Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 17
March 10, 2023

QDA

Multiclass LDA Suppose the data set is divided into $K > 2$ disjoint classes

Bayes's rule classifier Let

$$
\begin{aligned}
P(\mathbf{X} \in \Pi_i) &= \pi_i, i = 1, \ldots, K \\
P(\mathbf{X} = \mathbf{x} | \, \mathbf{X} \in \Pi_i) &= f_i(\mathbf{x}), \\
p(\Pi_i | \, \mathbf{x}) &= P(\mathbf{X} \in \Pi_i | \, \mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k}
\end{aligned}
$$

QDA

Multiclass LDA Suppose the data set is divided into $K > 2$ disjoint classes

Bayes's rule classifier Let

$$
\begin{aligned}
P(\mathbf{X} \in \Pi_i) &= \pi_i, i = 1, \ldots, K \\
P(\mathbf{X} = \mathbf{x} | \, \mathbf{X} \in \Pi_i) &= f_i(\mathbf{x}), \\
p(\Pi_i | \, \mathbf{x}) &= P(\mathbf{X} \in \Pi_i | \, \mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k}
\end{aligned}
$$

Assign $\mathbf{x}$ to Π_i if

$$f_i(\mathbf{x})\pi_i = \max_{1 \le j \le K} f_j(\mathbf{x})\pi_j$$

QDA

Multiclass LDA Suppose the data set is divided into $K > 2$ disjoint classes

Bayes's rule classifier Let

$$\begin{aligned} P(\mathbf{X} \in \Pi_i) &= \pi_i, i = 1, \ldots, K \\ P(\mathbf{X} = \mathbf{x} | \, \mathbf{X} \in \Pi_i) &= f_i(\mathbf{x}), \\ p(\Pi_i | \, \mathbf{x}) &= P(\mathbf{X} \in \Pi_i | \, \mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k} \end{aligned}$$

Assign $\mathbf{x}$ to Π_i if

$$f_i(\mathbf{x})\pi_i = \max_{1 \le j \le K} f_j(\mathbf{x})\pi_j$$

If the maximizer is not unique then assign $\mathbf{x}$ randomly to break the tie.

LDA

Thus $\mathbf{x}$ gets assigned to Π_i if $f_i(\mathbf{x})\pi_i > f_j(\mathbf{x})\pi_j$ for all $i \neq j$ i.e. $\log_e(f_i(\mathbf{x})\pi_i) > \log_e(f_j(\mathbf{x})\pi_j)$. Finally, define

$$L_{ij}(\mathbf{x}) = \log_e \left[\frac{f_i(\mathbf{x})\pi_i}{f_j(\mathbf{x})\pi_j} \right]$$

and assign $\mathbf{x}$ to Π_i if $L_{ij}(\mathbf{x}) > 0$, otherwise assign $\mathbf{x}$ to Π_j.

LDA

Thus $\mathbf{x}$ gets assigned to Π_i if $f_i(\mathbf{x})\pi_i > f_j(\mathbf{x})\pi_j$ for all $i \neq j$ i.e. $\log_e(f_i(\mathbf{x})\pi_i) > \log_e(f_j(\mathbf{x})\pi_j)$. Finally, define

$$L_{ij}(\mathbf{x}) = \log_e \left[\frac{f_i(\mathbf{x})\pi_i}{f_j(\mathbf{x})\pi_j} \right]$$

and assign $\mathbf{x}$ to Π_i if $L_{ij}(\mathbf{x}) > 0$, otherwise assign $\mathbf{x}$ to Π_j.

The classification regions in the feature space $\mathbb{R}^d$ are

$$R_i = \{\mathbf{x} \in \mathbb{R}^d : L_{ij}(\mathbf{x}) > 0, j = 1, 2 \ldots, K, j \neq i\},$$

$i = 1, 2, \ldots, K$

LDA

Assuming that $f_i(\mathbf{x}) \sim \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ we have

$$L_{ij}(\mathbf{x}) = b_{0ij} + \mathbf{b}_{ij}^T \mathbf{x}$$

where

$$\begin{aligned} \mathbf{b}_{ij} &= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} \\ b_{0ij} &= -\frac{1}{2}\left[\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j\right] + \log_e(\pi_i/\pi_j) \end{aligned}$$

LDA

Assuming that $f_i(\mathbf{x}) \sim \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ we have

$$L_{ij}(\mathbf{x}) = b_{0ij} + \mathbf{b}_{ij}^T \mathbf{x}$$

where

$$\begin{aligned} \mathbf{b}_{ij} &= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} \\ b_{0ij} &= -\frac{1}{2}\left[\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j\right] + \log_e(\pi_i/\pi_j) \end{aligned}$$

Conclusion Since $L_{ij}(\mathbf{x})$ is linear in $\mathbf{x}$, the regions $R_i, i = 1, \ldots, K$ partition the feature space by means of hyperplanes

LDA

Question How to implement in real data?

