Statistics 101C - Week 2 - Thursday

Shirong Xu

University of California, Los Angeles

shirong@stat.ucla.edu

October 9, 2024

Logistic Regression and Logistic Loss function

- Logistic Loss function:

$$L(f(\boldsymbol{x}_i), y_i) = \log\Big(1 + \exp(-f(\boldsymbol{x}_i)y_i)\Big)$$

- The expected logistic loss (Logistic Risk):

$$\begin{aligned} R_{\log}(f) =& \mathbb{E}_{\boldsymbol{X},Y}\left(L(f(\boldsymbol{X}), Y)\right) \\ =& \mathbb{E}_{\boldsymbol{X}}\Big[\mathbb{P}(Y=1|\boldsymbol{X})\log\Big(1+\exp(-f(\boldsymbol{X}))\Big) + \mathbb{P}(Y=-1|\boldsymbol{X})\log\Big(1+\exp(f(\boldsymbol{X}))\Big) \end{aligned}$$

- The optimal function minimizing $R_{\log}(f)$ is defined as

$$f^*_{log}(\boldsymbol{X}) = \log\left(\frac{\mathbb{P}(Y=1|\boldsymbol{X})}{1-\mathbb{P}(Y=1|\boldsymbol{X})}\right)$$

Logistic Loss for classification

- We can construct a classifier as $\text{sign}(f^*_{log}(\boldsymbol{x})) \in \{-1, 1\}$. This classifier is identical to the Bayes classifier.

$$\text{sign}(f^*_{log}(\boldsymbol{x})) = \text{sign}(\eta(\boldsymbol{x}) - 1/2)$$

Logistic Loss for classification

- Suppose we have a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and we employ the logistic loss for classification. Then the training error in terms of logistic loss can be written as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-f(\boldsymbol{x}_i)y_i)\right)$$

where $y_i \in \{-1, 1\}$.

Logistic Loss for classification

- Suppose we have a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and we employ the logistic loss for classification. Then the training error in terms of logistic loss can be written as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp(-f(\boldsymbol{x}_i)y_i)\right)$$

  where $y_i \in \{-1, 1\}$.

- Here we do not make any assumption of f.

Logistic Loss for classification

- Suppose we have a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and we employ the logistic loss for classification. Then the training error in terms of logistic loss can be written as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(1 + \exp(-f(\boldsymbol{x}_i)y_i)\Big)$$

  where $y_i \in \{-1, 1\}$.
- Here we do not make any assumption of f.
- **Question**: What if we assume that $f(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i$

If $f(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i$

- $R_n(f)$ can be further written as

$$
\begin{aligned}
R_n(f) =& \frac{1}{n} \sum_{i=1}^{n} \log\Big(1 + \exp(-f(\boldsymbol{x}_i) y_i)\Big) \\
=& \frac{1}{n} \sum_{i=1}^{n} \log\Big(1 + \exp\big(-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) y_i\big)\Big)
\end{aligned}
$$

- We suppose that $y_i = 1 \Leftrightarrow \widetilde{y}_i = 1$ and $y_i = -1 \Leftrightarrow \widetilde{y}_i = 0$

$$
\begin{aligned}
R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \Big[& \widetilde{y}_i \log\Big(1 + \exp\big(-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)\big)\Big) \\
& + (1 - \widetilde{y}_i) \log\Big(1 + \exp\big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i\big)\Big)\Big]
\end{aligned}
$$

If $f(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i$

- We suppose that $y_i = 1 \Leftrightarrow \widetilde{y}_i = 1$ and $y_i = -1 \Leftrightarrow \widetilde{y}_i = 0$

$$
\begin{aligned}
R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \Big[& \widetilde{y}_i \log\Big(1 + \exp\big(-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)\big)\Big) \\
& + (1 - \widetilde{y}_i) \log\Big(1 + \exp\big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i\big)\Big)\Big] \\
= \frac{1}{n}\sum_{i=1}^{n} \Big[& \widetilde{y}_i \log\Big(\exp\big(-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)\big)\Big) \\
& + \log\Big(1 + \exp\big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i\big)\Big)\Big]
\end{aligned}
$$

