



# Statistical Learning

## Lecture 4: Classification

---

Yi He

January 18, 2023

# Plan for Today

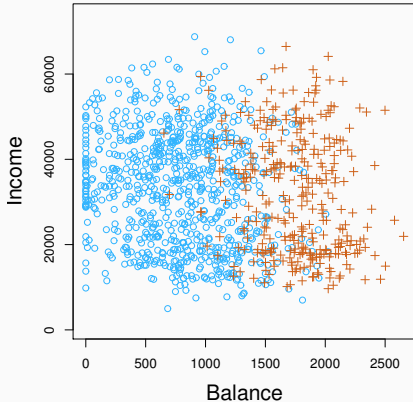
1. Introduction
2. Discriminant Analysis
3. Logistic Regression
4. Imbalanced Data

# Introduction

---

## Textbook Example: Default Data

- `default` = whether default on his or her credit card payment
- `default`  $\in \{\text{Yes}, \text{No}\}$  is binary
- Features: `balance`, `income`, `student`, ...



# Regression Function

- Encode the dummy target variable

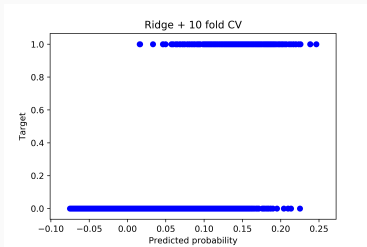
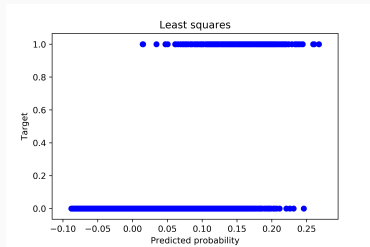
$$Y = \begin{cases} 0 & \text{default} = \text{No}, \\ 1 & \text{default} = \text{Yes}. \end{cases}$$

- The input vector  $X = (\text{balance}, \text{income}, \text{student})^T$
- Regression function

$$\begin{aligned} \mu(x) &= \mathbb{E}[Y|X = x] = 1 \cdot \mathbb{P}(Y = 1|X = x) + 0 \cdot \mathbb{P}(Y = 0|X = x) \\ &= \mathbb{P}(Y = 1|X = x) \end{aligned}$$

- Regression algorithms estimate the *posterior* probability.
- Shall we use linear regression?

# Why Not Linear Regression



- Fitted values are probabilities = not binary
- How to interpret negative estimates as probabilities?
- **Yes** if predicted probability  $> 0.5$
- ... then all predictions = 0?

# Zero-One Loss for Classification

- Binary target  $Y \in \{0, 1\}$ , features  $X \in \mathbb{R}^p$ ,  $Z = (Y, X^T)^T$
- *Classifier* or prediction rule  $g : \mathbb{R}^p \rightarrow \{0, 1\}$
- Consider the *zero-one loss*

$$\ell_{0-1}(g(X), Y) = \begin{cases} 0 & g(X) = Y \\ 1 & g(X) \neq Y \end{cases},$$

and the risk function

$$R(g) = \mathbb{E}[\ell_{0-1}(g(X), Y)] = \mathbb{P}(Y \neq g(X)), \quad g \in \mathcal{G},$$

and  $\mathcal{G}$  is a set of candidate classifiers.

# The Bayes Classifier

- Tute Q1: The *Bayes classifier* minimizes the expected zero-one loss and it is given by

$$g(x) = \operatorname{argmax}_{k \in \{0,1\}} \mathbb{P}(Y = k | X = x).$$

In other words (ignoring the decision boundary),

$$g(x) = \begin{cases} 1 & \mathbb{P}(Y = 1 | X = x) > \frac{1}{2} \\ 0 & \mathbb{P}(Y = 1 | X = x) < \frac{1}{2} \end{cases}$$

- The decision boundary  
 $\{x \in \mathbb{R}^p : \mathbb{P}(Y = 1 | X = x) = \mathbb{P}(Y = 0 | X = x) = \frac{1}{2}\}$

