Classification

Maria Jose Medina

Universidad de Santiago de Chile

Outline

Introduction

- The linear regression model discussed before assumes that the response variable Y is quantitative. But now we are going to study when the response variable is instead **qualitative**.

Introduction

- The linear regression model discussed before assumes that the response variable Y is quantitative. But now we are going to study when the response variable is instead **qualitative**.
- For this, linear regression methods are not effective:

Introduction

- The linear regression model discussed before assumes that the response variable Y is quantitative. But now we are going to study when the response variable is instead **qualitative**.
- For this, linear regression methods are not effective:
  1. A regression method cannot accommodate a qualitative response with more than two classes.

Introduction

- The linear regression model discussed before assumes that the response variable Y is quantitative. But now we are going to study when the response variable is instead **qualitative**.
- For this, linear regression methods are not effective:
  1. A regression method cannot accommodate a qualitative response with more than two classes.
  2. A regression method will not provide meaningful estimates of $Pr(Y|X)$, even with just two classes.

Introduction

- The linear regression model discussed before assumes that the response variable Y is quantitative. But now we are going to study when the response variable is instead **qualitative**.
- For this, linear regression methods are not effective:
  1. A regression method cannot accommodate a qualitative response with more than two classes.
  2. A regression method will not provide meaningful estimates of $Pr(Y|X)$, even with just two classes.

Table of Contents

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.
- The logistic model suggest to use, v

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.
- The logistic model suggest to use, v

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.
- The logistic model suggest to use, v

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

- The function outputs between 0 and 1 and will always produce an S-shaped curve of this form.

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.
- The logistic model suggest to use, v

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

- The function outputs between 0 and 1 and will always produce an S-shaped curve of this form.
- Regardless of the value of X, we will obtain a probability.

Logistic regression

Logistic Model

- Here we are going to deal with a binary classification, i.e. $Y = 0$ or $Y = 1$.
- The idea is to find $p(X)$ such as $p(X) = Pr(Y = 1|X)$.
- The logistic model suggest to use, v

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

- The function outputs between 0 and 1 and will always produce an S-shaped curve of this form.
- Regardless of the value of X, we will obtain a probability.
- To fit the model, we use a method called **maximum likelihood**

Logistic regression

Estimating the model coefficients

- With a little of manipulation, equation (1) becomes,

Logistic regression

Estimating the model coefficients

- With a little of manipulation, equation (1) becomes,

$$
\begin{aligned}
p(X) &= \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \\
\frac{p(X)}{1-p(X)} &= e^{\beta_0+\beta_1 X} \\
\log\left(\frac{p(X)}{1-p(X)}\right) &= \beta_0+\beta_1 X
\end{aligned}
$$

Logistic regression

Estimating the model coefficients

- With a little of manipulation, equation (1) becomes,

$$\begin{aligned} p(X) &= \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \\ \frac{p(X)}{1-p(X)} &= e^{\beta_0+\beta_1 X} \\ \log\left(\frac{p(X)}{1-p(X)}\right) &= \beta_0+\beta_1 X \end{aligned}$$

- The coefficients β_0 and β_1 are estimated based on the available *training data* using the **maximum likelihood** method.

Logistic regression

Estimating the model coefficients

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

Logistic regression

Estimating the model coefficients

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

- We seek for estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using (1), corresponds as closely as possible to the individual's observed status.

Logistic regression

Estimating the model coefficients

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

- We seek for estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using (1), corresponds as closely as possible to the individual's observed status.
- This intuition can be formalized using a mathematical equation called a likelihood function:

Logistic regression

Estimating the model coefficients

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

- We seek for estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using (1), corresponds as closely as possible to the individual's observed status.
- This intuition can be formalized using a mathematical equation called a likelihood function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{2}$$

Logistic regression

Estimating the model coefficients

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:

- We seek for estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using (1), corresponds as closely as possible to the individual's observed status.
- This intuition can be formalized using a mathematical equation called a likelihood function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{2}$$

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

Logistic regression

Notes

Some notes about logistic regression:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.

