Homework 5

1 Directions:

- **Due: Thursday March 24, 2022 at 9pm.** Late submissions will be accepted for 24 hours with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

Yuichi Hamamoto

2 Problems

Problem 1. [20 points] Suppose you are predicting a feature Y that can take on three values $Y \in \{+1, +2, +3\}$ and you can predict Y using two features X_1 and X_2. You decide to try LDA (i.e. we will estimate a different mean for each class but esimate a common covariance matrix). Suppose that the (common) covariance matrix you estimate is

$$\Sigma_{+1} = \Sigma_{+2} = \Sigma_{+3} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

the priors are equal

$$\pi_{+1} = \pi_{+2} = \pi_{+3} = \frac{1}{3},$$

and the mean vectors are

$$\mu_{+1} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mu_{+2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mu_{+3} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

a) Find the equation for the LDA boundary between Y = +1 and Y = +2.

$$0 = \log\left(\frac{P(x|y = +1)\,\pi_{+1}}{P(x|y = \ +2)\pi_{+2}}\right)$$

$$= \log(P(x|y = +1)) - \log(P(x|y = \ +2)) + log1$$

$$= -\mu_{+1}\Sigma^{-1}X + \mu_{+2}\Sigma^{-1}X - 0.5(\mu_{+1}\Sigma^{-1}\mu_{+1} - \mu_{+2}\Sigma^{-1}\mu_{+2})$$

$$\because (\mu_{+1}\Sigma^{-1}\mu_{+1} - \mu_{+2}\Sigma^{-1}\mu_{+2}) = 0$$

$$= -[-1\ -1]\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}(X_1, X_2) + [1\ 1]\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}(X_1, X_2)$$

$$= X_1 + X_2$$

$$\therefore \mathrm{X}_2 = -\mathrm{X}_1$$

$$(\mathrm{X}_2 > -\mathrm{X}_1 : Y = \ +1, \mathrm{X}_2 < -\mathrm{X}_1 : Y = \ +2)$$

b) Find the equation for the LDA boundary between Y = +1 and Y = +3.

$$0 = \log\left(\frac{P(x|y = +1)\, \pi_{+1}}{P(x|y = \ +3)\pi_{+3}}\right)$$

$$= \log\big(P(x|y = +1)\big) - \log\big(P(x|y = \ +3)\big) + log1$$

$$= -\mu_{+1}\Sigma^{-1}X + \mu_{+3}\Sigma^{-1}X - 0.5(\mu_{+1}\Sigma^{-1}\mu_{+1} - \mu_{+3}\Sigma^{-1}\mu_{+3})$$

$$\because (\mu_{+1}\Sigma^{-1}\mu_{+1} - \mu_{+3}\Sigma^{-1}\mu_{+3}) = 0$$

$$= -[-1\ -1]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}(X_1, X_2) + [-1\ 1]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}(X_1, X_2)$$

$$= X_2$$

$$(\mathrm{X}_2 > 0 : Y = \ +1,\ \mathrm{X}_2 < 0 : Y = \ +3)$$

c) Find the equation for the LDA boundary between Y = +2 and Y = +3.

$$0 = \log\left(\frac{P(x|y = +2)\, \pi_{+2}}{P(x|y = \ +3)\pi_{+3}}\right)$$

$$= \log\big(P(x|y = +2)\big) - \log\big(P(x|y = \ +3)\big) + log1$$

$$= -\mu_{+2}\Sigma^{-1}X + \mu_{+3}\Sigma^{-1}X - 0.5(\mu_{+2}\Sigma^{-1}\mu_{+2} - \mu_{+3}\Sigma^{-1}\mu_{+3})$$