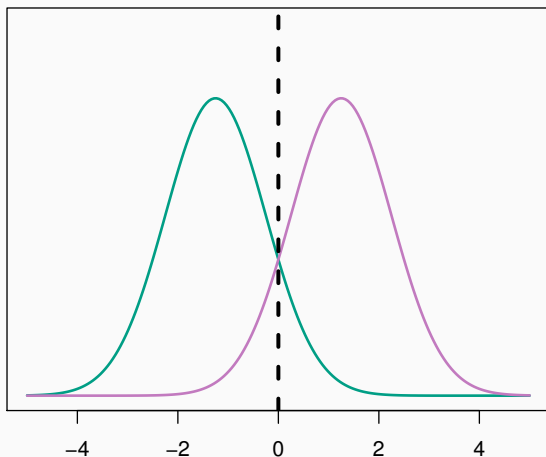


# Discriminant Analysis

---

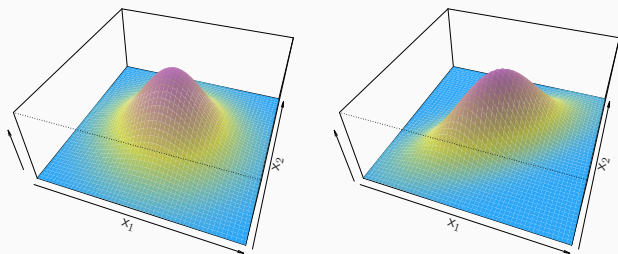
# Gaussian Models

Assume that  $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$  for  $k = 0, 1$ .



## Gaussian Models: More Illustration

Assume that  $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$  for  $k = 0, 1$ .



ISLR Figure 4.5 : Two multivariate Gaussian density functions are shown, with  $p = 2$ . Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

# Bayes' Theorem

The conditional density function of  $X$  given  $Y = k$  is Gaussian:

$$f_k(x) = \frac{1}{(\sqrt{2\pi})^p \sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right)$$

By law of iterated expectations, the unconditional density function is given by

$$f_X(x) = f_0(x) \cdot \pi_0 + f_1(x) \cdot \pi_1, \quad \pi_k = \mathbb{P}(Y = k).$$

Using the definition of conditional probability/density,

$$\mathbb{P}(Y = k | X = x) = \frac{f_{Y,X}(k, x)}{f_X(x)} = \frac{f_k(x)\pi_k}{f_0(x)\pi_0 + f_1(x)\pi_1}.$$

This is known as the *Bayes' theorem*.

# From Bayes' Theorem to Bayes' Classifier

Recall that

$$\mathbb{P}(Y = k|X = x) = \frac{f_{Y,X}(k, x)}{f_X(x)} = \frac{f_k(x)\pi_k}{f_X(x)}.$$

The denominator is  $f_X(x)$  common for all classes  $k \in \{0, 1\}$ .

The Bayes classifier

$$\begin{aligned} g(x) &= \operatorname{argmax}_{k \in \{0,1\}} \{f_k(x)\pi_k\} \\ &= \operatorname{argmax}_{k \in \{0,1\}} \{\log f_k(x) + \log \pi_k\}. \end{aligned}$$

# Quadratic Discriminant Analysis

Plugging in the Gaussian density function,

$$\begin{aligned} & \log f_k(x) + \log \pi_k \\ &= \underbrace{-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) - \frac{1}{2} \log(\det(\Sigma_k)) + \log \pi_k - \frac{p}{2} \log(2\pi)}_{\delta_k(x)}, \end{aligned}$$

The Bayes classifier  $g(x) = \operatorname{argmax}_{k \in \{0,1\}} \delta_k(x)$

- $\delta_k(x)$  is a quadratic discriminant function on  $x$
- The parameters  $\mu_k$ ,  $\Sigma_k$  and  $\pi_k$  are unknown in practice.
- In practice, QDA estimates the parameters  $\mu_k$ ,  $\Sigma_k$  and  $\pi_k$  by using *maximum likelihood* method: the resulting classifier is called the *Naive Bayes* classifier.

