BUAN 6341 - Machine Learning

Assignment 3

Yung-Kuei Chen

yxc177030

Data Description

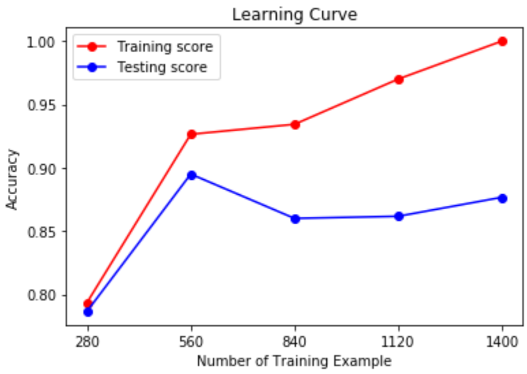
Nowadays, smart phones become an essential part of our life. To have a great phone with reasonable price is the first consideration when people are trying to buy a new phone. Therefore, it is very important to understand the relationship between price and performance about mobile phones to have the highest cost-performance ratio. This is the reason why I chose this mobile price classification dataset. The problem for this dataset is to correctly classify the price range of each cell phone from their features. There are 21 variables in the dataset. The target variable price_range is a categorical variable contained the integer number from 0 to 3. The price_range is 0 means that the lowest price group and the price_range is 3 means the highest price group. The independent variables are the features about the mobile phone. There are some numeric variables such as the size of RAM, the length of the talk time, the speed of CPU and so on. Besides, there are some dummy variables in this dataset, for example, the dual_sim presents whether the phone supports dual sim card or not.

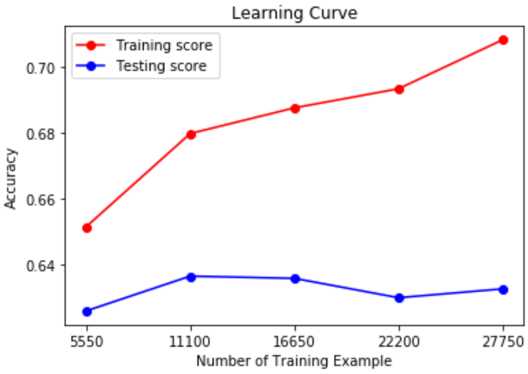
Another dataset is the online news popularity dataset from previous assignment which we want to classify the article as a high share (1) or low share (0) from the attributes of article. This time, I will put all the variable into the model to see how the model performs.

I will apply artificial neural network(ANN) and K-nearest neighbor(KNN) to approach these classification problems.

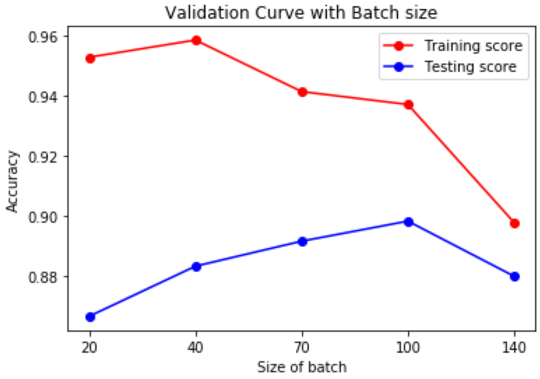
Error Rate

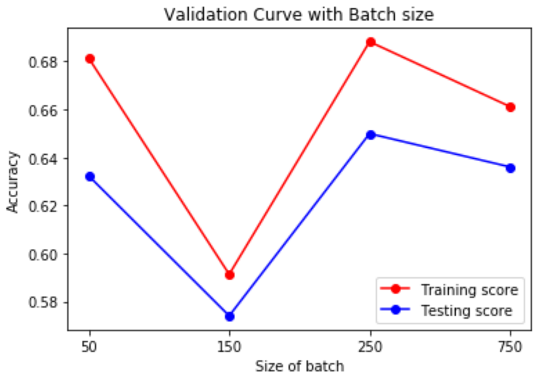
Artificial Neural Network (ANN)





