Homework #10
Linear Regression and *k*-NN

Overview

In this assignment, you are going to implement two learning algorithms, linear regression and *k*-NN, and do some experiments to compare their performances. You are provided with a program in which the linear regression part is already implemented, but the codes for *k*-NN are left incomplete. Your task is first to complete the codes for *k*-NN, then to improve the overall design of the program, and finally to rewrite the codes according to the new design. After that, you will conduct some experiments of testing the performances of the two learning schemes with the provided training and test data.

Implementing *k*-NN

Read carefully the provided codes for linear regression and then write the codes for *k*-NN. Note that the output of *k*-NN for a query example is made by first scanning the training set to identify the *k* examples that are closest to the query example, and then averaging their target values. A typical outcome of running the program is shown below. Notice that the root mean-squared error (RMSE) is the measure we use for evaluating each learning algorithm.
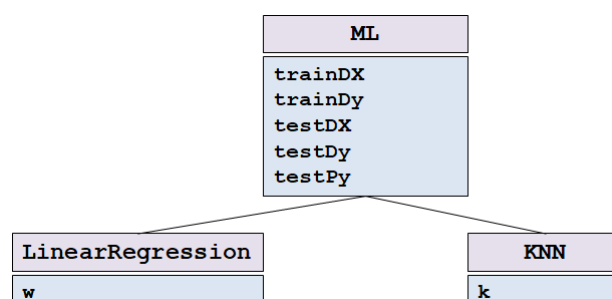
```
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 3

RMSE:  24.88
```

New Design

Considering later expansion of the program to include a variety of learning algorithms, it seems necessary to establish a subclass for each learning algorithm under the base class 'ML'. The following figure shows a possible class hierarchy with some class variables. You need to rearrange or rewrite the previous codes to fit to this new class design.

A typical outcome of running the revised program would look like the one shown below.

```
Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt
Enter the value for k: 5

RMSE:  1.31
```

Experiments

You are given two kinds of data: one is linear and the other is nonlinear. Each of both the data consists of a training set of 500 examples and a test set of 100 examples. Use the training set to build a model and test the model's performance using the test set. In the case of k-NN, you need to empirically determine the best value of k for each type of data.

Your report of experimental results should include the following numbers:
  - RMSE of linear regression with linear test data
  - The best value of k for k-NN for linear data
  - RMSE of k-NN with linear test data
  - RMSE of linear regression with nonlinear test data
  - The best value of k for k-NN for nonlinear data
  - RMSE of k-NN with nonlinear test data
More importantly, the report should also include the discussions on your findings.