Elucidata
www.elucidata.io

Assignment

Guidelines -

1. The solution to this assignment is expected to be submitted in the form of a python notebook or R .rmd file. [like Colab notebook]
2. Do not use absolute paths when loading data files in the notebook.
3. Make good use of markdowns to write proper section headers and explanations for the code and the results you would generate.
4. Your notebook should be readable and results reproducible.
5. Though part (III) is optional, we would like you to give it an attempt, especially if you have a background in ML modeling.

Data description -

The attached zipped file comprises two files – data.txt and features.txt representing the breast cancer data set. The file ‘data.txt’ has 32 columns, first one of which is the identifier, second column is the diagnosis (benign/malignant) and the rest of the 30 columns represent the features derived from the fine needle image of the breast mass. They describe the characteristics of the cell nuclei present in the image. Names of these features are provided in the ‘features.txt’ file for your reference.

PART - I

1. There are a few missing values in the data. Identify what fraction of observations has missing values and devise a strategy for the missing value imputation.

2. Are the scales of the features in the similar range? If not, normalise the data for downstream analysis. You could use either standard scaling (z-score transformation) or min-max scaling.

3. Create a heatmap of the normalised data with features in rows and observations in columns. Reorder the columns such that the samples from the same class are grouped together. What do you infer from the heatmap?

4. Generate a suitable visualization which can help examine the differences in the distribution of the features between the malignant and the benign tumors.
   [Hints] – Boxplot, Violinplot, Histograms

5. Can you pick up the top 5 features which are best discriminating between the two classes visually?

PART - II

6. Perform the dimensionality reduction on this dataset with PCA.
   a. Generate PCA scores plot with PC1 on x-axis and PC2 on y-axis and label the observations with diagnosis class.
   b. What do you infer from the PCA plot? Explain your findings.

PART – III

7. Now that we have identified the top features from the previous exercise, let’s build a very simple model using only the top variable to predict the diagnosis class. Essentially, given the value of the feature (x), we want to predict the diagnosis class y {Benign(B)/Malignant(M)}. The model can take the following form -

   y = {B if x <= x_threshold or M if x > x_threshold}

   a. Vary the x_threshold and assess how the model’s accuracy changes
   b. What’s a good threshold that minimizes both type I and type II errors
   c. Generate the precision-recall curve for this model
   d. What would you prefer – a model with high precision or with high recall? Justify your answer.