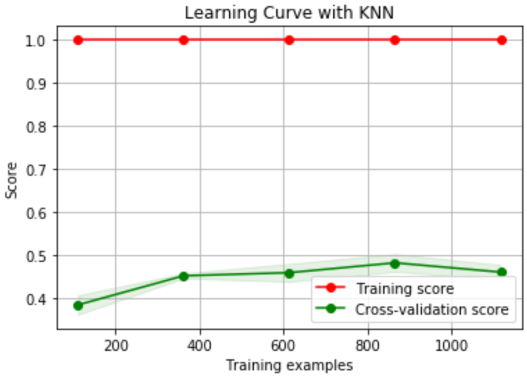
The accuracy is slightly related to the number of training example. The left-hand side picture tells us that the trend of the accuracy is increasing if we have more training example to train the model. The picture on the right-hand side tells us that the more training example leads to higher training score, however, the testing score remains the same. This means that the parameters of the model should be changed to prevent overfitting problem.

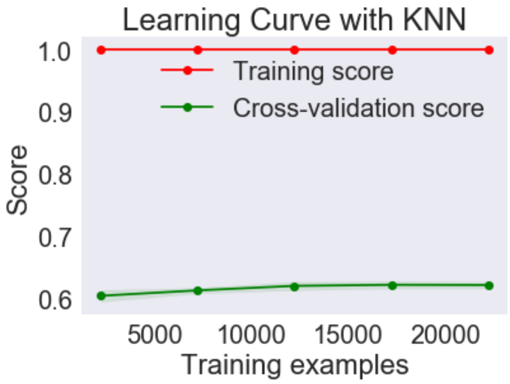




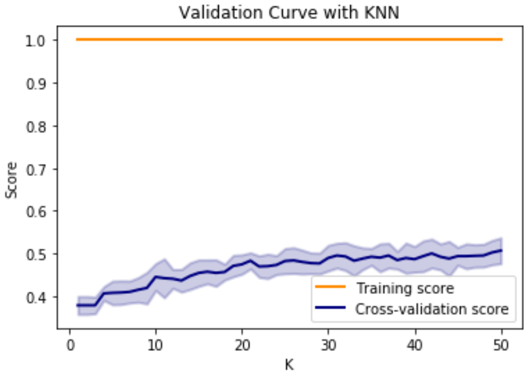
The reasonable batch size can lead to more accurate result. Choosing larger size of batch will give model a more general concept about the dataset. Therefore, the result will be more accurate. However, if choosing oversize batch, the accuracy will decrease because the information in the batch is too messy to decide the direction to find the optimal solution. (Ps: the result on the right-hand side has an unusual decrease is because of the seed setting in the Tensorflow package. Can be fixed if using other random seed)

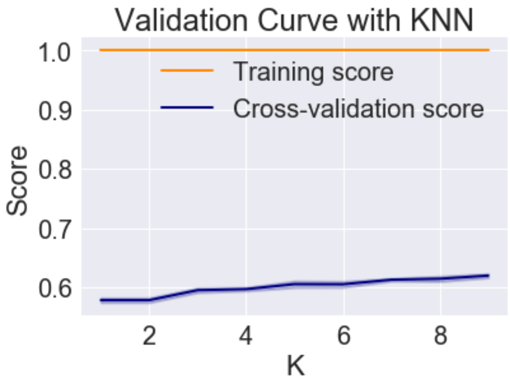
K Nearest Neighborhood (KNN)





The reason why training score in the KNN is always 100% is because the training point already exists in the dataset. Therefore, when KNN looks up in the dataset, it will find the correct label. From the result we can see that if we train the model with higher amount of training examples, the accuracy will increase. This is reasonable because KNN can find more accurate and close points when look up in the whole dataset.

