Tut 6: LDA (Bayes' Classifier)

Jan 2024

The general mathematical formulation of a generative model:

$$\begin{aligned} h_D(\boldsymbol{x}) &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j)\mathbb{P}(Y=j) \\ &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ [\ln \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j) + \ln \mathbb{P}(Y=j)] \end{aligned} \tag{6.1}$$

QDA (only works for numeric inputs which follows the normal distribution):

$$\boxed{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j) \approx \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{C}_j|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{C}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\}.}$$

LDA (only works for numeric inputs which follows the normal distribution):

$$\boxed{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j) \approx \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{C}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\}.}$$

$$\Rightarrow \boxed{h_D(\boldsymbol{x}) = \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}} \left\{\ln \mathbb{P}(Y=j) + \vec{\mu}_j^T \boldsymbol{C}^{-1}\left[\boldsymbol{x} - \frac{1}{2}\vec{\mu}_j\right]\right\}.}$$

Naive Bayes:

$$\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j) \approx \prod_{i=1}^{p} \mathbb{P}(X_i = x_i|Y=j)$$

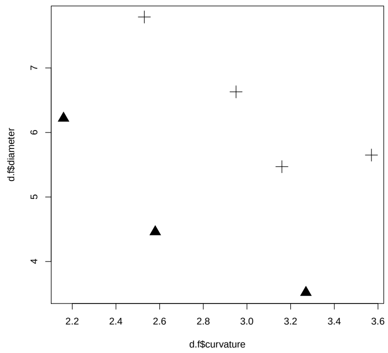
1. Factory XYZ produces very expensive and high quality golf balls that their qualities are measured in term of curvature and diameter. Result of quality control by experts is given in the table below:

Curvature	Diameter	Result
2.95	6.63	Passed
2.53	7.79	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	4.46	Not Passed
2.16	6.22	Not Passed
3.27	3.52	Not Passed

As a consultant to the factory, you get a task to set up the criteria for automatic quality control using LDA model. Then, the manager of the factory also wants to test your criteria upon a new type of golf ball which have curvature 2.81 and diameter 5.46.

(a) Plot the data with axes of curvature and diameter. Comment on the plot.

Solution. Both classes (Passed/Not Passed) are well-separated. Hence, LDA is suitable.



☐

(b) Write the data into matrix form by separating into "Passed" and "Not Passed".

Solution. Let 1 = Passed and 2 = Not Passed.

$$X_1 = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}$$

☐

(c) Compute the prior probability for both classes.

Solution. $\widehat{\pi}_1 = \frac{n_1}{n} = \frac{4}{7}, \quad \widehat{\pi}_2 = \frac{n_2}{n} = \frac{3}{7}$ ☐

(d) Compute the mean vectors for both classes.

Solution. The mean vectors are

$$\widehat{\mu}_1 = [\frac{2.95 + 2.53 + 3.57 + 3.16}{4}, \ \frac{6.63 + 7.79 + 5.65 + 5.47}{4}] = [3.0525, \ 6.3850]$$
$$\widehat{\mu}_2 = [\frac{2.58 + 2.16 + 3.27}{3}, \ \frac{4.46 + 6.22 + 3.52}{3}] = [2.67, \ 4.7333]$$

☐

(e) Compute the group covariance matrix.

Solution.

$$X_1 - \widehat{\mu}_1 = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix} - [3.0525, \ 6.3850] = \begin{bmatrix} -0.1025 & 0.2450 \\ -0.5225 & 1.4050 \\ 0.5175 & -0.7350 \\ 0.1075 & -0.9150 \end{bmatrix}$$

$$\Rightarrow C_1 = (X_1 - \widehat{\mu})^T (X_1 - \widehat{\mu}) = \begin{bmatrix} 0.562875 & -1.23795 \\ -1.237950 & 3.41150 \end{bmatrix}$$

$$X_2 - \widehat{\mu}_2 = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} - [2.67, \ 4.7333] = \begin{bmatrix} -0.0900 & -0.2733 \\ -0.5100 & 1.4867 \\ 0.6000 & -1.2133 \end{bmatrix}$$

$$C_2 = (X_2 - \widehat{\mu})^T (X_2 - \widehat{\mu}) = \begin{bmatrix} 0.6282 & -1.461600 \\ -1.4616 & 3.757067 \end{bmatrix}$$

The group covariance matrix

$$C = \frac{1}{7-2}(C_1 + C_2) = \begin{bmatrix} 0.238215 & -0.539910 \\ -0.539910 & 1.433713 \end{bmatrix}$$

□

(f) Write down the discriminant functions for both classes.

