DSC 140A - Homework 08

Due: Wednesday, March 5

Instructions: Write your solutions to the following problems either by typing them or handwriting them on another piece of paper or on an iPad/tablet. Show your work or provide justification unless otherwise noted; submissions that don't show work might lose credit. If you write code to solve a problem, include the code by copy/pasting or as a screenshot. You may use `numpy`, `pandas`, `matplotlib` (or another plotting library), and any standard library module, but no other third-party libraries unless specified. Submit homeworks via Gradescope by 11:59 PM.

A LaTeX template is provided at `http://dsc140a.com`, next to where you found this homework. Using it is totally optional, but encouraged if you plan to go to grad school. See this video for a quick introduction to LaTeX.

Problem 1.

The file at the link below contains a data set of 100 points from two classes (1 and -1).

`https://f000.backblazeb2.com/file/jeldridge-data/003-two_clusters/data.csv`

The first two columns contains features, and the last column contains the label of the point. Note that the labels are 1 and -1, not 1 and 0, and that there are no column headers.

In all parts of this problem you may use code to compute your answers. If you do, be sure to include your code.

a) Suppose two Gaussians with full covariance matrices are used to model the densities $p_X(x \mid Y = 1)$ and $p_X(x \mid Y = -1)$. What are the maximum likelihood estimates for the covariance matrices of each Gaussian?

(Allow each Gaussian to have its own covariance matrix; don't use the same covariance for both.)

Hint: the covariance matrix for the Gaussian fit to points from class 1 should have 12.29 in its top-left entry.

Solution: If you'd like to compute the covariances manually:

```
data = np.loadtxt('data.csv', delimiter=',')
X = data[:,:2]
y = data[:,-1]

X_1 = X[y == 1]
X_0 = X[y == -1]

y_1 = y[y == 1]
y_0 = y[y == -1] * 0

mu_1 = X_1.mean(axis=0)
Z_1 = (X_1 - mu_1)
n_1 = len(Z_1)

mu_0 = X_0.mean(axis=0)
Z_0 = (X_0 - mu_0)
```

```
n_0 = len(Z_0)

C_1 = 1 / n_1 * Z_1.T @ Z_1
C_0 = 1 / n_0 * Z_0.T @ Z_0

>>> C_1
array([[12.29584016,  0.28098224],
       [ 0.28098224, 16.06766736]])
>>> C_0
array([[10.91736224,  0.53015728],
       [ 0.53015728, 15.17320916]])
```

You could also have used numpy. If you did, you had to make sure to get to pass it an array whose *columns* are data points, instead of rows, and to use `bias = True`:

```
np.cov(X_1.T, bias=True)
```

b) Using the estimated Gaussians with the Bayes classification rule, what are the predicted labels of each of the following points?

- $(0, 0)^T$
- $(1, 1)^T$
- $(10, 5)^T$
- $(5, -5)^T$
- $(8, 5)^T$

Show your calculations.

Note: making predictions in this way (using Gaussians with unequal covariance matrices) is known as *Quadratic Discriminant Analysis*.

Solution: The predicted labels are -1, -1, 1, -1, -1.

The code below defines a function for making the predictions:

```
mvn = scipy.stats.multivariate_normal

py_1 = n_1 / (n_1 + n_0)
py_0 = n_0 / (n_1 + n_0)

def predict(x):
    if mvn.pdf(x, mean=mu_1, cov=C_1) * py_1 > mvn.pdf(x, mean=mu_0, cov=C_0) * py_0:
        return 1
    else:
        return -1
```

Problem 2.

In lecture, we derived Linear Discriminant Analysis (LDA) by starting with the Bayes classifier and modeling each class-conditional density as a multivariate Gaussian and using the same covariance matrix for each. We stated, but did not prove, that the decision boundary of an LDA classifier is linear.

Recall that, for a binary classifier based on the Bayes Classifier, the decision boundary is the set of all points $\vec{x}$ where

$$\hat{p}(\vec{x} \mid Y = 1)\hat{\mathbb{P}}(Y = 1) = \hat{p}(\vec{x} \mid Y = 0)\hat{\mathbb{P}}(Y = 0),$$

where the various $\hat{p}$ and $\hat{P}$ are estimated densities and probabilities.

Using this fact, prove that the decision boundary of an LDA classifier is linear. For simplicity, you may assume that $\vec{x} \in \mathbb{R}^2$ and that the shared covariance matrix is diagonal (although the result holds even if the covariance matrix is not diagonal).

Hint: since you may assume that $\vec{x} = (x_1, x_2)^T$, you can start from the above equality and solve for x_2 in terms of x_1, showing that you get the equation of a line.

