Machine Learning Final Project

**House Price Prediction**

2015-30194 김광훈

2015-20935 서효원

**How to use Program**

knn\_program.m: knn based classifier (regression)

1. initialize=1 (on command window)
2. testfile=’\_\_\_\_\_\_\_\_\_’ (on command window)
3. press ‘run’ on GUI
4. To use knn with additional 200 test set. Type MODE=2 on command window.

prediction: MLP based regression

classification: MLP based classification

**1. Problem Description**

Given part of Ames Housing Data, consisting of 80 variables directly related to the quality and quantity of many physical attributes of the property:  
1. Predict house price with regression model  
2. Classify the house price into 2 classes (price<160000 and price 160000)

**2. Data Description and Processing**

We used discrete and continuous attributes as it is.  
For nominal data, we one-hot encoded each item in every attribute.  
For ordinal data, we assigned an integer matching order of the items.

Example:

MSSubClass (Nominal)

Raw Data: 20, 30, 40, 45….

Processed: MSSubClass\_20 - 0 or 1

MSSubClass\_30 - 0 or 1

…

ExterQual (Ordinal)

Rawdata: Po, Fa, TA, Gd, Ex

Processed: 0, 1, 2, 3, 4

Before applying learning algorithm, standardized the data into the range -1 to 1.

To check our model performance without additional test data, we used K-fold cross validation method.

**3. Algorithm and Result**

**A. Price Prediction**

**i. Program your price prediction model that can predict the house model.**

We’ve made regression model for price prediction with k nearest neighbor algorithm and multi-layer perceptron artificial neural network model.

*k-nearest neighbor(knn) algorithm:*

The idea of using knn algorithm is based on the common sense that the price of houses with similar area in square feet and physical options would worth similarly.

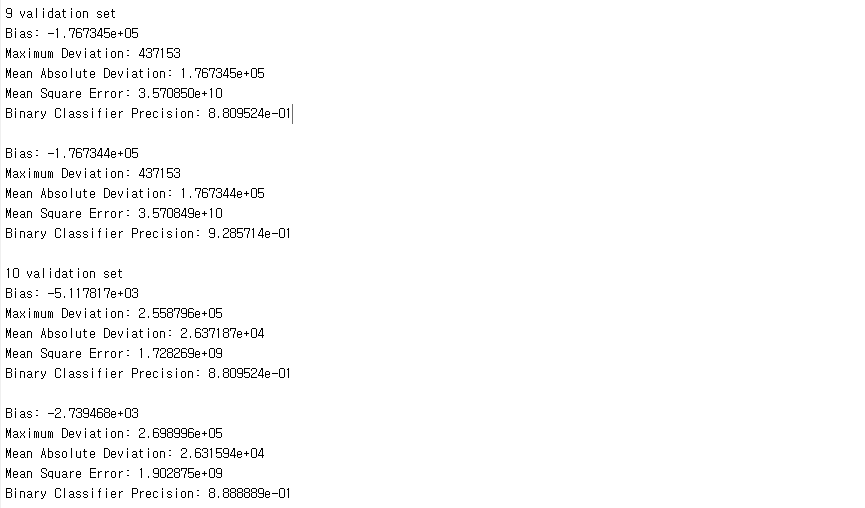
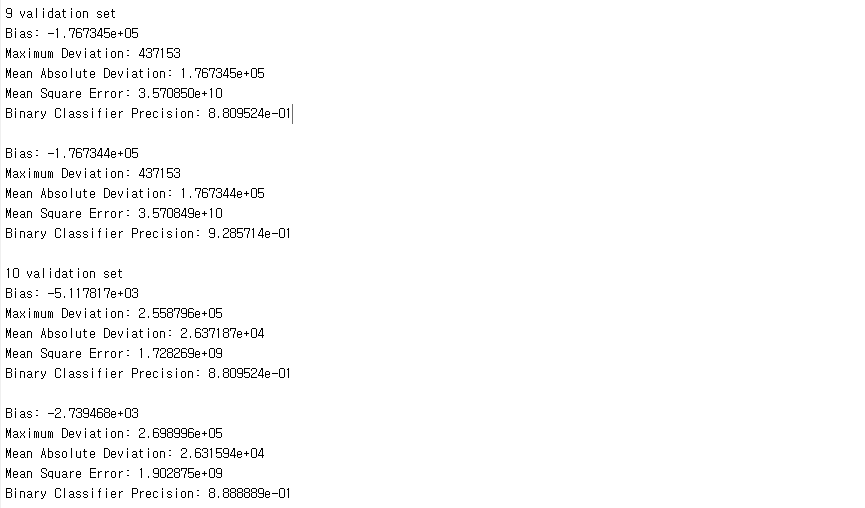


