*.

Example 1

Consider the equal cost situation:

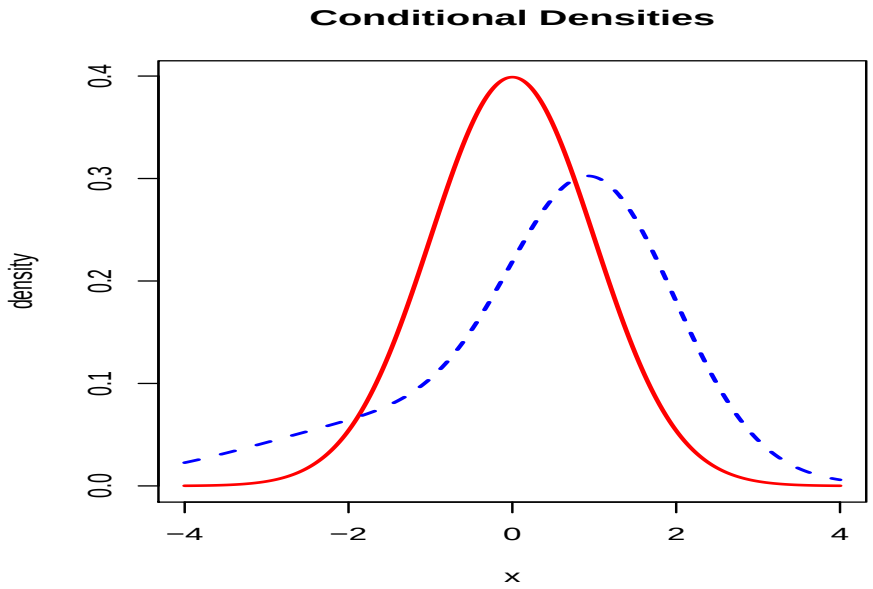
- $\pi_1 = \pi_0 = 0.5$ (balanced classes)
- $g_1(x) = \phi(x; \mu = 0, \sigma = 1)$
- $g_0(x) = 0.65\phi(x; \mu = 1, \sigma = 1) + 0.35\phi(x; \mu = -1, \sigma = 2)$

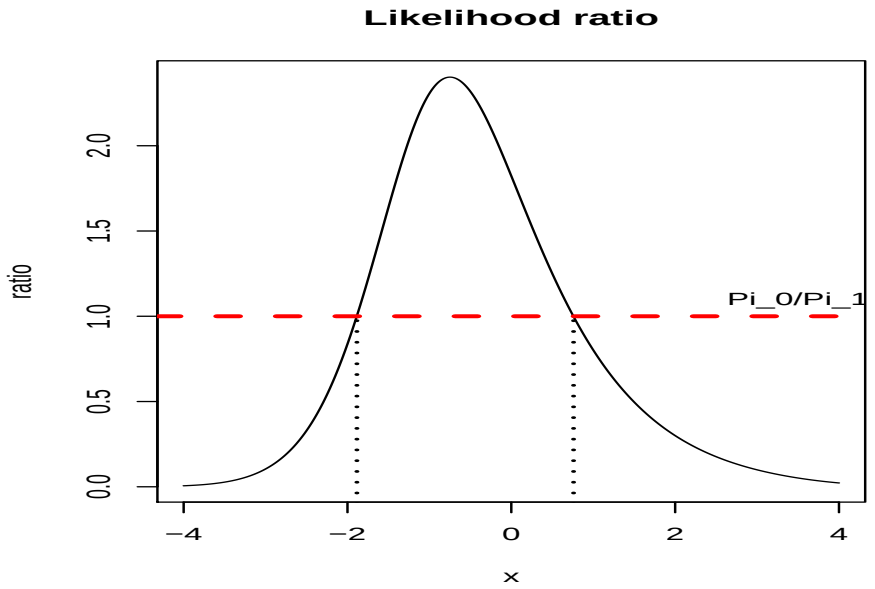
The Bayes boundary is

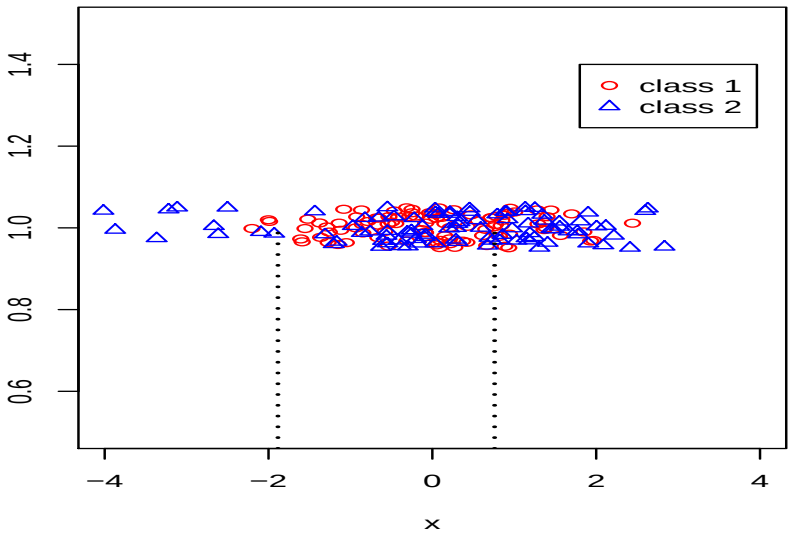
$$\{x : \frac{g_1(x)}{g_0(x)} = 1\} = \{-1.89, 0.76\}.$$