Logistic regression

Notes

Some notes about logistic regression:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.
- The z-statistic associated with β_1 is equal to $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.

Logistic regression

Notes

Some notes about logistic regression:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.
- The z-statistic associated with β_1 is equal to $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.
- With the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$.

Logistic regression
Notes

Some notes about logistic regression:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.
- The z-statistic associated with β_1 is equal to $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.
- With the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$.
- The null hypothesis implies that $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$.

Logistic regression

Notes

Some notes about logistic regression:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.
- The z-statistic associated with β_1 is equal to $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$.
- With the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$.
- The null hypothesis implies that $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$.
- To make predictions, we simply put the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ into the equation $\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$.

Table of Contents

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.
- This extension is sometimes known as *multinomial logistic regression*.

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.
- This extension is sometimes known as *multinomial logistic regression*.
- To do this, we first select a single multinomial class to serve as the *baseline*.

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.
- This extension is sometimes known as *multinomial logistic regression*.
- To do this, we first select a single multinomial class to serve as the *baseline*.
- Without loss of generality, we select the Kth logistic class for this role.

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.
- This extension is sometimes known as *multinomial logistic regression*.
- To do this, we first select a single multinomial class to serve as the *baseline*.
- Without loss of generality, we select the Kth logistic class for this role.
- The logistic model now becomes,

Multinomial logistic regression

Multinomial model

- It is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes.
- This extension is sometimes known as *multinomial logistic regression*.
- To do this, we first select a single multinomial class to serve as the *baseline*.
- Without loss of generality, we select the Kth logistic class for this role.
- The logistic model now becomes,

$$\Pr\left(Y = k|X = x\right) = \frac{e^{\beta_{k_0}+\beta_{k_1}x_1+\cdots+\beta_{k_p}x_p}}{1 + e^{\sum_{l=1}^{K-1}\beta_{l_0}+\beta_{l_1}x_1+\cdots+\beta_{l_p}x_p}}$$

for $k = 1, \cdots, K-1$, and

$$\Pr\left(Y = K|X = x\right) = \frac{1}{1 + e^{\sum_{l=1}^{K-1}\beta_{l_0}+\beta_{l_1}x_1+\cdots+\beta_{l_p}x_p}}.$$

Multinomial logistic regression

Softmax function

- An alternative coding for multinomial logistic regression, known as the *softmax coding*.

Multinomial logistic regression

Softmax function

- An alternative coding for multinomial logistic regression, known as the *softmax coding*.
- The softmax coding is equivalent to the coding just described in the sense that the fitted values, and other key model outputs will remain the same, regardless of coding.

Multinomial logistic regression

Softmax function

- An alternative coding for multinomial logistic regression, known as the *softmax coding*.
- The softmax coding is equivalent to the coding just described in the sense that the fitted values, and other key model outputs will remain the same, regardless of coding.
- In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, \cdots, K$,

Multinomial logistic regression

Softmax function

- An alternative coding for multinomial logistic regression, known as the *softmax coding*.
- The softmax coding is equivalent to the coding just described in the sense that the fitted values, and other key model outputs will remain the same, regardless of coding.
- In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, \cdots, K$,

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k_0} + \beta_{k_1} x_1 + \cdots + \beta_{k_p} x_p}}{e^{\sum_{l=1}^{K} \beta_{l_0} + \beta_{l_1} x_1 + \cdots + \beta_{l_p} x_p}}.$$

Multinomial logistic regression

Softmax function

- An alternative coding for multinomial logistic regression, known as the *softmax coding*.
- The softmax coding is equivalent to the coding just described in the sense that the fitted values, and other key model outputs will remain the same, regardless of coding.
- In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, \cdots, K$,

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k_0} + \beta_{k_1} x_1 + \cdots + \beta_{k_p} x_p}}{e^{\sum_{l=1}^{K} \beta_{l_0} + \beta_{l_1} x_1 + \cdots + \beta_{l_p} x_p}}.$$

- Thus, rather than estimating coefficients for $K - 1$ classes, we actually estimate coefficients for all K classes.