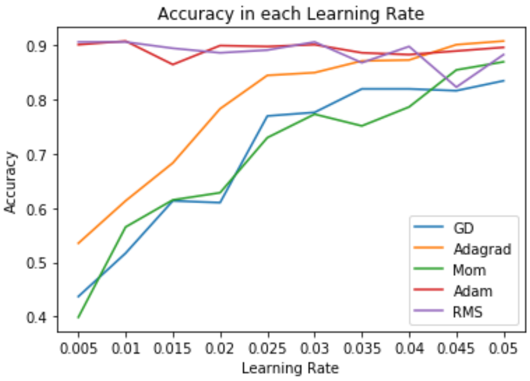


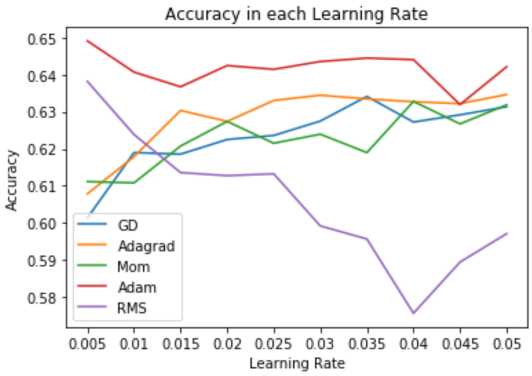


When choosing different number of K in the model, the bigger K will lead to computation cost and time spending. It is because algorithm should find more close points in the dataset. Therefore, it is important to compare the training cost and accuracy together to get the most suitable model.

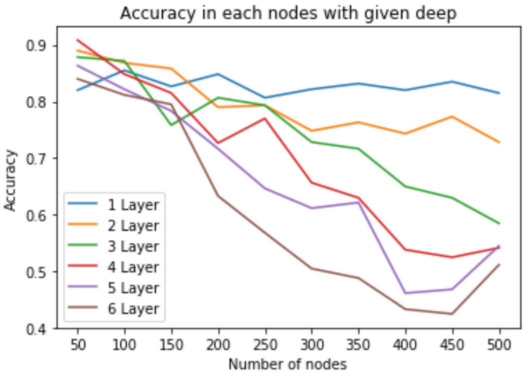
Performance Comparisons

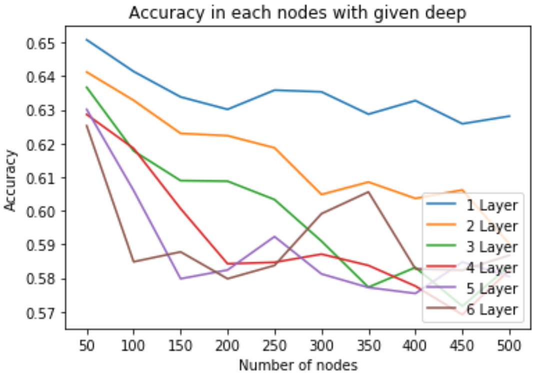
Artificial Neural Network (ANN)



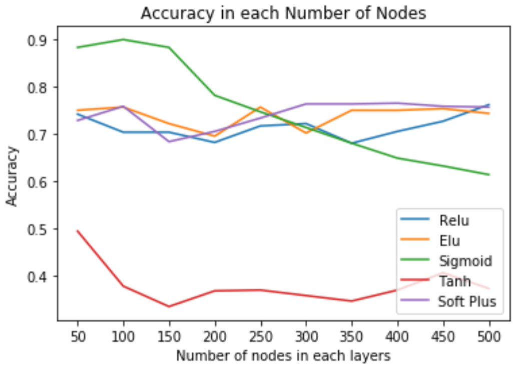


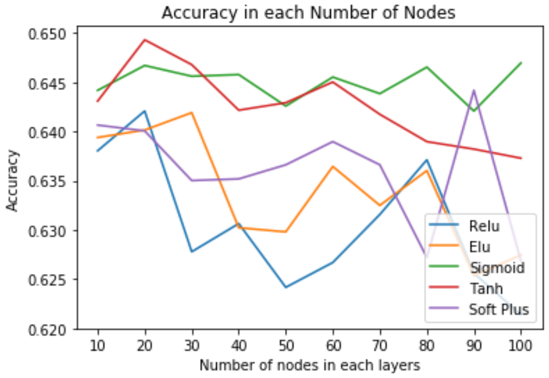
From the result on the left-hand side, we can see that some optimizers with lower learning rate are facing the underfitting problem. This is because that the low learning rate causes the process of convergence slow. The model will eventually converge but it cannot converge efficiently. Therefore, it is very important to choose the right learning rate to build the model. Moreover, it is also essential to choose the right optimizer. From the result, the Adam optimizer performs better than others. The reason why Adam is the best optimizer is that Adam optimizer will adjust the learning rate after each epoch. This makes converge more efficient. As a result, I will focus on Adam optimizer and learning rate 0.05.