Solution: Let C be the shared covariance matrix. The decision boundary consists of all $\vec{x} = (x_1, x_2)^T$ for which

$$\hat{p}(\vec{x}\,|\,Y = 1) \cdot \hat{\mathbb{P}}(Y = 1) = \hat{p}(\vec{x}\,|\,Y = 0)\dot{c}\hat{\mathbb{P}}(Y = 0).$$

Since this is an LDA classifier, $p(\vec{x}\,|\,Y = 0)$ and $p(\vec{x}\,|\,Y = 1)$ are both multivariate Gaussian densities with the same covariance matrix C. Let $\vec{\mu}_0$ and $\vec{\mu}_1$ be the means of the two classes. Then, substituting the Gaussian density function, the decision boundary consists of all $\vec{x}$ for which

$$\frac{1}{(2\pi|C|)^{d/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_0)^T C^{-1}(\vec{x}-\vec{\mu}_0)} \cdot \hat{\mathbb{P}}(Y = 0) = \frac{1}{(2\pi|C|)^{d/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T C^{-1}(\vec{x}-\vec{\mu}_1)} \cdot \hat{\mathbb{P}}(Y = 1).$$

We notice that the same normalization term appears on both sides of the equation, and so we can cancel it:

$$e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_0)^T C^{-1}(\vec{x}-\vec{\mu}_0)} \cdot \hat{\mathbb{P}}(Y = 0) = e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T C^{-1}(\vec{x}-\vec{\mu}_1)} \cdot \hat{\mathbb{P}}(Y = 1).$$

We can also take the natural logarithm of both sides to simplify, giving:

$$-\frac{1}{2}(\vec{x} - \vec{\mu}_0)^T C^{-1}(\vec{x} - \vec{\mu}_0) + \log \hat{\mathbb{P}}(Y = 0) = -\frac{1}{2}(\vec{x} - \vec{\mu}_1)^T C^{-1}(\vec{x} - \vec{\mu}_1) + \log \hat{\mathbb{P}}(Y = 1). \tag{1}$$

Here is where it might be convenient to write $\vec{x} = (x_1, x_2)^T$ and expand. Additionally, write $\vec{\mu}_0 = (\mu_{01}, \mu_{02})^T$ and $\vec{\mu}_1 = (\mu_{11}, \mu_{12})^T$, and let

$$C = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Since C is diagonal, its inverse is simply:

$$C^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Using these variables, we can expand:

$$\begin{aligned}
(\vec{x} - \vec{\mu}_0)^T C^{-1}(\vec{x} - \vec{\mu}_0) &= \begin{pmatrix} x_1 - \mu_{01} & x_2 - \mu_{02} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} x_1 - \mu_{01} \\ x_2 - \mu_{02} \end{pmatrix} \\
&= \begin{pmatrix} x_1 - \mu_{01} & x_2 - \mu_{02} \end{pmatrix} \begin{pmatrix} \frac{x_1 - \mu_{01}}{\sigma_1^2} \\ \frac{x_2 - \mu_{02}}{\sigma_2^2} \end{pmatrix} \\
&= \frac{(x_1 - \mu_{01})^2}{\sigma_1^2} + \frac{(x_2 - \mu_{02})^2}{\sigma_2^2}.
\end{aligned}$$

Similarly, carrying out the same expansion for the other term, we will get:

$$(\vec{x} - \vec{\mu}_1)^T C^{-1}(\vec{x} - \vec{\mu}_1) = \frac{(x_1 - \mu_{11})^2}{\sigma_1^2} + \frac{(x_2 - \mu_{12})^2}{\sigma_2^2}.$$

Plugging these expansions back into Equation (1), we get:

$$-\frac{1}{2}\left(\frac{(x_1-\mu_{01})^2}{\sigma_1^2}+\frac{(x_2-\mu_{02})^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=0)=-\frac{1}{2}\left(\frac{(x_1-\mu_{11})^2}{\sigma_1^2}+\frac{(x_2-\mu_{12})^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=1).$$

Since we wish to solve for x_2 in terms of x_1, we expand the squares and simplify:

$$\begin{aligned}&-\frac{1}{2}\left(\frac{x_1^2-2\mu_{01}x_1+\mu_{01}^2}{\sigma_1^2}+\frac{x_2^2-2\mu_{02}x_2+\mu_{02}^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=0)\\&\quad=-\frac{1}{2}\left(\frac{x_1^2-2\mu_{11}x_1+\mu_{11}^2}{\sigma_1^2}+\frac{x_2^2-2\mu_{12}x_2+\mu_{12}^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=1).\end{aligned}$$