So the optimal classification regions are

$$\Omega_1^* = (-1.89, 0.76), \quad \Omega_0^* = (-\infty, -1.89) \cup (0.76, \infty).$$


Conditional Densities
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


Likelihood ratio
ratio
0.0
0.5
1.0
1.5
2.0
Pi_0/Pi_1
−4
−2
0
2
4
x


class 1
class 2
0.6
0.8
1.0
1.2
1.4
−4
−2
0
2
4
x

Classification Error & Cost

For any decision function, there are two possible errors:

- misclassifying a sample in class 0 to 1 (false positive)
- misclassifying a sample in class 1 to 0 (false negative)

Each type of error is associated with a cost (the price to pay for the consequence):

$C(1,0)$ is the cost of misclassifying a sample in class 1 to 0
$C(0,1)$ is the cost of misclassifying a sample in class 0 to 1

Typically, we assume $C(i,j) \geq 0$ for any i, j. And $C(j,j) = 0$ for $j = 0, 1$.

Unequal Costs and Prediction Risk Function

Assume $C(0,1) \neq C(1,0)$. The loss becomes

$$L = C(1,0)I(Y=1, f(\boldsymbol{X})=0) + C(0,1)I(Y=0, f(\boldsymbol{X})=1).$$

For any learner f, its prediction risk is calculated as

$$\begin{aligned} R(f) &= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}} L(Y, f(\boldsymbol{X})) \\ &= E_{\boldsymbol{X}}[C(1,0)P(Y=1|\boldsymbol{X})I(f(\boldsymbol{X})=0) \\ &\quad + C(0,1)P(Y=0|X)I(f(\boldsymbol{X})=1)] \end{aligned}$$

Optimal Learner Under Unequal Costs

For fixed $\boldsymbol{x}$, the Bayes rule is given as

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } C(1,0)P(Y=1|\boldsymbol{X}=\boldsymbol{x}) > C(0,1)P(Y=0|\boldsymbol{X}=\boldsymbol{x}) \\ 0 & \text{if } C(1,0)P(Y=1|\boldsymbol{X}=\boldsymbol{x}) < C(0,1)P(Y=0|\boldsymbol{X}=\boldsymbol{x}). \end{cases}$$

Or, equivalently,

$$f^*(x) = \begin{cases} 1 & \text{if } \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} > \frac{C(0,1)}{C(1,0)}, \\ 0 & \text{if } \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} < \frac{C(0,1)}{C(1,0))}. \end{cases}$$

Unequal Costs for Two-Class Problems

Alternatively, the Bayes rule is expressed as

$$\phi_B(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad P(Y=1|\mathbf{X}=\mathbf{x}) > \frac{C(0,1)}{C(1,0)+C(0,1)} \\ 0 & \text{o.w.} \end{cases}.$$

- When $C(0,1) \gg C(1,0)$, the thresholding value is close to 1, thus tend to classify any object to class 0.
- When $C(1,0) \gg C(0,1)$, the thresholding value is close to 0, thus tend to classify any object to class 1.

Bayes Boundary for Unequal Costs

Using the likelihood ratio, the Bayes rule can be expressed as

$$f^*(x) = \begin{cases} 1 & \text{if } \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} > \frac{\pi_0 C(0,1)}{\pi_1 C(1,0)}, \\ 0 & \text{if } \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} < \frac{\pi_0 C(0,1)}{\pi_1 C(1,0)}. \end{cases}$$

The Bayes decision boundary is given by

$$\begin{aligned} & \{x : \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \frac{C(0,1)}{C(1,0)}\} \\ = & \{x : \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} = \frac{\pi_0 C(0,1)}{\pi_1 C(1,0)}\} \\ = & \{x : l(x) = \frac{\pi_0 C(0,1)}{\pi_1 C(1,0)}\} \end{aligned}$$

Revisit Example 1 (with unequal costs)

