ECE/CS/ME 539 – Fall 2024 — Homework 5

1. Please download **winequality-white.csv** dataset.

   AdaBoost (Adaptive Boosting) is a popular ensemble method that combines weak learners to create a strong classifier.

   (a) Partition the data into a 60/20/20 split for training/validation/testing and apply feature standardization (min-max standardization) to both the training and holdout data.

   (b) Start by fitting a single decision tree (as a weak learner) to the training data and report the prediction accuracy on the test data. Use `sklearn.tree.DecisionTreeClassifier` with default parameters.

   (c) Now train several AdaBoost classifiers with an increasing number of estimators (i.e., increasing number of weak learners). Use `sklearn.ensemble.AdaBoostClassifier`. Search this parameter between 1 and 100 estimators. Plot the accuracy as a function of the number of estimators.

   (d) AdaBoostClassifier allows you to configure other parameters for the training algorithm, including `learning_rate` and `algorithm`. Choose two additional parameters. Read the documentation and explain what they do. Try to improve on the model found in part (c) by searching for better values for those parameters.

   (e) For the best AdaBoost model you found in part (d), report its accuracy and confusion matrix on the test set.

   (f) Now, let's introduce XGBoost, another powerful boosting algorithm. Repeat steps (c), (d), and (e) using `xgboost.XGBClassifier` instead of AdaBoostClassifier. For XGBoost, consider parameters such as `max_depth`, `learning_rate`, `n_estimators`, `min_child_weight`, and `subsample`.

   (g) Compare the results of AdaBoost and XGBoost. Discuss the differences in terms of:
   - Accuracy on the test set
   - Training time
   - Model complexity (number of estimators needed for best performance)

   (h) Based on your findings, which boosting algorithm would you recommend for this dataset? Justify your answer.

2. (2 points) Load datafile **Xray.csv**. The label is at the last column.

   Develop a random forest classifier using 100 estimators. Perform 70/15/15 training, validation, testing stratified data partitioning. Please report validation accuracy, test accuracy and confusion matrix.

3. (4 points) Download **iris.csv** and perform the following:

(a) (2 pt) Use the 3rd feature x3 as the independent variable (feature) and the 4th feature x4 as the response, develop a linear regression model and plot the model.

(b) (2 points) Use (x3, x4) as the feature vector. Let Y = 1 if the feature has a class label = 3. Develop a logistic model, find the intercept and the coefficients **w** such that

logit = w_0 (intercept) + w_1· x3 + w_2· x4

estimate **w** in above equation. Next, denote xa = w_1· x3 + w_2· x4 and sort it from small to large (and Y and p accordingly). Plot xa vs Y (scatter plot) and xa vs p (line plot) in the same figure.

4. (4 pts) *Polynomial model order determination*

Download a dataset **re_dat.csv**. It contains two columns: the first column contains the indep. var. x_i and the second column contains the depend. var. y_i.

(a) (1 pt) Partition the data set into 90:10 training and testing datasets.

(b) (1 pts) Fitting the training dataset into a polynomial model with order P for P = 2, 3, ..., 15. Apply a 5-fold cross validation on the training dataset to fit the polynomial model. Choose the polynomial order P* that maximizes the validation accuracy.

(c) (2 pts) Develop a polynomial model with order P* using the entire training dataset obtained in (a). Evaluate the prediction error (R2 score) using the test dataset. Plotting the trained polynomial model using a line-plot over a uniform grid over the range of [-5, 5] in increment of 0.1. Then, overlay the line plot with a scatter plot of the testing data samples.