Logistic Regression

- **Logistic regression** estimates the conditional probability probability:

$$\mathbb{P}\left(\widetilde{Y} = 1 \middle| \boldsymbol{X}\right)$$

- In logistic regression, it is assumed that

$$\mathbb{P}\left(\widetilde{Y} = 1 \middle| \boldsymbol{X} = \boldsymbol{x}\right) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})},$$
$$\mathbb{P}\left(\widetilde{Y} = 0 \middle| \boldsymbol{X} = \boldsymbol{x}\right) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})},$$

where

- $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is a p-dimensional predictor
- β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are unknown parameters
- $\boldsymbol{\beta}^T \boldsymbol{x} = \sum_{l=1}^{p} \beta_i x_i$

Estimation in Logistic Regression

- Suppose a dataset in logistic regression is $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^{n}$, here $\widetilde{y}_i \in \{0, 1\}$.
- Likelihood function $L(\beta_0, \boldsymbol{\beta})$:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\mathbb{P}\left(\widetilde{Y} = 1 \middle| \boldsymbol{X} = \boldsymbol{x}\right)\right)^{\widetilde{y}_i} \left(\mathbb{P}\left(\widetilde{Y} = 0 \middle| \boldsymbol{X} = \boldsymbol{x}\right)\right)^{1-\widetilde{y}_i}$$

Negative Log-likelihood

Negative Log-likelihood:

$$
\begin{aligned}
& -\log L(\beta_0, \boldsymbol{\beta}) \\
= & -\sum_{i=1}^{n} \left[\widetilde{y}_i \log\left(\frac{\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right) + (1 - \widetilde{y}_i) \log\left(\frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)} \right) \right] \\
= & \sum_{i=1}^{n} \left[\widetilde{y}_i \log\Big(\exp(-\beta_0 - \boldsymbol{\beta}^T \boldsymbol{x}_i) \Big) + \log\Big(1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \Big) \right]
\end{aligned}
$$

If we take the expectation with respect to y_i, we have

$$
\begin{aligned}
& \mathbb{P}(\widetilde{y}_i = 1 | \boldsymbol{x}_i) \log(1 + \exp(-\beta_0 - \boldsymbol{\beta}^T \boldsymbol{x}_i)) \\
+ & \mathbb{P}(\widetilde{y}_i = 0 | \boldsymbol{x}_i) \log(1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i))
\end{aligned}
$$

Conclusion

- Logistic Regression is an special example of using **Logistic loss** for classification.
- Logistic Regression assumes that the $\mathbb{P}(\widetilde{Y} = 1|\boldsymbol{X} = \boldsymbol{x})$ as

$$\eta(\boldsymbol{x}) = \mathbb{P}(\widetilde{Y} = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

- The optimal function minimize the logistic risk is $f^*(\boldsymbol{x}) = \log(\frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$.

Relationship: Logistic Loss and Logistic Regression

Training Error (0-1): $\frac{1}{n}\sum_{i=1}^{n} I(f(x_i) \neq y_i)$

$\Downarrow$

Training Error (Logistic): $\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$

$\Downarrow$

Training Error (Logistic): $\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$ $\Leftarrow$ Consider: $f(x_i)=\beta_0 + \beta^T x$

Logistic Regression

Contents

Classification

- A typical dataset in classification $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.
  - $\boldsymbol{x}_i$: the covariate vector of i-th instance
  - $y_i \in \{0, 1\}$: binary label of i-th instance

- Bayes classifier f^*:

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > 1/2 \\ 0 & \text{if } \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) < 1/2 \end{cases}$$

- Minimal risk $R(f^*)$:

$$R(f^*) = \mathbb{E}\Big[f^*(\boldsymbol{X}) \neq Y\Big] = \mathbb{E}\Big[\min(\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X}))\Big],$$

where $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})$.

How can we construct classifier?

