Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 15
March 3, 2023

LDA

Observation The set of all vectors $\mathbf{x}$ such that $L(\mathbf{x}) = 0$ defines a hyperplane which divides the classes!! The term $\mathbf{b}^T\mathbf{x}$ in $L(\mathbf{x})$ is called the Fisher's linear discriminant function (LDF).

LDA

Observation The set of all vectors $\mathbf{x}$ such that $L(\mathbf{x}) = 0$ defines a hyperplane which divides the classes!! The term $\mathbf{b}^T\mathbf{x}$ in $L(\mathbf{x})$ is called the Fisher's linear discriminant function (LDF).

Let us denote the partition of the feature space $\mathbb{R}^d$ into two regions R_1 and R_2. If $\mathbf{x}$ falls into R_1 then it is classified as belonging to Π_1, whereas if $\mathbf{x}$ falls into region R_2, it is classified into Π_2.

LDA

Observation The set of all vectors $\mathbf{x}$ such that $L(\mathbf{x}) = 0$ defines a hyperplane which divides the classes!! The term $\mathbf{b}^T\mathbf{x}$ in $L(\mathbf{x})$ is called the Fisher's linear discriminant function (LDF).

Let us denote the partition of the feature space $\mathbb{R}^d$ into two regions R_1 and R_2. If $\mathbf{x}$ falls into R_1 then it is classified as belonging to Π_1, whereas if $\mathbf{x}$ falls into region R_2, it is classified into Π_2.

Total misclassification probability Misclassification happens if $\mathbf{x}$ is assigned to Π_2, but actually it belongs to Π_1 or if $\mathbf{x}$ is assigned to Π_1, but actually it belongs to Π_2.

LDA

Squared Mahalanobis distance The squared Mahalanobis distance between Π_1 and Π_2 is defined as

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

LDA

Squared Mahalanobis distance The squared Mahalanobis distance between Π_1 and Π_2 is defined as

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Question What does Mahalanobis distance measure?

LDA

Squared Mahalanobis distance The squared Mahalanobis distance between Π_1 and Π_2 is defined as

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Question What does Mahalanobis distance measure?

Recall Let X a random matrix. For the random matrix $Y = AXB^T + C$, where A, B, C are compatible matrices, $\mathbb{E}(Y) = A\mathbb{E}(X)B^T$ and the covariance matrix of $vec(Y)$ is $\Sigma_{YY} = (A \otimes B)\Sigma_{XX}(A \otimes B)^T$

LDA

Squared Mahalanobis distance The squared Mahalanobis distance between Π_1 and Π_2 is defined as

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Question What does Mahalanobis distance measure?

Recall Let X a random matrix. For the random matrix $Y = AXB^T + C$, where A, B, C are compatible matrices, $\mathbb{E}(Y) = A\mathbb{E}(X)B^T$ and the covariance matrix of $vec(Y)$ is $\Sigma_{YY} = (A \otimes B)\Sigma_{XX}(A \otimes B)^T$

Set $U = \mathbf{b}^T \mathbf{x}$. Then

$$\begin{aligned} \mathbb{E}(U|x \in \Pi_i) &= \mathbf{b}^T \mu_i = (\mu_1 - \mu_2)^T \Sigma^{-1} \mu_i \\ Var(U|x \in \Pi_i) &= \mathbf{b}^T \Sigma \mathbf{b} = (\mu_1 - \mu_2)^T \Sigma^{-1} \Sigma \Sigma^{-1} (\mu_1 - \mu_2) = \triangle^2. \end{aligned}$$

LDA

Squared Mahalanobis distance The squared Mahalanobis distance between Π_1 and Π_2 is defined as

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Question What does Mahalanobis distance measure?

Recall Let X a random matrix. For the random matrix $Y = AXB^T + C$, where A, B, C are compatible matrices, $\mathbb{E}(Y) = A\mathbb{E}(X)B^T$ and the covariance matrix of $vec(Y)$ is $\Sigma_{YY} = (A \otimes B)\Sigma_{XX}(A \otimes B)^T$

Set $U = \mathbf{b}^T \mathbf{x}$. Then

$$\begin{aligned} \mathbb{E}(U|x \in \Pi_i) &= \mathbf{b}^T \mu_i = (\mu_1 - \mu_2)^T \Sigma^{-1} \mu_i \\ Var(U|x \in \Pi_i) &= \mathbf{b}^T \Sigma \mathbf{b} = (\mu_1 - \mu_2)^T \Sigma^{-1} \Sigma \Sigma^{-1} (\mu_1 - \mu_2) = \triangle^2. \end{aligned}$$

Then the total misclassification probability is:

$$P(\mathbf{x} \in R_2 | \mathbf{x} \in \Pi_1)\pi_1 + P(\mathbf{x} \in R_1 | \mathbf{x} \in \Pi_2)\pi_2$$

LDA

Now

$$P(\mathbf{x} \in R_2 | \mathbf{x} \in \Pi_1) = P(L(\mathbf{x}) < 0 | \mathbf{x} \in \Pi_1).$$

