Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 14
March 2, 2023

Linear Discriminant Analysis

→ Suppose each observation $\boldsymbol{x}_i$ comes from several, say K classes having similar characteristics

Linear Discriminant Analysis

→ Suppose each observation $\boldsymbol{x}_i$ comes from several, say K classes having similar characteristics

→ Each class has a level and the observations are lebeled

Linear Discriminant Analysis

→ Suppose each observation $\boldsymbol{x}_i$ comes from several, say K classes having similar characteristics

→ Each class has a level and the observations are lebeled

Goal

△ **Discrimination** Use the information of the labeled observations in a learning set to construct a *classifier* which will separate the predefined classes

Linear Discriminant Analysis

→ Suppose each observation $\boldsymbol{x}_i$ comes from several, say K classes having similar characteristics

→ Each class has a level and the observations are lebeled

Goal

△ **Discrimination** Use the information of the labeled observations in a learning set to construct a *classifier* which will separate the predefined classes

△ **Classification** For a set of measurements on a new unlebeled observation, use the classifier to predict the class of the observation

LDA

Question How many classifiers are needed to differentiate the classes and to predict the class for a future observation?

LDA

Question How many classifiers are needed to differentiate the classes and to predict the class for a future observation?
Example: medical diagnosis problem, handwriting, gene expression data

LDA

Question How many classifiers are needed to differentiate the classes and to predict the class for a future observation?

Example: medical diagnosis problem, handwriting, gene expression data

→ Assume that the population $\mathcal{P}$ is partitioned into K unordered classes: $\Pi_1, \ldots, \Pi_K$

LDA

Question How many classifiers are needed to differentiate the classes and to predict the class for a future observation?

Example: medical diagnosis problem, handwriting, gene expression data

- $\rightarrow$ Assume that the population $\mathcal{P}$ is partitioned into K unordered classes: $\Pi_1, \ldots, \Pi_K$
- $\rightarrow$ Define *feature vector*

$$\mathbf{X} = (x_1, x_2, \ldots, x_d)^T$$

where x_i denotes a measurement

LDA

Binary classification: $K = 2$

→ Let $P(\mathbf{X} \in \Pi_i) = \pi_i$, $i = 1, 2$ denote the *prior probabilities* such that a randomly selected observation $\mathbf{X} = \mathbf{x}$ belongs to either Π_1 or Π_2.

LDA

Binary classification: $K = 2$

- $\rightarrow$ Let $P(\mathbf{X} \in \Pi_i) = \pi_i$, $i = 1, 2$ denote the *prior probabilities* such that a randomly selected observation $\mathbf{X} = \mathbf{x}$ belongs to either Π_1 or Π_2.
- $\rightarrow$ Let the conditional pdf of $\mathbf{X}$ for class i by

$$P(\mathbf{X} = \mathbf{x} \,|\, \mathbf{X} \in \Pi_i) = f_i(\mathbf{x}), i = 1, 2$$

LDA

Binary classification: $K = 2$

- $\rightarrow$ Let $P(\mathbf{X} \in \Pi_i) = \pi_i$, $i = 1, 2$ denote the *prior probabilities* such that a randomly selected observation $\mathbf{X} = \mathbf{x}$ belongs to either Π_1 or Π_2.
- $\rightarrow$ Let the conditional pdf of $\mathbf{X}$ for class i by

$$P(\mathbf{X} = \mathbf{x} \,|\, \mathbf{X} \in \Pi_i) = f_i(\mathbf{x}), i = 1, 2$$

- $\rightarrow$ By Bayes's theorem, the *posterior probability* is given by

$$p(\Pi_i|\mathbf{x}) = P(\mathbf{X} \in \Pi_i|\mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2}$$

that the observed $\mathbf{x}$ belongs to Π_i, $i = 1, 2$

LDA

Classification strategy

Bayes's rule classifier
Assign $\mathbf{x}$ to Π_1 if

$$r = \frac{p(\Pi_1|\mathbf{x})}{p(\Pi_2|\mathbf{x})} > 1 \text{ i.e. } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

and assign $\mathbf{x}$ to Π_2 otherwise.

LDA

Classification strategy

Bayes's rule classifier
Assign $\mathbf{x}$ to Π_1 if

$$r = \frac{p(\Pi_1|\mathbf{x})}{p(\Pi_2|\mathbf{x})} > 1 \text{ i.e. } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

and assign $\mathbf{x}$ to Π_2 otherwise.
If

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\pi_2}{\pi_1}$$

then decide the class by tossing a coin

LDA

Classification strategy

Bayes's rule classifier

Assign $\mathbf{x}$ to Π_1 if

$$r = \frac{p(\Pi_1|\mathbf{x})}{p(\Pi_2|\mathbf{x})} > 1 \text{ i.e. } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

and assign $\mathbf{x}$ to Π_2 otherwise.