Table of Contents

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.
- Let $f_k(X) \equiv Pr(X|Y = k)$ denote the *density function* of X for an observation that comes from the kth class.

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.
- Let $f_k(X) \equiv Pr(X|Y = k)$ denote the *density function* of X for an observation that comes from the kth class.
  $\rightarrow f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$.

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.
- Let $f_k(X) \equiv Pr(X|Y = k)$ denote the *density function* of X for an observation that comes from the kth class.
  $\rightarrow f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$.
  $\rightarrow f_k(x)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$.

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.
- Let $f_k(X) \equiv Pr(X|Y=k)$ denote the *density function* of X for an observation that comes from the kth class.
  $\rightarrow f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$.
  $\rightarrow f_k(x)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$.
- Then Bayes' theorem states that,

Generative models

Introduction

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_K represent the **prior** probability that a randomly chosen observation comes from the prior kth class.
- Let $f_k(X) \equiv Pr(X|Y = k)$ denote the *density function* of X for an observation that comes from the kth class.
  $\rightarrow f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$.
  $\rightarrow f_k(x)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$.
- Then Bayes' theorem states that,

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population.

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \quad (3)$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population.
  $\rightarrow$ Compute the fraction of the observations that belong to the kth class.

Generative models
Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population.
  $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \quad (3)$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population. $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.
- As we will see, to estimate $f_k(x)$, we will typically have to make some simplifying assumptions.

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population. $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.
- As we will see, to estimate $f_k(x)$, we will typically have to make some simplifying assumptions.
- We'll discuss three classifiers that use different estimates of $f_k(x)$:

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \quad (3)$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population.
  $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.
- As we will see, to estimate $f_k(x)$, we will typically have to make some simplifying assumptions.
- We'll discuss three classifiers that use different estimates of $f_k(x)$:
  - *Linear discriminant analysis*,

Generative models

Introduction

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population.
  $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.
- As we will see, to estimate $f_k(x)$, we will typically have to make some simplifying assumptions.
- We'll discuss three classifiers that use different estimates of $f_k(x)$:
  - *Linear discriminant analysis*,
  - *Quadratic discriminant analysis*,

Generative models

Introduction

$$p_k(x) = \mathrm{Pr}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{3}$$

- We only need the estimates of π_k and $f_k(x)$.
- Estimating π_k is easy if we have a random sample from the population. $\rightarrow$ Compute the fraction of the observations that belong to the kth class.
- Estimating the $f_k(x)$ is much more challenging.
- As we will see, to estimate $f_k(x)$, we will typically have to make some simplifying assumptions.
- We'll discuss three classifiers that use different estimates of $f_k(x)$:
  - *Linear discriminant analysis*,
  - *Quadratic discriminant analysis*,
  - *Naive Bayes*.

Table of Contents

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right) \tag{4}$$

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \text{exp}\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \tag{4}$$

- Where μ_k and σ_k^2 are the mean and variance parameters for the kth class.

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \text{exp} \left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2 \right) \tag{4}$$

- Where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- We'll assume constant variance so, $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma$.

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \tag{4}$$

- Where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- We'll assume constant variance so, $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma$.
- Plugging (4) into (3), results

Generative models

Linear discriminant analysis for $p = 1$

- For now, assume that $p = 1$: we have only one predictor.
- We want to obtain estimates for $f_k(x), \pi_k$ such as we can plug into (3) in order to estimate $p_k(x)$.
- We'll assume that $f_k(x)$ is *normal*,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \text{exp}\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \tag{4}$$

- Where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- We'll assume constant variance so, $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma$.
- Plugging (4) into (3), results

$$p_x(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \text{exp}\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \text{exp}\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)} \tag{5}$$

Generative models

Linear discriminant analysis for $p = 1$

$$p_x(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \text{exp}\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \text{exp}\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- Taking the log of (5) and rearranging the terms,

Generative models

Linear discriminant analysis for $p = 1$

$$p_x(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

- Taking the log of (5) and rearranging the terms,

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{6}$$

Generative models

Linear discriminant analysis for $p = 1$

$$p_x(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

- Taking the log of (5) and rearranging the terms,

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{6}$$

- To apply the Bayes classifier we still have to estimate the parameters π_k, μ_k and σ^2.