LDA

Now

$$P(\mathbf{x} \in R_2 | \mathbf{x} \in \Pi_1) = P(L(\mathbf{x}) < 0 | \mathbf{x} \in \Pi_1).$$

Note that $L(\mathbf{x}) = b_0 + U$. Also,

$$Z = \frac{U - \mathbb{E}(U|\mathbf{x} \in \Pi_i)}{\sqrt{var(U|\mathbf{x} \in \Pi_i)}} \sim \mathcal{N}(0,1)$$

LDA

Now

$$P(\mathbf{x} \in R_2 | \mathbf{x} \in \Pi_1) = P(L(\mathbf{x}) < 0 | \mathbf{x} \in \Pi_1).$$

Note that $L(\mathbf{x}) = b_0 + U$. Also,

$$Z = \frac{U - \mathbb{E}(U|\mathbf{x} \in \Pi_i)}{\sqrt{var(U|\mathbf{x} \in \Pi_i)}} \sim \mathcal{N}(0,1)$$

Then from using the expressions for $\mathbb{E}(U|x \in \Pi_i)$, $Var(U|x \in \Pi_i)$, and b_0 as derived above,

$$\begin{aligned} P(L(\mathbf{x}) < 0 | \mathbf{x} \in \Pi_1) &= P(U < b_0 | \mathbf{x} \in \Pi_1) \\ &= P(Z < \frac{b_0 - (\mu_1 - \mu_2)^T \Sigma^{-1} \mu_i}{\triangle}) \\ &= P(Z < -\frac{\triangle}{2} - \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}) \\ &= \Phi\left(-\frac{\triangle}{2} - \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}\right) \end{aligned}$$

LDA

Similarly we can obtain

$$\begin{aligned} P(\mathbf{x} \in R_1 | \mathbf{x} \in \Pi_2) &= P(L(\mathbf{x}) > 0 | \mathbf{x} \in \Pi_2) \\ &= P(Z < -\frac{\triangle}{2} - \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}) \\ &= \Phi\left(-\frac{\triangle}{2} + \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}\right) \end{aligned}$$

LDA

Similarly we can obtain

$$\begin{aligned} P(\mathbf{x} \in R_1 | \mathbf{x} \in \Pi_2) &= P(L(\mathbf{x}) > 0 | \mathbf{x} \in \Pi_2) \\ &= P(Z < -\frac{\triangle}{2} - \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}) \\ &= \Phi\left(-\frac{\triangle}{2} + \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}\right) \end{aligned}$$

If $\pi_1 = \pi_2 = 1/2$ then

$$P(\mathbf{X} \in R_2 | \mathbf{X} \in \Pi_1) = P(\mathbf{X} \in R_1 | \mathbf{X} \in \Pi_2) = \Phi(-\triangle/2)$$

LDA

Similarly we can obtain

$$\begin{aligned} P(\mathbf{x} \in R_1 | \mathbf{x} \in \Pi_2) &= P(L(\mathbf{x}) > 0 | \mathbf{x} \in \Pi_2) \\ &= P(Z < -\frac{\triangle}{2} - \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}) \\ &= \Phi\left(-\frac{\triangle}{2} + \frac{1}{\triangle} \log_e \frac{\pi_2}{\pi_1}\right) \end{aligned}$$

If $\pi_1 = \pi_2 = 1/2$ then

$$P(\mathbf{X} \in R_2 | \mathbf{X} \in \Pi_1) = P(\mathbf{X} \in R_1 | \mathbf{X} \in \Pi_2) = \Phi(-\triangle/2)$$

Observation Since miscalculation probability depends on $\triangle$, we can write the probability of miscalculation as $P(\triangle)$. Plotting the graph for $\pi_1 = \pi_2 = 1/2$, what is your conclusion?

LDA

Question How do we implement the method in real data?

LDA

Question How do we implement the method in real data?

Observations

$\rightarrow$ Note that $\mu_1, \mu_1, \mathbf{\Sigma}$ are not known

LDA

Question How do we implement the method in real data?

Observations

$\rightarrow$ Note that $\mu_1, \mu_1, \mathbf{\Sigma}$ are not known

$\rightarrow$ In general there are $2d + d(d+2)$ distinct parameters in $\mu_1, \mu_2, \mathbf{\Sigma}$ that can possibly be estimated from learning the data

LDA

Question How do we implement the method in real data?

Observations

→ Note that $\mu_1, \mu_1, \mathbf{\Sigma}$ are not known

→ In general there are $2d + d(d+2)$ distinct parameters in $\mu_1, \mu_2, \mathbf{\Sigma}$ that can possibly be estimated from learning the data

→ Suppose we have a random sample $\mathbf{X}_{1j}, 1 \leq j \leq n_1$, and $\mathbf{X}_{2l}, 1 \leq l \leq n_2$ with values $\mathbf{x}_{1j}$ and $\mathbf{x}_{2l}$ from Π_1 and Π_2 respectively

LDA

Question How do we implement the method in real data?