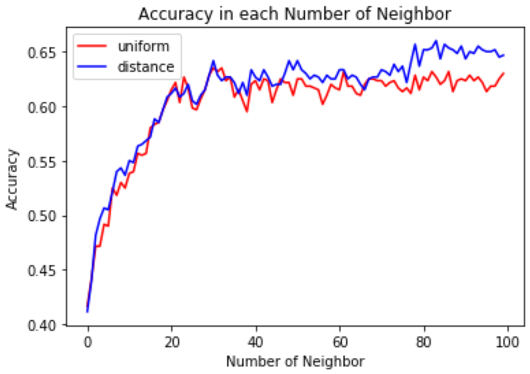
From the result after running 6 different layers with various numbers of node, the neural network with only one layer has the highest accuracy and consistency among all. Moreover, if the neural network has more and more layers, the accuracy will get lower. Therefore, we can focus on building fewer layer NN to approach this dataset. Since the variables in our data is very straight forward which means the variables have direct influence in the result, we do not need more than 2 layers to build our neural network. Thus, we chose 1 layer to build NN.

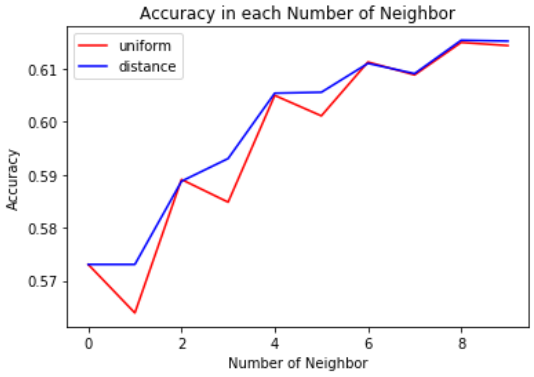




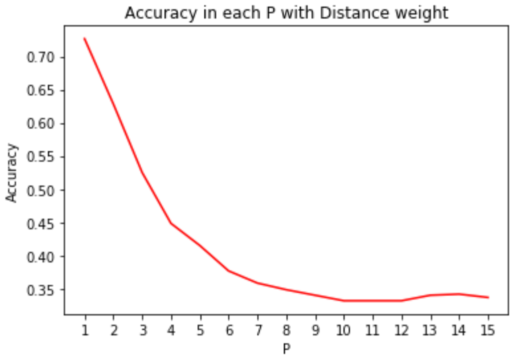
Examine what is the effect from different activation function. Choose five different activation functions: Relu, Elu, Sigmoid, Tanh and Soft plus to build five different models and to see how they perform. From the results, we can easily recognize that the Sigmoid function performs better in these two datasets with low number of nodes. The reason behind this is because that the data in our datasets has less negative values and huge amount of large positive values. The NN will perform better when restrict the data range in (0,1). Therefore, Sigmoid function is our first choice.

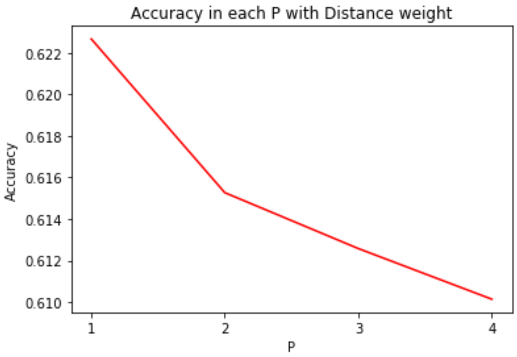
K Nearest Neighborhood (KNN)





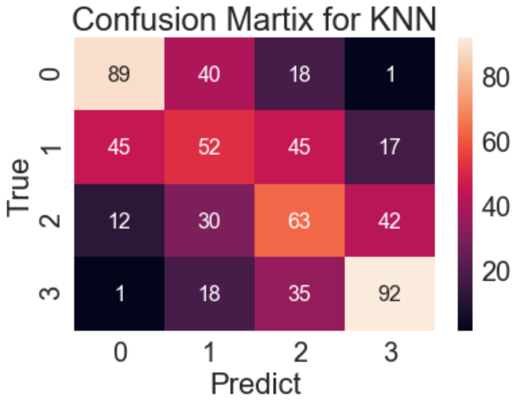
There are two main weight functions in KNN. One is uniform, the other is distance. These two functions provide different criteria to calculate the distance between points. From the result of these two datasets, the distance weight function is better than the uniform function. This is an expected result because the distance weight means the prediction is based on the weighted value from different distance. The closest point is more influential than other points