- **Discriminative models**
  - Discriminative modeling studies the $P(Y|\boldsymbol{X})$
  - Examples: Logistic regression (LR)
- **Generative models**
  - Generative models studies the joint probability distribution $\mathbb{P}(\boldsymbol{X}, Y)$
  - Examples: linear discriminant analysis and quadratic discriminant analysis

Basics of Generative models

- LDA and QDA are **generative models**, we need to consider the structure of $\mathbb{P}(\boldsymbol{X}, Y)$

$$\mathbb{P}(\boldsymbol{X}, Y) = \mathbb{P}(\boldsymbol{X}|Y)\mathbb{P}(Y)$$
$$\mathbb{P}(\boldsymbol{X}, Y) = \mathbb{P}(Y|\boldsymbol{X})\mathbb{P}(\boldsymbol{X})$$

An alternative look

Let $k \in \{0, 1\}$. We can develop an alternative formulation of $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x})$ from the definition of conditional probability.

$$\begin{aligned}\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, Y = k)}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})} &= \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k) \cdot \mathbb{P}(Y = k)}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})} \\ &= \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k) \cdot \mathbb{P}(Y = k)}{\sum_{k=0}^{1} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k) \cdot \mathbb{P}(Y = k)}\end{aligned}$$

- $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$ the marginal distribution
- $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x})$: given $\boldsymbol{X} = \boldsymbol{x}$ the probability that outcome $Y = k$.

Banknote Dataset

conterfeit	Length	Left	Right	Bottom	Top	Diagonal
0	214.70000	129.70000	129.30000	8.60000	9.60000	141.60000
0	215.40000	130.00000	129.90000	8.50000	9.70000	141.40000
0	214.90000	129.40000	129.50000	8.20000	9.90000	141.50000
0	214.50000	129.50000	129.30000	7.40000	10.70000	141.50000
0	214.70000	129.60000	129.50000	8.30000	10.00000	142.00000
0	215.60000	129.90000	129.90000	9.00000	9.50000	141.70000
0	215.00000	130.40000	130.30000	9.10000	10.20000	141.10000
0	214.40000	129.70000	129.50000	8.00000	10.30000	141.20000
0	215.10000	130.00000	129.80000	9.10000	10.20000	141.50000
0	214.70000	130.00000	129.40000	7.80000	10.00000	141.20000
1	214.40000	130.10000	130.30000	9.70000	11.70000	139.80000
1	214.90000	130.50000	130.20000	11.00000	11.50000	139.50000
1	214.90000	130.30000	130.10000	8.70000	11.70000	140.20000
1	215.00000	130.40000	130.60000	9.90000	10.90000	140.30000
1	214.70000	130.20000	130.30000	11.80000	10.90000	139.70000
1	215.00000	130.20000	130.20000	10.60000	10.70000	139.90000
1	215.30000	130.30000	130.10000	9.30000	12.10000	140.20000

Discriminant Analysis (DA)

Discriminant Analysis models $\mathbb{P}(Y|\boldsymbol{X})$ as follows:

- Step 1: Make assumptions on data structure

Discriminant Analysis (DA)

Discriminant Analysis models $\mathbb{P}(Y|\boldsymbol{X})$ as follows:

- Step 1: Make assumptions on data structure
  - Let $\pi_k = \mathbb{P}(Y = k)$ be the prior probability of category $k = 0, 1$

Discriminant Analysis (DA)

Discriminant Analysis models $\mathbb{P}(Y|\boldsymbol{X})$ as follows:

- Step 1: Make assumptions on data structure
  - Let $\pi_k = \mathbb{P}(Y = k)$ be the prior probability of category $k = 0, 1$
  - Suppose that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k)$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$

Discriminant Analysis (DA)

Discriminant Analysis models $\mathbb{P}(Y|\boldsymbol{X})$ as follows:

- Step 1: Make assumptions on data structure
  - Let $\pi_k = \mathbb{P}(Y = k)$ be the prior probability of category $k = 0, 1$
  - Suppose that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k)$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$