Observations

→ Note that $\mu_1, \mu_1, \mathbf{\Sigma}$ are not known

→ In general there are $2d + d(d+2)$ distinct parameters in $\mu_1, \mu_2, \mathbf{\Sigma}$ that can possibly be estimated from learning the data

→ Suppose we have a random sample $\mathbf{X}_{1j}, 1 \leq j \leq n_1$, and $\mathbf{X}_{2l}, 1 \leq l \leq n_2$ with values $\mathbf{x}_{1j}$ and $\mathbf{x}_{2l}$ from Π_1 and Π_2 respectively

Sampling methods from a population

→ Mixture sampling - a sample of $n = n_1 + n_2$ is selected so that n_1 and n_2 are randomly selected

LDA

Question How do we implement the method in real data?

Observations

→ Note that $\mu_1, \mu_1, \mathbf{\Sigma}$ are not known

→ In general there are $2d + d(d+2)$ distinct parameters in $\mu_1, \mu_2, \mathbf{\Sigma}$ that can possibly be estimated from learning the data

→ Suppose we have a random sample $\mathbf{X}_{1j}, 1 \leq j \leq n_1$, and $\mathbf{X}_{2l}, 1 \leq l \leq n_2$ with values $\mathbf{x}_{1j}$ and $\mathbf{x}_{2l}$ from Π_1 and Π_2 respectively

Sampling methods from a population

→ Mixture sampling - a sample of $n = n_1 + n_2$ is selected so that n_1 and n_2 are randomly selected

→ Separate sampling - a sample of n_i is randomly selected from $\Pi_i, i = 1, 2$ and $n = n_1 + n_2$

LDA

Estimation of parameters The ML estimates of μ_i, $i = 1, 2$ and $\mathbf{\Sigma}$ are

LDA

Estimation of parameters The ML estimates of $\mu_i,\ i = 1, 2$ and $\mathbf{\Sigma}$ are

$$\begin{aligned} \widehat{\mu}_i &= \bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, i = 1, 2 \text{ and} \\ \widehat{\Sigma} &= \frac{1}{n}S,\ S = S_1 + S_2,\ S_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \end{aligned}$$

LDA

Estimation of parameters The ML estimates of $\mu_i,\ i = 1, 2$ and $\mathbf{\Sigma}$ are

$$\begin{aligned}\widehat{\mu}_i &= \bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, i = 1, 2 \text{ and} \\ \widehat{\Sigma} &= \frac{1}{n}S,\ S = S_1 + S_2,\ S_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T\end{aligned}$$

Note that for unbiased estimator of $\mathbf{\Sigma}$, we can divide S by its degree of freedom $n - 2 = n_1 + n_2 - 2$ rather than n to make $\widehat{\mathbf{\Sigma}}$

LDA

The probabilities π_1, π_2 can be chosen based on past experiences or can be estimated as

$$\widehat{\pi}_i = \frac{n_i}{n}, i = 1, 2$$

LDA

The probabilities π_1, π_2 can be chosen based on past experiences or can be estimated as

$$\widehat{\pi}_i = \frac{n_i}{n}, i = 1, 2$$

Then $\widehat{L}(\mathbf{x}) = \widehat{b}_0 + \widehat{\mathbf{b}}^T\mathbf{x}$, where

$$\begin{aligned} \widehat{\mathbf{b}} &= \widehat{\mathbf{\Sigma}}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \\ \widehat{b}_0 &= -\frac{1}{2}\left[\overline{\mathbf{x}}_1^T\widehat{\mathbf{\Sigma}}^{-1}\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2^T\widehat{\mathbf{\Sigma}}^{-1}\overline{\mathbf{x}}_2\right] + \log_e\frac{n_1}{n} - \log_e\frac{n_2}{n} \end{aligned}$$

are ML estimates of $\mathbf{b}$ and b_0 respectively.

LDA

The probabilities π_1, π_2 can be chosen based on past experiences or can be estimated as

$$\widehat{\pi}_i = \frac{n_i}{n}, i = 1, 2$$

Then $\widehat{L}(\mathbf{x}) = \widehat{b}_0 + \widehat{\mathbf{b}}^T\mathbf{x}$, where

$$\begin{aligned} \widehat{\mathbf{b}} &= \widehat{\mathbf{\Sigma}}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \\ \widehat{b}_0 &= -\frac{1}{2}\left[\overline{\mathbf{x}}_1^T\widehat{\mathbf{\Sigma}}^{-1}\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2^T\widehat{\mathbf{\Sigma}}^{-1}\overline{\mathbf{x}}_2\right] + \log_e\frac{n_1}{n} - \log_e\frac{n_2}{n} \end{aligned}$$

are ML estimates of $\mathbf{b}$ and b_0 respectively.

Classification rule The classification rule assigns $\mathbf{x}$ to Π_1 if $\widehat{L}(\mathbf{x}) > 0$, and assigns $\mathbf{x}$ to Π_2 otherwise.