2023 Spring, INE5008 Machine Learning with Data Mining, by Kichun Lee

HW 5

Due 11 am, May 11, 2023

**PART #1**

Refer to the attached *data\_1.txt*. We are considering the following three regression models for Price:

(Model 1) Price =

(Model 2) Price =

(Model 3) Price =

Use (1) 5-fold cross-validation, (2) randomized 80%training-20%testing validation, (3) AIC, (4) BIC, and (5) adjusted R2 find the best model among the three.

**PART #2**

(Open question) In a multivariate regression analysis, does multi-collinearity affect the model’s significance? Does it affect each variable’s significance?

Describe what to do if multi-collinearity is suspected.

**PART #3**

Suppose you have the following training data set:

|  |  |  |
| --- | --- | --- |
| x1 | x2 | Y |
| 1 | 1 | 0 |
| 1 | 3 | 0 |
| 3 | 3 | 0 |
| 3 | 1 | 0 |
| 3 | 2 | 1 |
| 2 | 3 | 1 |
| 1 | 2 | 1 |
| 2 | 1 | 1 |

Let us use k-nn with k=3 and Manhattan (L-1) distance. Compute the probability for the label, Y, being 1 for an observation (1.5, 1.7).

**PART #4**

The file *x\_train.txt* contains 7000 observations (rows) of some frequency features for individuals using an application of accelerometer and gyroscope in Samsung Galaxy S II. For the app, refer to the following:

https://www.youtube.com/watch?v=XOEN9W05\_4A&t=2s

https://www.youtube.com/watch?v=ti4HEgd4Fgo

The number of features (variables) for a row is 561. For the side information of the 561 features, refer to the attached inputvariables.txt and inputvariables\_info.txt.

The associated file *y\_train.txt* contains the corresponding 7000 response statuses: 1 WALKING, 2 WALKING\_UPSTAIRS, 3 WALKING\_DOWNSTAIRS, 4 SITTING, 5 STANDING, 6 LAYING.

Predict the statuses of the observations in X\_test.txt using the k-nn method. For simplicity, try to find a good ‘k’ value among k=3,4,5,6,7,8 by a randomized cross validation approach of 60% training and 40% testing. Present a result file, *y\_test\_pred.txt*, that contains the predicted statuses of the 1400 observations along with your code file.

(The code can be in R, Python, Excel or MATLAB, or JAVA or whatever computing language you like)