AUA, Machine Learning
Midterm III

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. We recommend to use the same random_state value in all functions across your notebook when such parameter exists.

Problem I. Regression by Tree Based Methods (score = 50)

Dataset is:

$$X_1, y_1 = make\_regression(n_{samples} = 1000, n_{features} = 7, n_{informative} = 3, random_{state} = XY, noise = 10)$$

Use **train_test_split** function and split the dataset into 60% - 40% portions, respectively.

1. **(score = 15)** Apply DecisionTreeRegressor() to the **train data**. Tune parameter "max_leaf_nodes" via 3-fold cross validation to maximize the "score". Show the optimal value of the parameter and the "score" on **train data**. Denote the optimal model as "Reg_A".
2. **(score = 15)** Apply RandomForestRegressor() to the **train data**. Simultaneously tune parameters "n_estimators" and "max_features" via 3-fold cross validation to maximize the "score". Show the optimal values of parameters and the "score" on **train data**. Denote the optimal model as "Reg_B".
3. **(score = 15)** Apply AdaBoostRegressor() to the **train data**. Simultaneously tune parameters "n_estimators" and "max_depth" = 1,2,3 via 3-fold cross validation to maximize the "score". Show the optimal values of parameters and the "score" on **train data**. Denote the optimal model as "Reg_C".
4. **(score = 5)** Perform comparison of models "Reg_A", "Reg_B", and "Reg_C" by **test scores**. Which models overfit **train data**?

Problem II. Classification by SVC (score = 50)

Dataset is:

$$X_2, y_2 = make\_gaussian\_quantiles(n_{samples} = 1000, n_{features} = 12, n_{classes} = 2, random_{state} = XY)$$

Use **train_test_split** function and split the dataset into 70% - 30% portions, respectively.

1. **(score 15)** Apply SVC with kernel = "poly" to the **train data**. Simultaneously tune parameters "degree"=1,2,3 and "C" to get the model with the maximum "accuracy" via 3-fold cross validation. Show the optimal values of parameters and the "accuracy". Denote the optimal model as "Class_A".
2. **(score 15)** Apply SVC with kernel = "rbf" to the **train data**. Simultaneously tune parameters "C" and "gamma" to get the model with the maximum "accuracy" via 3-fold cross validation. Show the optimal values of parameters and the "accuracy". Denote it as "Class_B".
3. **(score 20)** Compare models "Class_A" and "Class_B" on the **test data**:
    a. By accuracies
    b. By ROC curves and the corresponding AUCs
    c. By PR curves for each class and the corresponding PR-AUCs
    d. Which models overfit **train data**?