If

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\pi_2}{\pi_1}$$

then decide the class by tossing a coin

r is referred to as the 'odds-ratio' that Π_1 is the correct class than Π_2 given the information of $\mathbf{x}$

LDA

Recall: A random d-vector $\mathbf{X}$ is said to have d-variate Gaussian (Normal) distribution with mean vector $\boldsymbol{\mu}$ and positive-definite $d \times d$ covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

LDA

Recall: A random d-vector $\mathbf{X}$ is said to have d-variate Gaussian (Normal) distribution with mean vector $\boldsymbol{\mu}$ and positive-definite $d \times d$ covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Then we write $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

LDA

Recall: A random d-vector $\mathbf{X}$ is said to have d-variate Gaussian (Normal) distribution with mean vector $\boldsymbol{\mu}$ and positive-definite $d \times d$ covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Then we write $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\triangle = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

is referred to as the *Mahalanobis distance* from $\mathbf{x}$ to $\boldsymbol{\mu}$

LDA

Gaussian LDA, Fisher 1936

Let $f_1(x)$ and $f_2(x)$ be multivariate Gaussian having arbitrary mean vectors and a common covariance matrix Σ:

$$f_1(\cdot) \sim \mathcal{N}_d(\mu_1, \Sigma), \text{ and } f_2(\cdot) \sim \mathcal{N}_d(\mu_2, \Sigma).$$

LDA

Gaussian LDA, Fisher 1936

Let $f_1(x)$ and $f_2(x)$ be multivariate Gaussian having arbitrary mean vectors and a common covariance matrix Σ:

$$f_1(\cdot) \sim \mathcal{N}_d(\mu_1, \Sigma), \text{ and } f_2(\cdot) \sim \mathcal{N}_d(\mu_2, \Sigma).$$

Then

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma^{-1} (\mathbf{x} - \mu_2)\right\}}$$

where the normalization factor $(2\pi)^{-d/2}|\Sigma|^{-1/2}$ got cancelled.

LDA

Gaussian LDA, Fisher 1936

Let $f_1(x)$ and $f_2(x)$ be multivariate Gaussian having arbitrary mean vectors and a common covariance matrix Σ:

$$f_1(\cdot) \sim \mathcal{N}_d(\mu_1, \Sigma), \text{ and } f_2(\cdot) \sim \mathcal{N}_d(\mu_2, \Sigma).$$

Then

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\mu_2)\right\}}$$

where the normalization factor $(2\pi)^{-d/2}|\Sigma|^{-1/2}$ got cancelled.
Then

$$\log_e \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = (\mu_1-\mu_2)^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)$$

LDA

Gaussian LDA, Fisher 1936

Let $f_1(x)$ and $f_2(x)$ be multivariate Gaussian having arbitrary mean vectors and a common covariance matrix Σ:

$$f_1(\cdot) \sim \mathcal{N}_d(\mu_1, \Sigma), \text{ and } f_2(\cdot) \sim \mathcal{N}_d(\mu_2, \Sigma).$$

Then

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\mu_2)\right\}}$$

where the normalization factor $(2\pi)^{-d/2}|\Sigma|^{-1/2}$ got cancelled.
Then

$$\begin{aligned}\log_e \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= (\mu_1-\mu_2)^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2) \\ &= (\mu_1-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\overline{\mu}), \quad \overline{\mu} = (\mu_1+\mu_2)/2\end{aligned}$$

LDA

Further,

$$L(\mathbf{x}) = \log_e \left\{ \frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{z})\pi_2} \right\} = b_0 + \mathbf{b}^T \mathbf{x},$$

a linear function of **x**, where

$$\begin{aligned} \mathbf{b} &= \Sigma^{-1}(\mu_1 - \mu_2) \\ b_0 &= -\frac{1}{2}\left\{\mu_1^T \Sigma^{-1} \mu_1 - \mu_2 \Sigma^{-1} \mu_2\right\} + \log_e(\pi_2/\pi_1) \end{aligned}$$

LDA

Further,

$$L(\mathbf{x}) = \log_e \left\{ \frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{z})\pi_2} \right\} = b_0 + \mathbf{b}^T \mathbf{x},$$

a linear function of $\mathbf{x}$, where

$$\begin{aligned} \mathbf{b} &= \Sigma^{-1}(\mu_1 - \mu_2) \\ b_0 &= -\frac{1}{2}\left\{\mu_1^T \Sigma^{-1}\mu_1 - \mu_2 \Sigma^{-1}\mu_2\right\} + \log_e(\pi_2/\pi_1) \end{aligned}$$

Classification rule Gaussian LDA: Assign $\mathbf{x}$ to Π_1 if $L(\mathbf{x}) > 0$, otherwise assign $\mathbf{x}$ to Π_2.