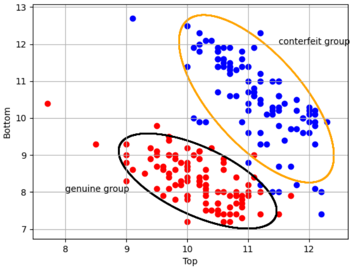


Figure: Black ellipsoid: covariance structure of genuine group. Green ellipsoid: covariance structure of the counterfeit group

Discriminant Analysis (DA)

- $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\Big(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \Big),$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}, \boldsymbol{\mu}_k = \begin{pmatrix} \mu_{1,k} \\ \mu_{2,k} \\ \vdots \\ \mu_{p,k} \end{pmatrix}, \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma^2_{1,1,k} & \sigma^2_{1,2,k} & \cdots & \sigma^2_{2,2,k} \\ \sigma^2_{2,1,k} & \sigma^2_{2,2,k} & \cdots & \sigma^2_{2,p,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{p,1,k} & \sigma^2_{p,2,k} & \cdots & \sigma^2_{p,p,k} \end{pmatrix}.$$

Discriminant Analysis (DA)

- Step 2: We use the Bayes' theorem to compute $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}), k = 0, 1.$

$$\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k)}{\pi_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1) + \pi_0 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 0)}$$

Discriminant Analysis (DA)

- Step 2: We use the Bayes' theorem to compute $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}), k = 0, 1.$

$$\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k)}{\pi_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1) + \pi_0 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 0)}$$

Question: What is the difference between Linear and Quadratic discriminant analyses?

Discriminant Analysis (DA)

- Step 2: We use the Bayes' theorem to compute $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}), k = 0, 1$.

$$\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = k)}{\pi_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1) + \pi_0 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 0)}$$

Question: What is the difference between Linear and Quadratic discriminant analyses?

- **Linear** Discriminant Analysis (LDA) assumes that the classes have a common covariance matrix. In other words, that is $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$
- **Quadratic** Discriminant Analysis (QDA) does not assumes this. So, we have a covariance matrix $\boldsymbol{\Sigma}_0$ for class 0 and $\boldsymbol{\Sigma}_1$ for class 1.

Linear Discriminant Analysis (LDA)

Three Assumptions in LDA

Linear Discriminant Analysis (LDA)

Three Assumptions in LDA

1. Multivariate normal distribution for each group, that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ is multivariate normal

Linear Discriminant Analysis (LDA)

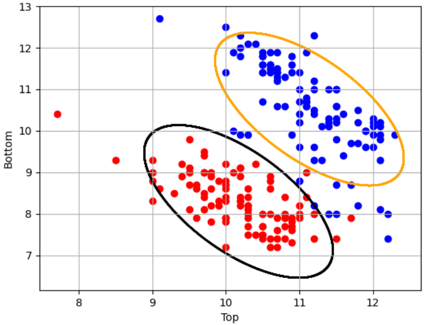
Three Assumptions in LDA

1. Multivariate normal distribution for each group, that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ is multivariate normal
2. They have different mean vectors

Linear Discriminant Analysis (LDA)

Three Assumptions in LDA

1. Multivariate normal distribution for each group, that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ is multivariate normal
2. They have different mean vectors
3. **Same covariance matrices**



Use LDA for classification

We make predictions using LDA as follows:

$$f_{LDA}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} > 0.5 \\ 0, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} \leq 0.5 \end{cases}$$

Use LDA for classification

We make predictions using LDA as follows:

$$f_{LDA}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} > 0.5 \\ 0, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} \leq 0.5 \end{cases}$$

Conclusions we can make

1. Similar to the Bayes classifier, we classify to the most probable class using the posterior probability

Use LDA for classification

We make predictions using LDA as follows:

$$f_{LDA}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} > 0.5 \\ 0, & \text{if } \frac{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\pi_1 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)+\pi_0 \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} \leq 0.5 \end{cases}$$