For Example 1, assume unequal costs are used as:

$$C(0,1) = 2, C(1,0) = 1.$$

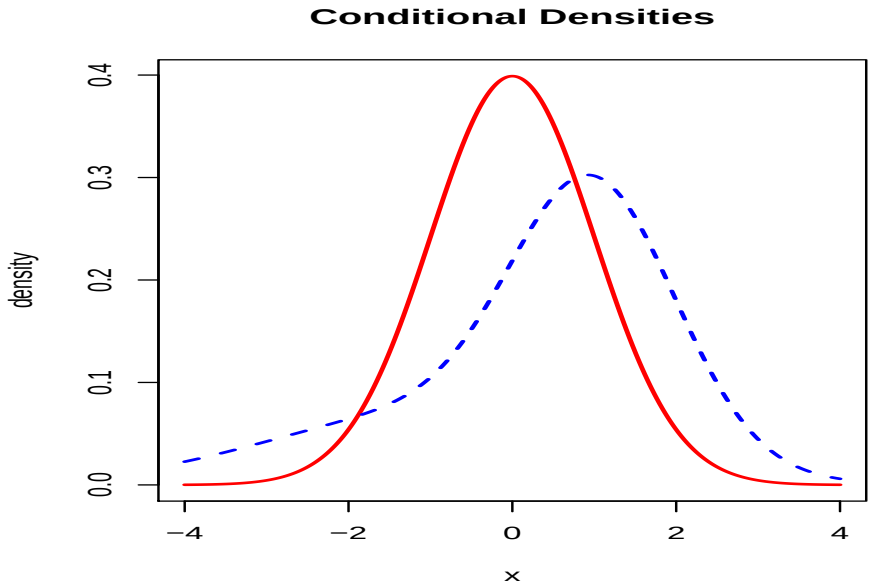
The Bayes boundary becomes

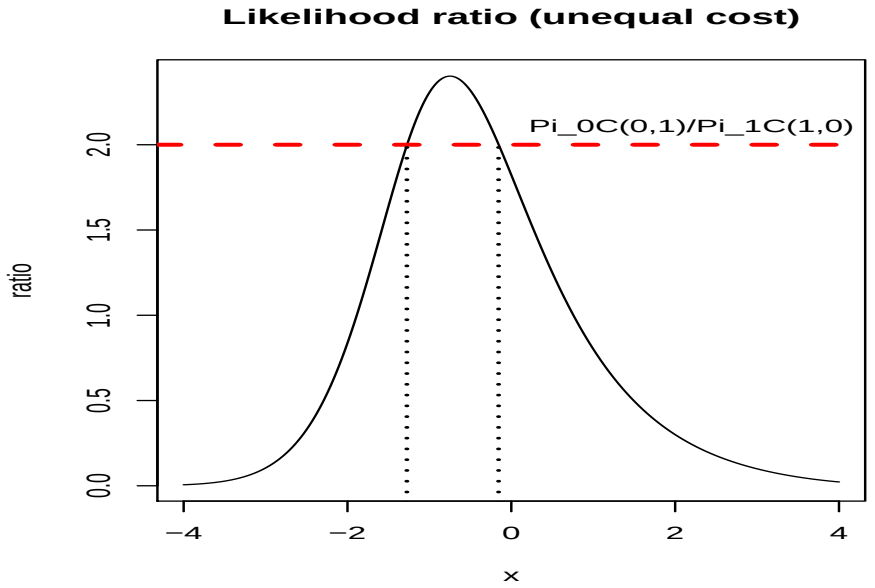
$$\{x : \frac{g_1(x)}{g_0(x)} = 2\} = \{-1.28, -0.16\}.$$

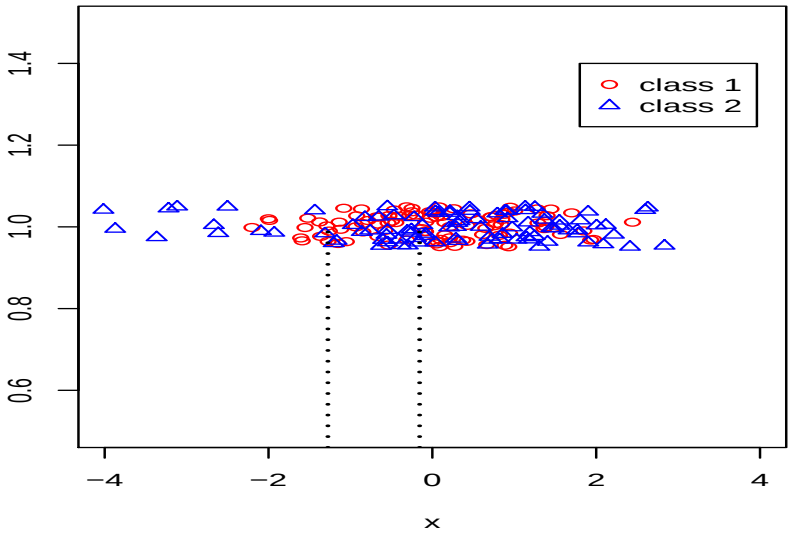
So the optimal classification regions are

$$\Omega_1^* = (-1.28, -0.16), \quad \Omega_0^* = (-\infty, -1.28) \cup (-0.16, \infty).$$

Comment: Due to the larger cost for misclassifying class 0 to 1, the decision is more protective class 0: shrinking Ω_1^* and expanding Ω_0^*.