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2. \tag{8}$$

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2. \tag{8}$$

where n is the total number of training observations, and n_k is the number of training observations in the kth class.

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2. \tag{8}$$

  where n is the total number of training observations, and n_k is the number of training observations in the kth class.

- Sometimes we have knowledge of the class membership probabilities $\pi_1, \cdots, \pi_K$, which can be used directly.

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2. \tag{8}$$

  where n is the total number of training observations, and n_k is the number of training observations in the kth class.
- Sometimes we have knowledge of the class membership probabilities $\pi_1, \cdots, \pi_K$, which can be used directly.
- In the absence of any additional information, LDA estimates π_k using

Generative models

Linear discriminant analysis for $p = 1$

- The following estimates are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2. \tag{8}$$

  where n is the total number of training observations, and n_k is the number of training observations in the kth class.

- Sometimes we have knowledge of the class membership probabilities $\pi_1, \cdots, \pi_K$, which can be used directly.
- In the absence of any additional information, LDA estimates π_k using

$$\hat{\pi}_k = \frac{n_k}{n} \tag{9}$$

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu\hat{k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right) \tag{10}$$

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu\hat{k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right) \tag{10}$$

- The Bayes decision boundary is the point for which $\hat{\delta}_1(x) = \hat{\delta}_2(x) = \cdots = \hat{\delta}_K(x)$

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu\hat{k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right) \tag{10}$$

- The Bayes decision boundary is the point for which $\hat{\delta}_1(x) = \hat{\delta}_2(x) = \cdots = \hat{\delta}_K(x)$
- The Bayes classifier involves assigning an observation $X = x$ to the class for which (10) is largest.

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu\hat{k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right) \tag{10}$$

- The Bayes decision boundary is the point for which $\hat{\delta}_1(x) = \hat{\delta}_2(x) = \cdots = \hat{\delta}_K(x)$
- The Bayes classifier involves assigning an observation $X = x$ to the class for which (10) is largest.
- The word *linear* in the classifier's name stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ in (10) are linear functions of x.

Generative models

Linear discriminant analysis for $p = 1$

- Now, equation (6) can be rewritten as

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu\hat{k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right) \tag{10}$$

- The Bayes decision boundary is the point for which $\hat{\delta}_1(x) = \hat{\delta}_2(x) = \cdots = \hat{\delta}_K(x)$
- The Bayes classifier involves assigning an observation $X = x$ to the class for which (10) is largest.
- The word *linear* in the classifier's name stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ in (10) are linear functions of x.

Table of Contents

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu_k, \Sigma)$:

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu_k, \Sigma)$:
  $\rightarrow \mu_k$ is a class-specific mean vector.

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu_k, \Sigma)$:
  $\rightarrow$ μ_k is a class-specific mean vector.
  $\rightarrow$ $cov(X) = \Sigma$ is common to all K classes.

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu_k, \Sigma)$:
  $\rightarrow \mu_k$ is a class-specific mean vector.
  $\rightarrow cov(X) = \Sigma$ is common to all K classes.
- Formally, the multivariate Gaussian density is defined as

Generative models

Linear discriminant analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, X_2, \cdots, X_p)$ is drawn from a *multivariate Gaussian distribution*
- We assume that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.
- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu_k, \Sigma)$:
  $\rightarrow$ μ_k is a class-specific mean vector.
  $\rightarrow$ $cov(X) = \Sigma$ is common to all K classes.
- Formally, the multivariate Gaussian density is defined as

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \text{exp} \left(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right) \tag{11}$$

Generative models

Linear discriminant analysis for $p > 1$

- Plugging the density function $f_k(X = x)$, into (3), taking logs and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

Generative models

Linear discriminant analysis for $p > 1$

- Plugging the density function $f_k(X = x)$, into (3), taking logs and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \tag{12}$$

is largest.

