AUA, Machine Learning
Midterm II

Check your AUA ID = ******XY and take the last 2 numbers. Use them as the value of random state parameter in the functions that simulate data. Before starting your Jupyter Notebook write your name, surname and AUA ID.

Problem I. Binary Classification Problem – Logistic Regression (score = 30)

Dataset is:

$$X_1, y_1 = make\_classification(n\_samples = 4000, n\_features = 15, n\_informative = 2, n\_clusters\_per\_class = 1, random\_state = XY, class\_sep = 1, flip\_y = 0.05, n\_classes = 2)$$

Use **train_test_split** function and split the dataset into 80% - 20% portions.

1. **(score=15)** Apply **Logistic Regression** to the training set and find the optimal value of parameter C via 10-fold cross validation while maximizing the Accuracy. Draw the average cv-score (average Accuracies) across the values of C. Show the optimal value of C and the corresponding Accuracy.
2. **(score=15)** Apply the trained model with the optimal parameter C to the test set. Perform comparison of the results for both training and test sets. Comment on the ROC curves, AUCs, and classification reports.

Problem II. Multiclass Classification Problem - LDA, QDA (score = 30)

Dataset is:

$$X_2, y_2 = make\_blobs(n\_samples = [5000,300,150], n\_features = 10, random\_state = XY, cluster\_std = [6,3,5])$$

Apply **LDA** and **QDA** to the entire dataset:

1. **(score 15)** Show the ROC curves for both models on the same plot. Calculate the AUCs. Show and compare the accuracies for both models. Which model performs better? Explain why?
2. **(score 15)** Show the PR curves for both models and for each class. Which class is predicted better and by which model? Do you see connection with the class imbalance problem?

Problem III. Binary Classification - kNN (score = 40)

Dataset is the same as for Problem I. Use the same (exactly the same) training and test sets.

1. **(score=10)** Apply **k-NN** to the training set and find optimal **k** via 10-fold cv while maximizing the Accuracy. Draw the average cv-scores across different **k**. Show the optimal **k** and the corresponding accuracy.
2. **(score = 10)** Apply the trained model with optimal **k** to the test set. Compare the results obtained both for training and test sets.
3. **(score = 20)** Perform comparison (**for test data**) with the optimal model derived in Problem I. Comment on the ROC curves plotted together and PR curves plotted together for each class.