Conclusions we can make

1. Similar to the Bayes classifier, we classify to the most probable class using the posterior probability
2. The decision boundary can be easily derived as

$$\frac{\pi_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1)}{\pi_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1) + \pi_0 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 0)} = 1/2$$

$$\Leftrightarrow \log \frac{\pi_1}{\pi_0} + \log \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 1)}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|Y = 0)} = 0.$$

Decision boundary in LDA

A closer look at the decision boundary.

$$\log \frac{\pi_1}{\pi_0} + \log \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = 1)}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = 0)} = 0$$

$$\Updownarrow$$

$$\log \frac{\pi_1}{\pi_0} + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 = 0$$

$$\Updownarrow$$

$$\log \frac{\pi_1}{\pi_0} + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = 0.$$

Decision boundary in LDA

A closer look at the decision boundary.

$$\log\frac{\pi_1}{\pi_0} + \log\frac{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=1)}{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=0)} = 0$$

$$\Updownarrow$$

$$\log\frac{\pi_1}{\pi_0} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) - \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 = 0$$

$$\Updownarrow$$

$$\log\frac{\pi_1}{\pi_0} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) - \frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu}_0) = 0.$$

The decision boundary can be written as (a linear equation)

$$\boldsymbol{x}^T C_1(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}) + C_2(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}) = 0,$$

where $C_1(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$ and
$C_2(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}) = \log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\boldsymbol{\mu_1}+\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu}_0)$.

Parameter estimation in LDA

Thanks to the formulation of LDA, we can easily estimate its parameters.

- The prior probability π_0 and π_1.

$$\widehat{\pi}_0 = \frac{n_0}{n_0 + n_1} \text{ and } \widehat{\pi}_1 = \frac{n_1}{n_0 + n_1},$$

  where n_k is the number of observations in the training data set that belong to class.

Parameter estimation in LDA

Thanks to the formulation of LDA, we can easily estimate its parameters.

- The prior probability π_0 and π_1.

$$\widehat{\pi}_0 = \frac{n_0}{n_0 + n_1} \text{ and } \widehat{\pi}_1 = \frac{n_1}{n_0 + n_1},$$

  where n_k is the number of observations in the training data set that belong to class.
- The means are estimated as

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i:y_i=k} \boldsymbol{x}_i, k = 0, 1$$

Parameter estimation in LDA

Thanks to the formulation of LDA, we can easily estimate its parameters.

- The prior probability π_0 and π_1.

$$\widehat{\pi}_0 = \frac{n_0}{n_0 + n_1} \text{ and } \widehat{\pi}_1 = \frac{n_1}{n_0 + n_1},$$

  where n_k is the number of observations in the training data set that belong to class.

- The means are estimated as

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i:y_i=k} \boldsymbol{x}_i, k = 0, 1$$

- The covariance matrices are estimated as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \sum_{k=0}^{1} \sum_{i:y_i=k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T$$

Quadratic Discriminant Analysis (QDA)

Three Assumptions in QDA

1. Multivariate normal distribution for each group, that $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ is multivariate normal
2. They have different mean vectors
3. **Different covariance matrices**

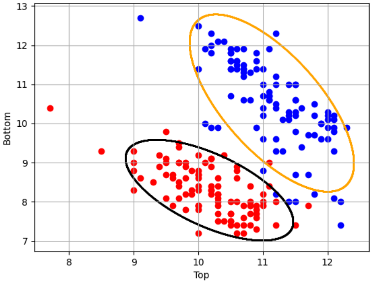


Figure: Different covariance structures

Decision boundary in QDA

We follow a similar analysis of QDA as with LDA. After some algebra, we arrive to the following (interesting) equation:

$$\log \frac{\pi_1}{\pi_0} - \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} + \boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) + \cdots = 0$$

Conclusion

- The decision boundary in QDA is a quadratic function

LDA vs QDA

The difference between LDA and QDA can be summarized as

- LDA is simpler than QDA. (LDA is a special case of QDA)
- QDA needs to estimate more parameters. One covariance matrix for each class.
- LDA is much less flexible than QDA, but this also means that it has low variance
- If the assumptions of LDA do not hold, then it can lead to poor estimates and so, a high bias.