Generative models

Linear discriminant analysis for $p > 1$

- Plugging the density function $f_k(X = x)$, into (3), taking logs and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \tag{12}$$

is largest. This is the matrix version of (6)

Generative models

Linear discriminant analysis for $p > 1$

- Plugging the density function $f_k(X = x)$, into (3), taking logs and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \tag{12}$$

  is largest. This is the matrix version of (6)
- The Bayes decision boundaries are the set of values x for which $\delta_k(x) = \delta_l(x)$ for $k \neq l$.

Generative models

Linear discriminant analysis for $p > 1$

- Plugging the density function $f_k(X = x)$, into (3), taking logs and performing a little bit of algebra reveals that the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \tag{12}$$

  is largest. This is the matrix version of (6)
- The Bayes decision boundaries are the set of values x for which $\delta_k(x) = \delta_l(x)$ for $k \neq l$.
- Once again, we need to estimate the unknown parameters $\mu_1, \cdots, \mu_K$, $\pi_1, \cdots, \pi$, and Σ.
- The formulas are similar to those used in the one-dimensional case.

Generative models

Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.

Generative models

Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.
- But in some problems, this threshold is no longer optimal.

Generative models

Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.
- But in some problems, this threshold is no longer optimal.
- To help to decide which threshold value is the best, we can use the **ROC curve**.

Generative models

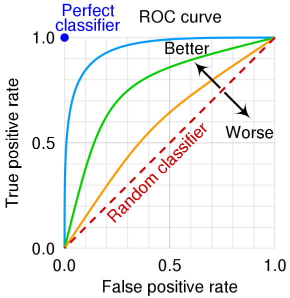
Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.
- But in some problems, this threshold is no longer optimal.
- To help to decide which threshold value is the best, we can use the **ROC curve**.
- The performance of the classifier, summarized over all possible thresholds, is given by the **Area Under the ROC Curve (AUC)**.

Generative models

Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.
- But in some problems, this threshold is no longer optimal.
- To help to decide which threshold value is the best, we can use the **ROC curve**.
- The performance of the classifier, summarized over all possible thresholds, is given by the **Area Under the ROC Curve (AUC)**.

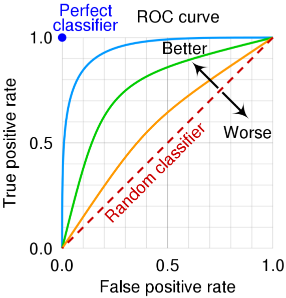


- An ideal ROC curve will hug the top left corner.

Generative models

Linear discriminant analysis for $p > 1$

- In a binary context, the Bayes classifier, and by extension LDA, uses a *threshold* of 0.5 to assign an observation to a particular class.
- But in some problems, this threshold is no longer optimal.
- To help to decide which threshold value is the best, we can use the **ROC curve**.
- The performance of the classifier, summarized over all possible thresholds, is given by the **Area Under the ROC Curve (AUC)**.



- An ideal ROC curve will hug the top left corner.
- The larger the AUC the better the classifier.

Table of Contents

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.
- However, QDA assumes that each class has its **own covariance matrix**.

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.
- However, QDA assumes that each class has its **own covariance matrix**.
  $\rightarrow$ Assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class.

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.
- However, QDA assumes that each class has its **own covariance matrix**.
  $\rightarrow$ Assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class.
- Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.
- However, QDA assumes that each class has its **own covariance matrix**.
  $\rightarrow$ Assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class.
- Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \tag{13}$$

is largest.