$\rightarrow$ The mean vectors $\boldsymbol{\mu}_i, 1 \leq i \leq K$ and the common covariance matrix $\boldsymbol{\Sigma}$ are not known

LDA

Question How to implement in real data?

→ The mean vectors $\boldsymbol{\mu}_i, 1 \leq i \leq K$ and the common covariance matrix $\boldsymbol{\Sigma}$ are not known

→ There are $Kd + d(d+1)/2$ parameters we need to estimate from learning the sampled feature vectors

LDA

Question How to implement in real data?

- → The mean vectors $\boldsymbol{\mu}_i, 1 \leq i \leq K$ and the common covariance matrix $\boldsymbol{\Sigma}$ are not known
- → There are $Kd + d(d+1)/2$ parameters we need to estimate from learning the sampled feature vectors
- → Let $\mathbf{x}_{ij}$ denote the data points in the Π_i, $1 \leq j \leq n_i$

LDA

Question How to implement in real data?

- → The mean vectors $\boldsymbol{\mu}_i, 1 \leq i \leq K$ and the common covariance matrix $\boldsymbol{\Sigma}$ are not known
- → There are $Kd + d(d+1)/2$ parameters we need to estimate from learning the sampled feature vectors
- → Let $\mathbf{x}_{ij}$ denote the data points in the Π_i, $1 \leq j \leq n_i$
- → Set $\mathcal{X}_i = [\mathbf{x}_{i1}\, \mathbf{x}_{i2}\, \ldots \mathbf{x}_{in_i}]_{d \times n_i}$, $1 \leq i \leq K$

LDA

Question How to implement in real data?

- → The mean vectors $\boldsymbol{\mu}_i, 1 \le i \le K$ and the common covariance matrix $\boldsymbol{\Sigma}$ are not known
- → There are $Kd + d(d+1)/2$ parameters we need to estimate from learning the sampled feature vectors
- → Let $\mathbf{x}_{ij}$ denote the data points in the Π_i, $1 \le j \le n_i$
- → Set $\mathcal{X}_i = [\mathbf{x}_{i1}\, \mathbf{x}_{i2}\, \ldots \mathbf{x}_{in_i}]_{d \times n_i}$, $1 \le i \le K$
- → Set $n = \sum_{i=1}^{K} n_i$ and

$$\mathcal{X} = [\mathcal{X}_1\, \mathcal{X}_2\, \ldots \mathcal{X}_K] = [\mathbf{x}_{11}\, \ldots\, \mathbf{x}_{1n_1} \ldots \mathbf{x}_{K1}\, \ldots\, \mathbf{x}_{Kn_K}]$$

- → Set $\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n_i} \mathcal{X}_i \mathbf{1}_{n_i}$, $1 \le i \le K$ and

$$\overline{\mathcal{X}} = [\underbrace{\overline{x}_1, \ldots, \overline{x}_1}_{n_1}, \ldots, \underbrace{\overline{x}_K, \ldots, \overline{x}_K}_{n_K}]_{d \times n}$$

LDA

$\rightarrow$ Let

$$\mathcal{X}_c = \mathcal{X} - \overline{\mathcal{X}} = (\mathcal{X}_1 C_{n_1} \ldots \mathcal{X}_K C_{n_K}) \in \mathbb{R}^{d \times n}$$

where $C_{n_j}, j = 1, 2, \ldots, K$ is the centering matrix.

LDA

→ Let

$$\mathcal{X}_c = \mathcal{X} - \overline{\mathcal{X}} = (\mathcal{X}_1 C_{n_1} \ldots \mathcal{X}_K C_{n_K}) \in \mathbb{R}^{d\times n}$$

where $C_{n_j}, j = 1, 2, \ldots, K$ is the centering matrix.