Conditional Densities
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


Likelihood ratio (unequal cost)
ratio
0.0
0.5
1.0
1.5
2.0
Pi_0C(0,1)/Pi_1C(1,0)
−4
−2
0
2
4
x


0.6
0.8
1.0
1.2
1.4
class 1
class 2
−4
−2
0
2
4
x

Fisher Consistency

Desired properties of a classifier: *Fisher consistency*

- A binary classifier produced from loss $L(Y, f(\boldsymbol{X}))$ is *Fisher consistent*, the minimizer of $E[L(Y, f(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]$ has the same sign as the Bayes rule.
- In other words, Fisher consistency requires the loss function asymptotically yields the Bayes decision boundary.

We will talk about Fisher consistency of various losses later.

Linear Regression Models

Assume $E(Y|\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1$.

Let $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$.

Ordinary least square (OLS) estimator is minimizer of

$$RSS(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}),$$

- $\mathbf{y} = (y_1, \cdots, y_n)^T$, with each $y_i \in \{0, 1\}$.
- X is the design matrix, with the first column being 1's.

The minimizer is

$$\widehat{\boldsymbol{\beta}}^{OLS} = (X^TX)^{-1}X^T\mathbf{y}.$$

Linear Classification Rule

Classification Rule: given any input $\boldsymbol{x}$, compute

$$\hat{b}(\boldsymbol{x}) = \hat{\beta}_0^{OLS} + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}_1^{OLS} - 0.5$$

The, the predicted label is given by

$$\hat{y} = \hat{f}(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \hat{b}(\boldsymbol{x}) > 0 \\ 0 & \text{if} \quad \hat{b}(\boldsymbol{x}) < 0, \end{cases}$$

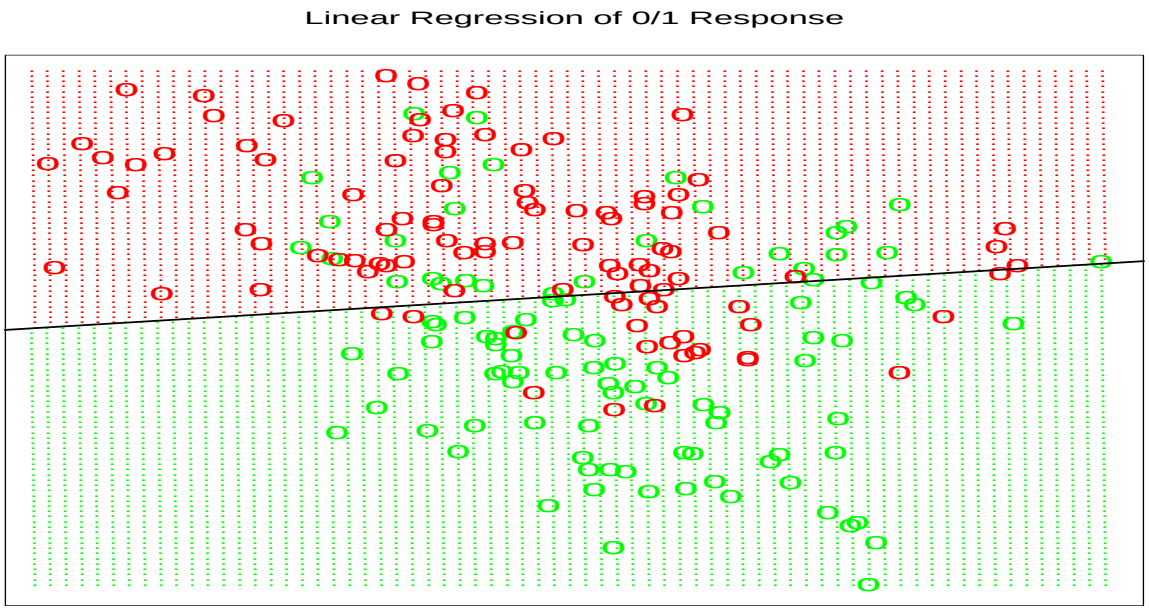


Figure 2.1: *A classification example in two dimensions. The classes are coded as a binary variable—* GREEN $= 0$, RED $= 1$ *—and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The red shaded region denotes that part of input space classified as* RED, *while the green region is classified as* GREEN.

Comments on Linear Models

Advantages:

- simple - fitting and making inferences
- the estimate is smooth
- in general, low variance

Limitations:

- rely heavily on linear assumption
- not robust against outliers
- potentially high bias