Solution. The discriminant functions are $\delta_j(X) = \ln(\pi_j) - \frac{1}{2}\mu_j C^{-1}\mu_j^T + \mu_j C^{-1}\boldsymbol{x}^T$, $j = 1, 2$:

$$\mu_1 C^{-1} = \begin{bmatrix} 3.0525 & 6.3850 \end{bmatrix} \begin{bmatrix} 0.238215 & -0.539910 \\ -0.539910 & 1.433713 \end{bmatrix}^{-1} = \begin{bmatrix} 156.38334 & 63.34455 \end{bmatrix}$$

$$\Rightarrow \delta_1(X) = \ln\frac{4}{7} - \frac{1}{2}\mu_1 C^{-1} \begin{bmatrix} 3.0525 \\ 6.3850 \end{bmatrix} + \mu_1 C^{-1}\boldsymbol{x}^T = \begin{bmatrix} 156.38334 & 63.34455 \end{bmatrix} \boldsymbol{x}^T - 441.4672$$

$$\mu_2 C^{-1} = \begin{bmatrix} 2.6700 & 4.7333 \end{bmatrix} \begin{bmatrix} 0.238215 & -0.539910 \\ -0.539910 & 1.433713 \end{bmatrix}^{-1} = \begin{bmatrix} 127.59686 & 51.35205 \end{bmatrix}$$

$$\delta_2(X) = \ln\frac{3}{7} - \frac{1}{2}\mu_2 C^{-1} \begin{bmatrix} 2.6700 \\ 4.7333 \end{bmatrix} + \mu_2 C^{-1}\boldsymbol{x}^T = \begin{bmatrix} 127.59686 & 51.35205 \end{bmatrix} \boldsymbol{x}^T - 292.7214$$

□

(g) Transform all the given data into discriminant functions.

Solution. For the first data, $\boldsymbol{x}_1 = [2.95,\ 6.63]$,

$\delta_1(\boldsymbol{x}_1) = \begin{bmatrix} 156.38334 & 63.34455 \end{bmatrix} \begin{bmatrix} 2.95 & 6.63 \end{bmatrix}^T - 441.4672 = 439.8380$

$\delta_2(\boldsymbol{x}_1) = \begin{bmatrix} 127.59686 & 51.35205 \end{bmatrix} \begin{bmatrix} 2.95 & 6.63 \end{bmatrix}^T - 292.7214 = 424.1534$, etc.

X_1	X_2	$\delta_1(X)$	$\delta_2(X)$	Class
2.95	6.63	439.8380	424.1534	1
2.53	7.79	447.6367	430.1311	1
3.57	5.65	474.7180	452.9385	1
3.16	5.47	399.1988	391.3804	1
2.58	4.46	244.5185	265.5086	2
2.16	6.22	290.3239	302.2976	2
3.27	3.52	292.8791	305.2795	2

□

(h) Locate the new golf ball in the plot as well as the functions to classify it.

Solution. $x_{new} = [2.81,\ 5.46]$

$$\delta_1(x_{new}) = \begin{bmatrix} 156.38334 & 63.34455 \end{bmatrix} \begin{bmatrix} 2.81 \\ 5.46 \end{bmatrix} - 441.4672 = 343.8312$$

$$\delta_2(x_{new}) = \begin{bmatrix} 127.59686 & 51.35205 \end{bmatrix} \begin{bmatrix} 2.81 \\ 5.46 \end{bmatrix} - 292.7214 = 346.208$$

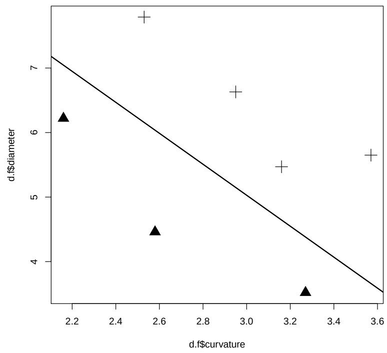
Hence, the new golf ball should be classified into class 2, i.e. "not passed". □

(i) Plot the discriminant line into plot of $\delta_1(X)$ versus $\delta_2(X)$.

