# DSC 140A - Homework

Due:

**Instructions:** Write your solutions to the following problems either by typing them or handwriting them on another piece of paper. Show your work or provide justification unless otherwise noted. 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 .

**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.

**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*.

**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.

**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.