Figure Result using knn (above without PCA dimension reduction, below with PCA dimension reduction)

For price prediction, find k nearest neighbors of target test data from given training set. The predicted price is calculated by simply averaging the sale price of the found neighbors.

The maximum deviation is given quite large meaning that there exists a great outlier that we could not find using knn model. Mean Absolute Deviation was about 26000. It isn’t small value, but is acceptable for simple algorithm like knn.

**ii. Show which variables are comparatively more related to the house price**

From knn result, it is hard to say applying PCA made model to work better with knn.

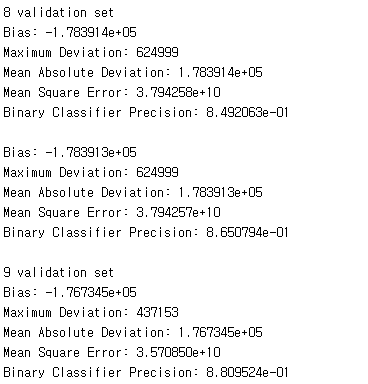
B. Price Range Classification

**i. Program Classification models that can classify the house price into 2 classes (price<160000 and price>=160000)**

For classification, we’ve used knn algorithm and MLP ANN model.

*knn algorithm:*

The thought behind using knn is described in the same section of regression model.



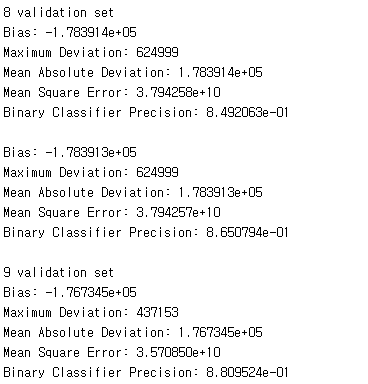


Figure Binary Classification using KNN model

Naïve knn approach showed about 0.85 and 0.88 precision rate with two of 10-folds model. Average of all 10-fold cross-validation data set is given as:



Figure Average Binary Classification Precision of 10-fold cross validation set

The way of improving above rate is explained next section.

**ii. Show which variables are comparatively more related to the house price class**

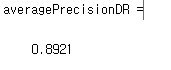
Applying PCA (dim reduction from approx. 200 attributes to 40 attributes) to the training data sets and then do knn classification resulted in:

Figure Average Binary Classfication Precision of 10-fold cross validation set after applying dimension reduction using PCA

It can be seen that there was approximately 1.5% point increase in average precision compared to the naïve knn algorithm. However, the 1.5% point seems too small and does not really mean much.

**ii. Improve the model performance**

To improve knn model, we have utilized ensemble method. Ensemble is composed of 100 bootstrap sampled training set. We have used bootstrap sampled training set to get 100 many classification result then major voted the result using knn model. For each bootstrap samples, 20 attributes are randomly selected. The result is given below:

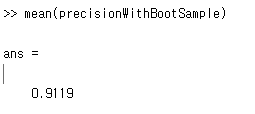


Figure Average Binary Classification Prediction of 10-fold corss validation set with bootstrap sampled ensemble method

Compared to the 87% average precision, it seems that using ensemble somewhat increased the precision of classification model.