Solution. The discriminant functions fail when $\delta_1(X) = \delta_2(X)$, i.e.

$$(156.38334 - 127.59686)x_1 + (63.34455 - 51.35205)x_2 - (441.4672 - 292.7214) = 0$$
$$\Rightarrow x_2 = 12.40324 - 2.4x_1$$

which can be used to plot a line using `abline(12.40324, -2.4)`.



□

2. (Final Assessment Jan 2021 Q5(a)) The data in Table 5.1 contains size measurements for two penguin species, i.e. Adelie and Gentoo, observed on three islands in the Palmer Archipelago, Antarctica.

Table 5.1: Palmer Archipelago Penguin data

Flipper length (mm)	Body mass (gram)	Species
196	4400	Adelie
188	3050	Adelie
219	5250	Gentoo
193	4200	Adelie
208	4200	Gentoo
215	5000	Gentoo
197	4775	Adelie

Construct a linear discriminant analysis (LDA) model for the given data in Table 5.1 by following the following steps.

(i) Write down the general mathematical formula of the LDA model. (2 marks)

Solution. The general mathematical formulation is [2 marks]

$$
\begin{aligned}
h_D(\boldsymbol{x}) &= \underset{j\in\{1,\ldots,K\}}{\operatorname{argmax}} \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{C}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\vec{\mu}_j)^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\vec{\mu}_j)\right\}\mathbb{P}(Y=j) \\
&= \underset{j\in\{1,\ldots,K\}}{\operatorname{argmax}} \left\{\ln\mathbb{P}(Y=j) + \vec{\mu}_j^T\boldsymbol{C}^{-1}\left[\boldsymbol{x}-\frac{1}{2}\vec{\mu}_j\right]\right\}.
\end{aligned}
$$

□

(ii) Write down the estimates of all the parameters of discriminant functions of the LDA model in part (i). (11 marks)

Solution. The prior probability estimates for both classes are

$$
\widehat{\mathbb{P}}(Y=\text{Adelie}) = \frac{4}{7}, \quad \widehat{\mathbb{P}}(Y=\text{Gentoo}) = \frac{3}{7}. \qquad [2 \text{ marks}]
$$

The mean vectors for both classes are

$$
\begin{aligned}
\vec{\mu}_{\text{Adelie}} &= \frac{1}{4}((196,4400)+(188,3050)+(193,4200)+(197,4775)) \\
&= (193.50,\ 4106.25) \\
\vec{\mu}_{\text{Gentoo}} &= \frac{1}{3}((219,5250)+(208,4200)+(215,5000)) \\
&= (214,\ 4816.667)
\end{aligned}
\qquad [3 \text{ marks}]
$$

We now estimate the "unscaled" covariance matrix estimate for Adelie:

$$X_{\text{Adelie}} - \vec{\mu}_{\text{Adelie}} = \begin{bmatrix} 196 & 4400 \\ 188 & 3050 \\ 193 & 4200 \\ 197 & 4775 \end{bmatrix} - [193.50,\ 4106.25] = \begin{bmatrix} 2.5 & 293.5 \\ -5.5 & -1056.5 \\ -0.5 & 93.5 \\ 3.5 & 668.5 \end{bmatrix}$$

$$C_{\text{Adelie}} = (X_{\text{Adelie}} - \vec{\mu}_{\text{Adelie}})^T (X_{\text{Adelie}} - \vec{\mu}_{\text{Adelie}}) = \begin{bmatrix} 49.0 & 8837.5 \\ 8837.5 & 1657969.0 \end{bmatrix}$$

[1.5 marks]