Quadratic discriminant analysis

- Assumes that the discriminant observations from each class are drawn from a **Gaussian distribution**.
- The idea is to compute each estimate and then plug them Bayes' theorem in order to perform a prediction.
- However, QDA assumes that each class has its **own covariance matrix**.
  $\rightarrow$ Assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class.
- Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \qquad (13)$$

  is largest.
- So the QDA classifier involves plugging estimates for μ_k, Σ_k and π_k into (13), and then assigning an observation $X = x$ to the class for which this quantity is **largest**.

Table of Contents

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.

With quadratic discriminant analysis:

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.

With quadratic discriminant analysis:

- Estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.
- LDA is a much less flexible classifier.

With quadratic discriminant analysis:

- Estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.
- LDA is a much less flexible classifier.

With quadratic discriminant analysis:

- Estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.
- Recommended if the training set is very large, so that the variance is not a major concern.

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.
- LDA is a much less flexible classifier.
- LDA can suffer from high bias.

With quadratic discriminant analysis:

- Estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.
- Recommended if the training set is very large, so that the variance is not a major concern.

When to use Linear vs quadratic discriminant analysis?

With linear discriminant analysis:

- Assumes common covariance matrix and becomes linear in x, for a total of Kp linear coefficients to estimate.
- LDA is a much less flexible classifier.
- LDA can suffer from high bias.

With quadratic discriminant analysis:

- Estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.
- Recommended if the training set is very large, so that the variance is not a major concern.
- Use it when the assumption of a common covariance matrix is clearly untenable.

Table of Contents

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \text{Pr}(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \text{Pr}(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \text{Pr}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \Pr(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- As we saw in previous sections, estimating the prior probabilities $\pi_1, \cdots, \pi_K$ is typically straightforward: for instance, we can estimate $\hat{\pi}_k$ as $\hat{\pi}_k = n_k/n$.

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \Pr(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- As we saw in previous sections, estimating the prior probabilities $\pi_1, \cdots, \pi_K$ is typically straightforward: for instance, we can estimate $\hat{\pi}_k$ as $\hat{\pi}_k = n_k/n$.
- However, estimating $f_1(x), \cdots, f_K(x)$ is more subtle.

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \Pr(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- As we saw in previous sections, estimating the prior probabilities $\pi_1, \cdots, \pi_K$ is typically straightforward: for instance, we can estimate $\hat{\pi}_k$ as $\hat{\pi}_k = n_k/n$.
- However, estimating $f_1(x), \cdots, f_K(x)$ is more subtle.
- The naive Bayes classifier assumes,

$$f_k(x) = f_{k_1}(x_1) \times f_{k_2}(x_2) \times \cdots \times f_{k_p}(x_p) \tag{14}$$

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \Pr(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- As we saw in previous sections, estimating the prior probabilities $\pi_1, \cdots, \pi_K$ is typically straightforward: for instance, we can estimate $\hat{\pi}_k$ as $\hat{\pi}_k = n_k/n$.
- However, estimating $f_1(x), \cdots, f_K(x)$ is more subtle.
- The naive Bayes classifier assumes,

$$f_k(x) = f_{k_1}(x_1) \times f_{k_2}(x_2) \times \cdots \times f_{k_p}(x_p) \tag{14}$$

where f_{k_j} is the density function of the jth predictor among observations in the kth class.

Naive Bayes

Introduction

- Recall that Bayes' theorem provides an expression for the posterior probability, $p_k(x) = \Pr(Y = k|X = x)$ in terms of $\pi_1, \cdots, \pi_K$ and $f_1(x), \cdots, f_K(x)$.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- As we saw in previous sections, estimating the prior probabilities $\pi_1, \cdots, \pi_K$ is typically straightforward: for instance, we can estimate $\hat{\pi}_k$ as $\hat{\pi}_k = n_k/n$.
- However, estimating $f_1(x), \cdots, f_K(x)$ is more subtle.
- The naive Bayes classifier assumes,

$$f_k(x) = f_{k_1}(x_1) \times f_{k_2}(x_2) \times \cdots \times f_{k_p}(x_p) \tag{14}$$

where f_{k_j} is the density function of the jth predictor among observations in the kth class. $\rightarrow$ Within the kth class, the p predictors are independent.

