Assignment: Homework 4

How to Hand It In

1. Put all your solutions in one folder. Compress this folder and name it <firstname>_<lastname>_HW4.zip. For example, "Zhiyao_Duan_HW4.zip".
2. Make sure you have a report to show graphs and explain your answers, in addition to source code.
3. Submit to the corresponding entry on Blackboard.

When to Hand It In

It is due at 11:59 PM on the date specified on the course calendar. **Late assignments will receive a 20% deduction of the full grade each day.**

Problems (10 points in total)

1. (3 points) Multi-class classification and evaluation with the Glass dataset "glass.csv". The documentation of the glass classification dataset can be found at https://raw.githubusercontent.com/jbrownlee/Datasets/master/glass.names
    a. (1.5 points) Randomly partition the dataset into training, validation, and test subsets with the ratio of 6:2:2. Train a Decision Tree, Logistic Regression, and SVM classifiers using scikit-learn to classify them as 6 classes, tune the hyper-parameters on the validation set. Plot the colored confusion matrix on the test subset and report the accuracy.
    b. (0.5 points) Use SMOTE to resample the training subset, so that each class has 100 training samples. Retrain the three classifiers you have and test the performance. How are the results compared with those from a)? Describe and analyze your results.
    c. (1 point) Consider binary classification: window glass (classes 1-4) and non-window glass (classes 5-7). Choose one best classification algorithm from the previous question and train a binary classifier. Predict a score for each test sample. Draw the ROC curve and compute AUC. Choose a threshold for the scores, report the true positive rate, true negative rate, false positive rate, false negative rate, precision, recall, accuracy, F1-score of the classifier.

2. (3 points) Regression on 1-d toy data. "Q2_data.npz" contains a 1-d regression dataset, with a training set and two test sets. The first column is the input feature x, and the second column is the target value y. **Note:** you need to implement the regression models in this problem.
    a. (0.5 points) Load the data and visualize the training and test data points on the same graph with different markers using a scatter plot.
    b. (0.5 points) Fit the training data using linear regression minimizing the squared error loss. You can solve it using the normal equation and

pseudoinverse. Compute the Root Mean Squared Error (RMSE) and Mean Absolute Deviations (MAD) on the training and two test sets, respectively . Draw the regression function on top of the data scatter plot. Describe and analyze your results.

   c. (1 point) Fit the training data using polynomial regression with a p-th order polynomial, using the squared error loss **without** L2 regularization. You can again solve it using the normal equation and pseudoinverse. Vary p from 2 to 10. Report the RMSE and MAD on the training and test sets for each p. Draw the best three regression curves in terms of RMSE on the training set on top of the data scatter plot. Describe and analyze your results.
   d. (0.5 points) Fit the dataset using polynomial regression with a p-th order polynomial, using the squared error loss **with** L2 regularization. Repeat the remaining steps described in c).
   e. (0.5 points) Design some nonlinear features (e.g., polynomials, sinusoids, exponentials, and some kinds of combinations of them) based on x, and then run linear regression with squared error and L2 regularization. Report the RMSE and MAD metrics on the training and test sets. Draw the regression curve on top of the data scatter plot. Do you think it is a better fit than polynomial regression in c) and d)?

3. (4 points) Regression on a real-world dataset. “winequality-white.csv” is the Wine Quality Dataset that involves predicting the quality of white wines on a scale given chemical measures of each wine. **Note:** you are expected to use your own implementation of linear regression and kernel ridge regression in this problem. If you use external packages, points will be deducted. For support vector regression and the multi-class classifier, you can use external packages.
   a. (0.5 points) Preprocessing. Check whether there are duplicate values in the dataset. If yes, drop the duplicate data samples. Then partition the dataset into training, validation, and test subsets. Use the validation set to tune the hyperparameters (e.g., which kernel to use, hyper-parameters of the kernels, which classifier to use, etc.) and report the MSE, MAE, RMSE performance on the test subset for the following b)-d) questions.
   b. (1 point) Train a linear regression model and report the performance.
   c. (0.5 point) Train a kernel ridge regression model and report the performance. Apparently, you need to design your kernel. There are many kernels to try, e.g., polynomial kernels, Gaussian RBF kernels, and combinations of them. You can also try to define kernels using other nonlinear function such as sinusoids, exponentials, etc.
   d. (0.5 point) Train a support vector regression with different kernels and report the performance.

e. (0.5 point) Train a multi-class classifier to solve this problem, treating each integer rating as a class. You can choose any classifier until you get satisfying performance on the validation set. Report the accuracy and confusion matrix on the test set.
f. (1 point) Compare regression and multi-class classification. Choose the best regression model from your b)-d) and quantize the output prediction to integers. Report the classification accuracy and confusion matrix of the regression model. How is it compared to your classifier in e)? Analyze your results.