Lab 7 *(due: Mar 14)*
MACHINE LEARNING - COSC 4360

Department of Computer Science and Electrical Engineering

Spring 2025

Exercises

*Create a **New Project** for every exercise. Take a screenshot of the source code along with its output and place the **source code** and the **screenshot** in a **zipped folder** named **LastNameFirstName_Lab7***

Exercise 1

Given the following dataset: *wdbc.data.csv*, ignore the first column (ID number), assign the second column to **y** and the rest of the columns to **X**. **Scale** your data using the **Standardization** method and perform **Principal Component Analysis** with **PCA=2**. **Plot** the **2** *Principal Components* and print the **variance ratio**. In addition, given the data point: dataPoint = np.array([7.76, 24.54, 47.92, 181, 0.05263, 0.04362, 0, 0, 0.1587, 0.05884, 0.3857, 1.428, 2.548, 19.15, 0.007189, 0.00466, 0, 0, 0.02676, 0.002783, 9.456, 30.37, 59.16, 268.6, 0.08996, 0.06444, 0, 0, 0.2871, 0.07039]), **plot** it on the existing plot, as shown in Fig. 1 below. In addition, using **logistic regression**, plot the **decision boundary** and **predict** the class to which the data point belongs. Use two different colors in the plot. You may use any *built-in* functions you wish.

Note: You will have to solve for x_2, as shown in eqn. 1 below, to plot the decision boundary.

$$w_0 + w_1x_1 + w_2x_2 = 0 \Longrightarrow x_2 = -\frac{w_0 + w_1x_1}{w_2} \tag{1}$$

Exercise 2

Given the following dataset: *golf.csv*, use **Naïve Bayes** classifier to **predict** whether a golf game will be played or not given the following three data points: [Rainy, Hot, High, True], [Sunny, Mild, Normal, False], [Sunny, Cool, High, False].

Note: Use *LabelEncoder()* to convert the three data points into numerals.

Exercise 3 *(Optional)*

Given the same dataset, compute the **probabilities** yourself using only one **predictor** variable, e.g., *Outlook*. Your algorithm should be able to compute the probabilities for any **predictor** variable without any modifications. Compare your results with the output of the *built-in* function *predict_proba()*.

Note: Submit through **Canvas**

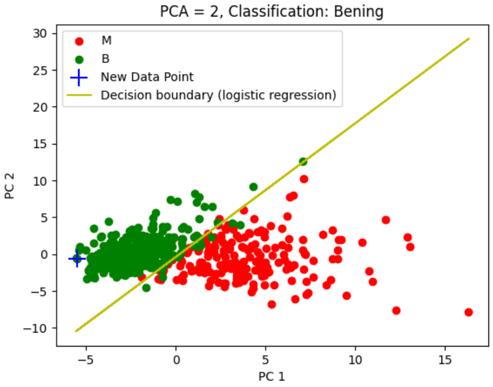

Figure 1: PCA plot with logistic regression decision boundary