Go to [www.menti.com](https://www.menti.com) and use the code 6815 0114

# Shall we use QDA in high dimensions?

 Mentimeter

0

Yes, it works in both low and high dimensions

0

No, the maximum likelihood method fails in high dimensions

0

I have no idea



# Linear Discriminant Analysis

- Suppose that  $\Sigma_k = \Sigma$  for all classes  $k \in \{0, 1\}$ , that is,

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$$

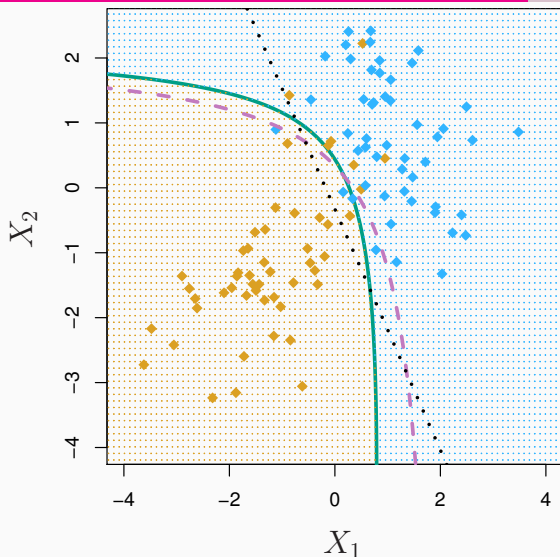
- We can then use linear discriminant functions

$$\begin{aligned} \delta_k(x) &= \underbrace{-\frac{1}{2}x^T \Sigma^{-1}x}_{\text{common}} + x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k - \underbrace{\frac{1}{2} \log \det(\Sigma)}_{\text{common}} + \log \pi_k \\ &= x^T a_k + b_k, \end{aligned}$$

- $a_k = \Sigma^{-1} \mu_k$ ,  $b_k = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$
- Again, in practice, LDA estimates the parameters  $\mu_k$ ,  $\Sigma$  and  $\pi_k$  by the maximum likelihood method.



## QDA and LDA: Simulated Example



ISLR Figure 3.8

Green=QDA  
(Estimate)

Purple=Bayes  
(True)

Black=LDA  
(Estimate)

Is QDA always better than LDA?

## QDA VS LDA: Bias-Variance Tradeoff

- Different estimators of the covariance matrices.
- Even when  $\Sigma_k$  are not equal over all classes: one may interpret LDA as a regularized version of QDA.
- The LDA estimators  $\hat{\Sigma}_0 = \hat{\Sigma}_1 = \hat{\Sigma}$  may suffer from modelling bias as the  $\hat{\Sigma}_0$  and  $\hat{\Sigma}_1$  cannot converge to different limits
- ... but with a smaller estimation variance by using the data over all classes.
- If  $\Sigma_k$  are close to each other, LDA can outperform QDA.
- If  $\Sigma_k$  are very different, QDA often performs better.
- In high dimensions, one may even use diagonal LDA that estimate  $\hat{\Sigma}_0 = \hat{\Sigma}_1 = \text{diag}(\hat{\Sigma})$  that removes the off-diagonal elements of  $\hat{\Sigma}$ : beyond our course.

# Logistic Regression

---

# Regression Function and Bayes Classifier

The regression function

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

fully characterizes the Bayes classifier

$$g(x) = \begin{cases} 0 & \mu(x) < \frac{1}{2} \\ 1 & \mu(x) > \frac{1}{2}. \end{cases}$$

- In principle, substituting the regression function  $\mu(x)$  with some estimator  $\hat{f}(x)$  yields an estimator of the Bayes classifier
- However, as discussed in the introduction, the linear regression techniques do not exploit the fact that the true regression function represents probabilities.

# Logit Transformation

The *log-odds* or *logit*

$$\begin{aligned}h(x) &\equiv \log \left( \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) \\&= \log \left( \frac{\mathbb{P}(Y = 1|X = x)}{1 - \mathbb{P}(Y = 1|X = x)} \right) = \log \left( \frac{\mu(x)}{1 - \mu(x)} \right)\end{aligned}$$

is a *logit transformation* of the regression function  
 $\mu(x) = \mathbb{P}(Y = 1|X = x)$ .