Naive Bayes

Introduction

- Once we have made the naive Bayes assumption, we can plug (14) into (3) to obtain an expression for the posterior probability,

$$p_k(x) = \frac{\pi_k f_{k_1}(x_1) \times f_{k_2}(x_2) \times \cdots \times f_{k_p}(x_p)}{\sum_{l=1}^{K} \pi_l f_{l_1}(x_1) \times f_{l_2}(x_2) \times \cdots \times f_{l_p}(x_p)} \tag{15}$$

Naive Bayes

Introduction

- Once we have made the naive Bayes assumption, we can plug (14) into (3) to obtain an expression for the posterior probability,

$$p_k(x) = \frac{\pi_k f_{k_1}(x_1) \times f_{k_2}(x_2) \times \cdots \times f_{k_p}(x_p)}{\sum_{l=1}^{K} \pi_l f_{l_1}(x_1) \times f_{l_2}(x_2) \times \cdots \times f_{l_p}(x_p)} \tag{15}$$

  for $k = 1, \cdots, K$.
- To estimate the one-dimensional density function f_{kj} using training data $x_{1j}, \cdots, x_{nj}$, we have a few options.

Naive Bayes

Estimating the density function

1. X_j **is quantitative and parametric**

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.

Naive Bayes

Estimating the density function

1. X_j **is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.
2. X_j **is quantitative and non-parametric**

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.
2. **X_j is quantitative and non-parametric**
   - A very simple way to do this is by making a histogram for the observations of the jth predictor within each class.

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.
2. **X_j is quantitative and non-parametric**
   - A very simple way to do this is by making a histogram for the observations of the jth predictor within each class.
   - Then, $f_{kj}(x_j)$ is the fraction of the training observations in the kth class that belong to the **same histogram bin** as x_j.

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.
2. **X_j is quantitative and non-parametric**
   - A very simple way to do this is by making a histogram for the observations of the jth predictor within each class.
   - Then, $f_{kj}(x_j)$ is the fraction of the training observations in the kth class that belong to the **same histogram bin** as x_j.
   - Alternatively, we can use a *kernel density estimator*, which is essentially a smoothed version of a histogram
3. **X_j is qualitative**

Naive Bayes

Estimating the density function

1. **X_j is quantitative and parametric**
   - We can assume that $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$.
     $\rightarrow$ We assume that within each class, the jth predictor is drawn from a (univariate) normal distribution.
   - But predictors are independent from each other.
     $\rightarrow$ the class-specific covariance matrix is diagonal.
2. **X_j is quantitative and non-parametric**
   - A very simple way to do this is by making a histogram for the observations of the jth predictor within each class.
   - Then, $f_{kj}(x_j)$ is the fraction of the training observations in the kth class that belong to the **same histogram bin** as x_j.
   - Alternatively, we can use a *kernel density estimator*, which is essentially a smoothed version of a histogram
3. **X_j is qualitative**
   - Simply count the proportion of training estimator observations for the jth predictor corresponding to each class.

A final note

None of these methods uniformly dominates the others: in any setting, the choice of method will depend on:

A final note

None of these methods uniformly dominates the others: in any setting, the choice of method will depend on:

- The true distribution of the predictors in each of the K classes.

A final note

None of these methods uniformly dominates the others: in any setting, the choice of method will depend on:

- The true distribution of the predictors in each of the K classes.
- The values of n and p.

A final note

None of these methods uniformly dominates the others: in any setting, the choice of method will depend on:

- The true distribution of the predictors in each of the K classes.
- The values of n and p.
- The bias-variance trade-off, etc.

Thank you!

Any question?