We recognize some terms that appear on both sides of the equation, namely x_1^2/σ_1^2 and x_2^2/σ_2^2. We can cancel these terms:

$$\begin{aligned}&-\frac{1}{2}\left(-\frac{2\mu_{01}x_1}{\sigma_1^2}+\frac{\mu_{01}^2}{\sigma_1^2}-\frac{2\mu_{02}x_2}{\sigma_2^2}+\frac{\mu_{02}^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=0)\\&\quad=-\frac{1}{2}\left(-\frac{2\mu_{11}x_1}{\sigma_1^2}+\frac{\mu_{11}^2}{\sigma_1^2}-\frac{2\mu_{12}x_2}{\sigma_2^2}+\frac{\mu_{12}^2}{\sigma_2^2}\right)+\log\hat{\mathbb{P}}(Y=1).\end{aligned}$$

We now try to solve for x_2 in terms of x_1. We start by moving all of the terms involving x_2 to the left-hand side of the equation, and all of the other terms to the right-hand side:

$$\left(-\frac{\mu_{02}}{\sigma_2^2}+\frac{\mu_{12}}{\sigma_2^2}\right)x_2=\left(-\frac{\mu_{01}}{\sigma_1^2}+\frac{\mu_{11}}{\sigma_1^2}\right)x_1+\frac{\mu_{01}^2}{2\sigma_1^2}-\frac{\mu_{11}^2}{2\sigma_1^2}+\frac{\mu_{02}^2}{2\sigma_2^2}-\frac{\mu_{12}^2}{2\sigma_2^2}-\log\hat{\mathbb{P}}(Y=0)+\log\hat{\mathbb{P}}(Y=1).$$

This is a linear function of x_1 and x_2, and so the decision boundary is linear. We can see this a little more clearly if we give names to the coefficients of x_1 and x_2 and the constants:

$$\underbrace{\left(-\frac{\mu_{02}}{\sigma_2^2}+\frac{\mu_{12}}{\sigma_2^2}\right)}_{\alpha}x_2=\underbrace{\left(-\frac{\mu_{01}}{\sigma_1^2}+\frac{\mu_{11}}{\sigma_1^2}\right)}_{\beta}x_1+\underbrace{\frac{\mu_{01}^2}{2\sigma_1^2}-\frac{\mu_{11}^2}{2\sigma_1^2}+\frac{\mu_{02}^2}{2\sigma_2^2}-\frac{\mu_{12}^2}{2\sigma_2^2}-\log\hat{\mathbb{P}}(Y=0)+\log\hat{\mathbb{P}}(Y=1)}_{\gamma}.$$

Therefore: $\alpha x_2=\beta x_1+\gamma$, or: $x_2=\frac{\beta}{\alpha}x_1+\frac{\gamma}{\alpha}$, which is the equation of a line.

Problem 3.

You've been hired by a generic online retailer named after a rainforest named after a river. Your job is to build a model to predict whether or not a particular item will sell. You are provided with a dataset of outcomes for a collection of products:

Brand	Price Range	Condition	Sold
A	High	Used	No
A	High	New	Yes
B	Low	New	Yes
C	Medium	New	Yes
B	Low	Used	No
A	High	New	No
C	High	Used	Yes
A	Medium	Used	Yes
B	Medium	Used	No
C	Low	New	No
B	Low	Used	Yes

Using a Naïve Bayes classifier and the data above, predict if a product with Brand = B, Price Range = Medium, Condition = Used will sell or not. Show your calculations.

Solution:

We start by calculating the class conditional probabilities:

$$P(\text{Brand} = \text{B} \,|\, \text{Sold} = \text{Yes}) = \frac{2}{6}$$

$$P(\text{Brand} = \text{B} \,|\, \text{Sold} = \text{No}) = \frac{2}{5}$$

$$P(\text{Price Range} = \text{Medium} \,|\, \text{Sold} = \text{Yes}) = \frac{2}{6}$$

$$P(\text{Price Range} = \text{Medium} \,|\, \text{Sold} = \text{No}) = \frac{1}{5}$$

$$P(\text{Condition} = \text{Used} \,|\, \text{Sold} = \text{Yes}) = \frac{3}{6}$$

$$P(\text{Condition} = \text{Used} \,|\, \text{Sold} = \text{No}) = \frac{3}{5}$$

The prior probabilities are $P(\text{Sold} = \text{Yes}) = 6/11$ and $P(\text{Sold} = \text{No}) = 5/11$. Therefore:

$$
\begin{aligned}
&P(\text{Sold} = \text{Yes} \,|\, \text{Brand} = \text{B}, \text{Price Range} = \text{Medium}, \text{Condition} = \text{Used}) \\
&\qquad \propto \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{3}{6} \cdot \frac{6}{11} \\
&= \frac{72}{2376} \approx 0.03
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{Sold} = \text{No} \,|\, \text{Brand} = \text{B}, \text{Price Range} = \text{Medium}, \text{Condition} = \text{Used}) \\
&\qquad \propto \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{11} \\
&= \frac{30}{1375} \approx 0.021
\end{aligned}
$$

Because the former is larger, we predict that the product will be **sold**.