We now estimate the "unscaled" covariance matrix estimate for Gentoo:

$$X_{\text{Gentoo}} - \vec{\mu}_{\text{Gentoo}} = \begin{bmatrix} 219 & 5250 \\ 208 & 4200 \\ 215 & 5000 \end{bmatrix} - [214,\ 4816.667] = \begin{bmatrix} 5 & 433.333 \\ -6 & -616.667 \\ 1 & 183.333 \end{bmatrix}$$

$$C_{\text{Gentoo}} = (X_{\text{Gentoo}} - \vec{\mu}_{\text{Gentoo}})^T (X_{\text{Gentoo}} - \vec{\mu}_{\text{Gentoo}}) = \begin{bmatrix} 62 & 6050.0 \\ 6050 & 601666.7 \end{bmatrix}$$

[1.5 marks]

The group covariance matrix estimate is

$$\boldsymbol{C} = \frac{1}{7-2}(C_{\text{Adelie}} + C_{\text{Gentoo}}) = \begin{bmatrix} 22.2 & 2977.5 \\ 2977.5 & 451927.1 \end{bmatrix} \qquad \text{[1 mark]}$$

Note that using scientific calculator or linear algebra, we can obtain

$$\mu_{\text{Adelie}}^T \boldsymbol{C}^{-1} = \begin{bmatrix} 64.4419433 \\ -0.4154865 \end{bmatrix}, \quad \mu_{\text{Gentoo}}^T \boldsymbol{C}^{-1} = \begin{bmatrix} 70.566664 \\ -0.454267 \end{bmatrix}. \qquad \text{[1 mark]}$$

The discriminant functions (of the LDA model) for both classes are .... [1 mark]

$$\delta_{\text{Adelie}}(\boldsymbol{x}) = \ln\frac{4}{7} + \begin{bmatrix} 64.4419433 \\ -0.4154865 \end{bmatrix} (\boldsymbol{x} - \begin{bmatrix} 96.750 \\ 2053.125 \end{bmatrix}),$$

$$\delta_{\text{Gentoo}}(\boldsymbol{x}) = \ln\frac{3}{7} + \begin{bmatrix} 70.566664 \\ -0.454267 \end{bmatrix} (\boldsymbol{x} - \begin{bmatrix} 107 \\ 2408.334 \end{bmatrix})$$

□

(iii) Use the discriminant functions of the LDA model in part (ii) to determine the species of a penguin with a flipper length of 218 mm and a body mass of 4590 gram. (4 marks)

Solution.

$$\delta_{\text{Adelie}}(\begin{bmatrix} 218 \\ 4590 \end{bmatrix}) = \ln\frac{4}{7} + \begin{bmatrix} 64.4419433 \\ -0.4154865 \end{bmatrix} (\begin{bmatrix} 218 \\ 4590 \end{bmatrix} - \begin{bmatrix} 96.750 \\ 2053.125 \end{bmatrix}),$$
$$= -0.5596 - (-6759.548) = 6758.989$$

$$\delta_{\text{Gentoo}}(\begin{bmatrix} 218 \\ 4590 \end{bmatrix}) = \ln\frac{3}{7} + \begin{bmatrix} 70.566664 \\ -0.454267 \end{bmatrix} (\begin{bmatrix} 218 \\ 4590 \end{bmatrix} - \begin{bmatrix} 107 \\ 2408.334 \end{bmatrix})$$
$$= -0.8473 - (-6841.841) = 6840.994$$

[1.5+1.5=3 marks]

Since $6840.994 > 6758.989$, the LDA model predicts the species to be Gentoo. [1 mark]

□

(iv) (Not part of the final) Try to write down the QDA model.

Solution. Note that there are two inputs Flipper length and Body mass, therefore, $p = 2$ and $p/2 = 1$.

Part of the QDA model associated with Adelie is

$$\mathbb{P}(Y = \text{Adelie}|X_1 = x_1, X_2 = x_2) \propto P(Y = \text{Adelie})\times$$
$$\frac{1}{(2\pi)\sqrt{|C_{\text{Adelie}}|}} \exp\left\{-\frac{1}{2}\begin{bmatrix} x_1 - 193.50 \\ x_2 - 4106.25 \end{bmatrix} C_{\text{Adelie}}^{-1} \begin{bmatrix} x_1 - 193.50 & x_2 - 4106.25 \end{bmatrix}\right\}.$$

Try to write down the QDA model associated with Gentoo and then the complete model.

□