CITS5508 Machine Learning
Semester 1, 2024
Lab sheet 7
(Not assessed)

You will develop Python code to perform classification tasks using principal components analysis (PCA). Certify that the presentation of your Python notebook is good and that you used the Markdown cells well. Make sure you properly format your plots and results. For instance, all your diagrams/plots should have proper axis labels and titles to help the reader understand what you are plotting. Another example is the confusion matrix; not showing the class names makes the confusion matrix completely useless. Use the lab sheets to learn how to improve the presentation of your notebook, as you will need this in the assessments.

A model for diagnosing cancer

Determining whether a tumour is malignant or benign is one of the challenging aspects when treating cancer. Machine learning techniques can help identify cancer types by extracting the differences in cell nucleus features. In this lab sheet, you will extend the analysis on the Breast Cancer Wisconsin (diagnostic)[1] dataset. The `breast-cancer.csv` data set provided contains:

1. Patient ID number
2. Diagnosis (M = malignant, B = benign)
3. Ten cell nucleus features, namely:
   - radius (mean of distances from centre to points on the perimeter)
   - texture (standard deviation of grayscale values)
   - perimeter
   - area
   - smoothness (local variation in radius lengths)
   - compactness (perimeter^2 / area - 1.0)
   - concavity (severity of concave portions of the contour)
   - concave points (number of concave portions of the contour)
   - symmetry
   - fractal dimension ('coastline approximation' - 1)

Tasks

1. Build a logistic regression and a decision tree model to predict the tumour status. The presentation should include a comparison of the two models and a recommendation regarding which would be more appropriate in a clinical setting. Note: You should use the fundamental steps of a machine learning project (e.g. hyperparameters fine-tuning, cross-validation, etc.).
2. Describe the features that have a higher chance of impacting the prediction of the tumour status according to each of the two models. Discuss their similarities/differences.

[1] More information about the data set can be found at https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data and https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

3. Using PCA, present the scatter plot of the data on the first two principal components. Add to your scatter plot different colours to represent the two classes in the data. What proportion of data variance is explained using the first two principal components?

4. Considering the first two principal components from Task 3, present the biplot with the variables vectors and the observed data projected on the first two principal components (with the colours for the two categories). Give your interpretation of the results.

5. Using the plot of Task 4, which variables are more related to the tumour status? Justify your answer. Compare the results obtained with the results obtained in Task 2.

6. Using PCA, determine the number of components to retain 95% of the explained variance. Use as new features the resulting principal components scores and repeat task 1 on these new features. You can choose one of the models (logistic regression or the decision tree). What is the dimension of the new (projected) data set? Comment on the performance resulting from using the original and principal components features.