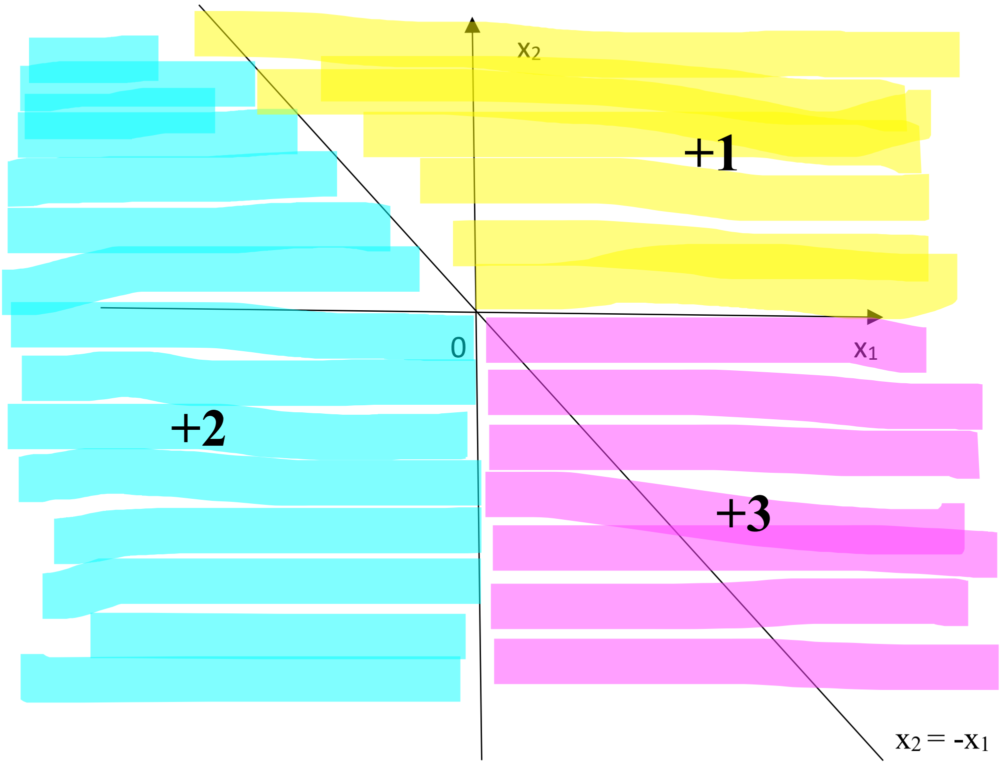
$$\because (\mu_{+2}\Sigma^{-1}\mu_{+2} - \mu_{+3}\Sigma^{-1}\mu_{+3}) = 0$$

$$= -[1\ 1]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}(X_1, X_2) + [-1\ 1]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}(X_1, X_2)$$

$$= X_1$$

$$(\mathrm{X}_1 > 0 : Y = \ +3,\ \mathrm{X}_1 < 0 : Y = \ +2)$$

d) Make a plot with x_1 along horizontal axis, x_2 along vertical axis and draw each of the LDA boundaries you found. For each region, write the class label (e.g. "+2") that would be chosen for a new sample that appeared in that region.



e) If you wanted to use Naive Bayes in addition to LDA (thus still assuming Gaussianity) for this problem, explain what would change (if anything).

In case the features are not independent, then the result could be bad.

Problem 2. [15 points]

You intern for a university's athletics program and are tasked with predicting whether their rugby team will make it to the playoffs ($Y \in \{\text{yes}, \text{no}\}$) based on their score in the first game of the season, X.

You collect data and estimate that for the years that the rugby team made it to the playoffs, their score in the first game had a mean of $\widehat{\mu}_{\text{yes}} \approx 30$ and variance $\widehat{\sigma}^2_{\text{yes}} \approx 50$. For the years that they did not make it to the playoffs, their score in the first game had a mean of $\widehat{\mu}_{\text{no}} \approx 15$ and variance $\widehat{\sigma}^2_{\text{no}} \approx 80$. The team made it to the playoffs 30% of the years.

This year, they scored 20 points in their first game. Predict the probability that the team will make it to the playoffs. Use Bayes' theorem, and model the distribution of the first game scores, conditioned on whether or not they make it to the playoffs that year, as Gaussian.

P(The team makes it to the playoffs) = 0.3
P(The team does not make it to the playoffs) = 0.7
Distribution of whether the rugby team makes it to the playoffs: X~N(30, $\sqrt{50}$)
Distribution of whether the rugby team does not make it to the playoffs: X~N(15, $\sqrt{80}$)

∴P(Score 20 points | Finish in playoffs)
= X~N(30,$\sqrt{50}$)
$= \frac{1}{\sqrt{2\pi}\cdot 7.0711} \cdot e^{-\frac{1}{2}\left(\frac{20-30}{7.0711}\right)^2}$

= 0.020755371

Similarly,
P(Score 20 points | Not finish in playoffs) = X~N(15,$\sqrt{80}$)
$= \frac{1}{\sqrt{2\pi}\cdot 8.94427191} \cdot e^{-\frac{1}{2}\left(\frac{20-15}{8.94427191}\right)^2}$

= 0.038151055

∵bayes theorem
Aside: P(Scores 20 points)
= 0.020755371 * 0.3 + 0.038151055 * 0.7
= 0.03293235

P(The team makes it to the playoffs | Scores 20 points)
= (P(Scores 20 points The team makes it to the playoffs)*P(The team makes it to the playoffs)) / P(Scores 20 points)
(0.020755371 * 0.3) / 0.03293235 = 0.18907279

∴The probability is 18.9%