Exercise: Prediction of counterfeit banknotes

conterfeit	Length	Left	Right	Bottom	Top	Diagonal
0	214.70000	129.70000	129.30000	8.60000	9.60000	141.60000
0	215.40000	130.00000	129.90000	8.50000	9.70000	141.40000
0	214.90000	129.40000	129.50000	8.20000	9.90000	141.50000
0	214.50000	129.50000	129.30000	7.40000	10.70000	141.50000
0	214.70000	129.60000	129.50000	8.30000	10.00000	142.00000
0	215.60000	129.90000	129.90000	9.00000	9.50000	141.70000
0	215.00000	130.40000	130.30000	9.10000	10.20000	141.10000
0	214.40000	129.70000	129.50000	8.00000	10.30000	141.20000
0	215.10000	130.00000	129.80000	9.10000	10.20000	141.50000
0	214.70000	130.00000	129.40000	7.80000	10.00000	141.20000
1	214.40000	130.10000	130.30000	9.70000	11.70000	139.80000
1	214.90000	130.50000	130.20000	11.00000	11.50000	139.50000
1	214.90000	130.30000	130.10000	8.70000	11.70000	140.20000
1	215.00000	130.40000	130.60000	9.90000	10.90000	140.30000
1	214.70000	130.20000	130.30000	11.80000	10.90000	139.70000
1	215.00000	130.20000	130.20000	10.60000	10.70000	139.90000
1	215.30000	130.30000	130.10000	9.30000	12.10000	140.20000

- Length: length of banknote (mm)
- Left: length of left edge (mm)
- Right: length of right edge (mm)
- Top: distance from the image to top edge
- Bottom: distance from image to bottom
- Diagonal: length of diagonal (mm)
- counterfeit: 1 means counterfeit and 0 means genuine

Exercise: Prediction of counterfeit banknotes using R

- Step 1: Loading the dataset and split the dataset into training set and testing set:

```
library(mclust)
# Load the data set.
data(banknote)
banknote$Status<-factor(banknote$Status,levels=c("genuine", "counterfeit"))
# Split into training and test data.
set.seed(123) # Set seed to reproduce results.
i <- 1:dim(banknote)[1]
# Generate a random sample.
i.train <- sample(i, 130, replace = F) # 130 samples are used for training
bn.train <- banknote[i.train,] # training dataset
bn.test <- banknote[-i.train,] # testing dataset
```

- Step 2: Implement LDA and make prediction by LDA

```
library(MASS)
lda.mod <- lda(Status~Length + Right + Left + Top, data = bn.train) # Fit a LD
pred.lda.test <- predict(lda.mod,bn.test[,-1])

table('Reference' = bn.test[,1], "Predicted" =
       pred.lda.test$class)
```

Exercise: Prediction of counterfeit banknotes using R

- Result:

```
> table('Reference' = bn.test[,1], "Predicted" =
+         pred.lda.test$class)
             Predicted
Reference     genuine counterfeit
  genuine          30           3
  counterfeit       4          33
```

- Conclusion: The prediction accuracy of LDA is $(30+33)/70=0.9$.

Exercise: Prediction of counterfeit banknotes using R

- Implementation of QDA

```
qda.mod <- qda(Status~Length + Right + Left + Top, data = bn.train)
pred.qda.test <- predict(qda.mod, bn.test[,-1])
table('Reference' = bn.test[,1],
      "Predicted" = pred.qda.test$class)
```

- Result:

```
> table('Reference' = bn.test[,1],
+       "Predicted" = pred.qda.test$class)
             Predicted
Reference     genuine counterfeit
  genuine          29           4
  counterfeit       3          34
```

- Conclusion: The prediction accuracy of LDA is $(29+34)/70=0.9$. No improvement is observed.

Some questions

- Can you finish an implementation of LDA and QDA in R or Python?
- Can you summarize the difference between Logistic regression, LDA, and QDA?
- What is the difference between generative model and discriminative model?
- Is K-nearest neighbor classifier a generative model or a discriminative model?