→ Then compute

$$S = \mathcal{X}_c \mathcal{X}_c^T = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T.$$

LDA

→ Let

$$\mathcal{X}_c = \mathcal{X} - \overline{\mathcal{X}} = (\mathcal{X}_1 C_{n_1} \dots \mathcal{X}_K C_{n_K}) \in \mathbb{R}^{d \times n}$$

where $C_{n_j}, j = 1, 2, \dots, K$ is the centering matrix.

→ Then compute

$$S = \mathcal{X}_c \mathcal{X}_c^T = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T.$$

→ Consider $\mathbf{x}_{ij} - \overline{\mathbf{x}} = (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i) + (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, where

$$\overline{\mathbf{x}} = \frac{1}{n} \mathcal{X} \mathbf{1}_n = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \mathbf{x}_{ij} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \dots, \overline{\mathbf{x}}_d)^T$$

is the overall mean vector ignoring class identifiers

LDA

Then

Source of variation	df	Sum of squares matrix
Between classes	$K-1$	$S_b = \sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$
Within classes	$n-K$	$S_w = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$
Total	$n-1$	$S_{total} = S_b + S_w$ $= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T$

Table: Multivariate analysis of variance (MANOVA)

LDA

Then

Source of variation	df	Sum of squares matrix
Between classes	$K-1$	$S_b = \sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$
Within classes	$n-K$	$S_w = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$
Total	$n-1$	$S_{total} = S_b + S_w$ $=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T$

Table: Multivariate analysis of variance (MANOVA)

The total covariance matrix of the observations, S_{total}, having $n-1$ degrees of freedom (df) and calculated by ignoring the class identity, formed by the between-class covariance/scatter matrix S_b and the pooled within-class covariance/scatter matrix, S_w

LDA

An unbiased estimator of the common covariance matrix is then given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-K} S_w$$

LDA

An unbiased estimator of the common covariance matrix is then given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-K} S_w$$

Thus setting $\widehat{L}_{ij}(\mathbf{x}) = \widehat{b}_{0ij} + \widehat{\mathbf{b}}_{ij}^T \mathbf{x}$, where

$$\begin{aligned} \widehat{\mathbf{b}}_{ij} &= (\overline{\mathbf{x}}_i - {}_j)^T \widehat{\boldsymbol{\Sigma}}^{-1} \\ \widehat{b}_{0ij} &= -\frac{1}{2}\left\{\overline{\mathbf{x}}_i^T \widehat{\boldsymbol{\Sigma}}^{-1} \overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j^T \widehat{\boldsymbol{\Sigma}}^{-1} \overline{\mathbf{x}}_j\right\} + \log_e \frac{n_i}{n} - \log_e \frac{n_j}{n} \end{aligned}$$

LDA

An unbiased estimator of the common covariance matrix is then given by

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-K} S_w$$

Thus setting $\widehat{L}_{ij}(\mathbf{x}) = \widehat{b}_{0ij} + \widehat{\mathbf{b}}_{ij}^T \mathbf{x}$, where

$$\begin{aligned} \widehat{\mathbf{b}}_{ij} &= (\overline{\mathbf{x}}_i - {}_j)^T \widehat{\mathbf{\Sigma}}^{-1} \\ \widehat{b}_{0ij} &= -\frac{1}{2}\left\{\overline{\mathbf{x}}_i^T \widehat{\mathbf{\Sigma}}^{-1} \overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j^T \widehat{\mathbf{\Sigma}}^{-1} \overline{\mathbf{x}}_j\right\} + \log_e \frac{n_i}{n} - \log_e \frac{n_j}{n} \end{aligned}$$

The classification rule Assign $\mathbf{x}$ to Π_i if $\widehat{L}_{ij}(\mathbf{x}) > 0$, $j = 1, 2, \ldots, K, j \neq i$

QDA

For Quadratic Discriminant Analysis, set

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T, i = 1, 2, \ldots, K$$

QDA

For Quadratic Discriminant Analysis, set

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T, i = 1, 2, \ldots, K$$

Note that if d is large then the number of distinct parameters $Kd + kd(d+1)/2$ are to be estimated, hence this could be a huge increase compare to LDA. The $Q(\mathbf{x})$ will be similar to the 2-class problem.

QDA

For Quadratic Discriminant Analysis, set

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T, i = 1, 2, \ldots, K$$

Note that if d is large then the number of distinct parameters $Kd + kd(d+1)/2$ are to be estimated, hence this could be a huge increase compare to LDA. The $Q(\mathbf{x})$ will be similar to the 2-class problem.

The method that we have followed is called maximum likelihood estimation