The transformation is invertible:

$$\mu(x) = \frac{\exp(h(x))}{1 + \exp(h(x))} \in (0, 1)$$

Now, replacing  $h(x)$  with any real-valued estimator  $\hat{h}(x)$  yields an estimator  $\hat{f}(x) \in (0, 1)$  that represents a probability.

# From LDA to Logistic Regression

Suppose that  $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$ ,  $k \in \{0, 1\}$

Tute 4, Q4: The Bayes' theorem implies that the logit

$$\begin{aligned}h(x) &= \delta_1(x) - \delta_0(x) \\ &= b_1 + x^T a_1 - b_0 - x^T a_0 \equiv \beta_0 + x^T \beta,\end{aligned}$$

is a linear function.

- Logistic regression fits  $\beta_0$  and  $\beta$  by maximizing the likelihood function numerically:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i: y_i=1} f(x_i) \prod_{i: y_i=0} (1 - f(x_i))$$

- In general, LR does not require the Gaussian assumption. It only requires that the log-odds function is linear in features.
- For example, one may easily incorporate a dummy feature variable.
- We may add model complexity to the log-likelihood function for feature selection or/and shrinkage purpose in high dimensions: beyond our course.

# Imbalanced Data

---



# Posterior Probabilities Using LDA/QDA

For QDA and LDA:

$$\log(f_k(x)\pi_k) = c(x) + \delta_k(x)$$

$$\Rightarrow f_k(x)\pi_k = \exp(c(x)) \cdot \exp(\delta_k(x)) \equiv C(x) \exp(\delta_k(x)),$$

where  $c(x) \in \mathbb{R}$  and  $C(x) \in (0, \infty)$  do not depend on  $k$ .

The *Bayes's theorem* gives that

$$\mathbb{P}(Y = k|X = x) = \frac{\cancel{C(x)} \exp(\delta_k(x))}{\cancel{C(x)} \exp(\delta_0(x)) + \cancel{C(x)} \exp(\delta_1(x))}$$

- Replacing  $\delta_k(x)$  with the (maximum likelihood) estimators  $\hat{\delta}_k(x)$  yields the estimated posterior probabilities.
- The Bayes classifier checks whether

$$\hat{\delta}_1(x) > \hat{\delta}_0(x) \Leftrightarrow \hat{\mathbb{P}}(Y = 1|X = x) > 0.5$$

## Shall We Always Use The Bayes Classifier?

- The estimated probabilities based on logistic regression

$$\hat{\mathbb{P}}(Y = 1|X = x) = \frac{\exp(\hat{\beta}_0 + x^T \hat{\beta})}{1 + \exp(\hat{\beta}_0 + x^T \hat{\beta})}$$

- Again, the Bayes classifier checks whether

$$\hat{\mathbb{P}}(Y = 1|X = x) > 0.5$$

- This is due to the relation between regression function and Bayes classifier.

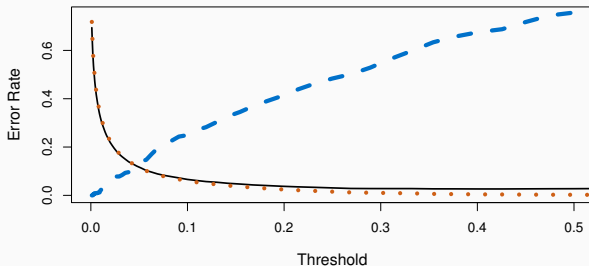
## Confusion Matrix: Accuracy Paradox

- Split default data into a training set and test set
- 4990 training data, 5010 test data

Predicted	Test observations		
	No	Yes	Total
No	4830	128	4958
Yes	9	43	52
Total	4839	171	5010

- False positive rate:  $9/4839 = 0.19\%$
- False negative rate:  $128/171 = 74.8\%$ !
- Overall error:  $(9 + 128)/5010 = 2.73\%$
- Accuracy  $1 - (9 + 128)/5010 = 97.27\%$

# Threshold Method: Textbook Example



ISLR Figure 4.7

- Predict  $\hat{Y} = 1$  when  $\hat{\mathbb{P}}(Y = 1|X = x) > c$  for a different choice of threshold  $c$
- blue = False negative, orange=False positive
- black=overall error