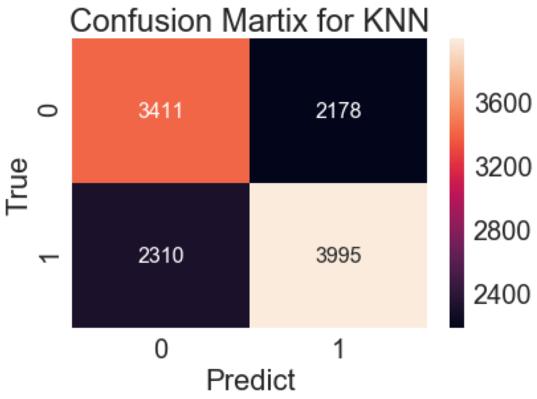




Then examined the different P for the Minkowski metric. From the result, we noticed that the higher P value will lead to lower accuracy. Therefore, we choose the low P for the distance metric.

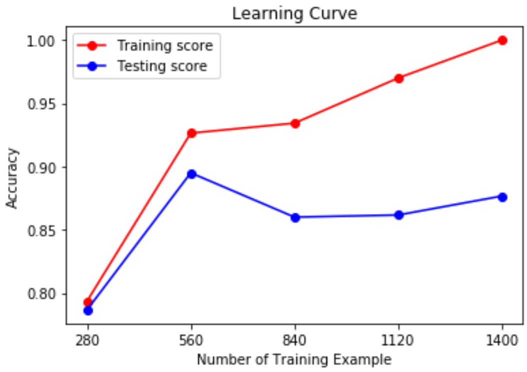
The results from the KNN are as follow.

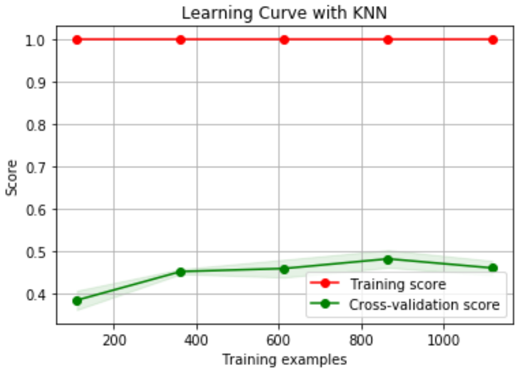




Comparison

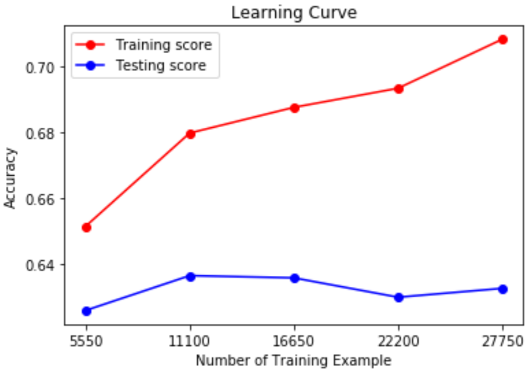
Mobile Price Dataset

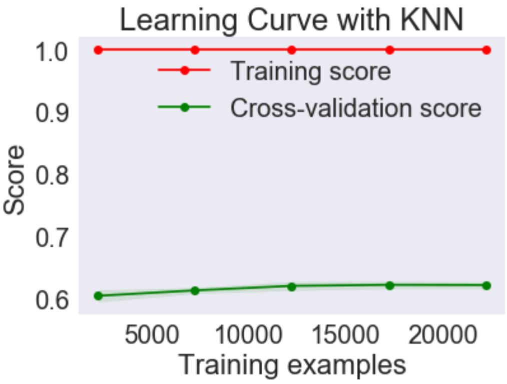




In ANN model, the accuracy reaches incredible 90% when choosing the correct parameters which is a very good result. On the other hand, KNN model does not perform well in this dataset. The accuracy is around 40%~50%. This may be solved by adding more valid data point into the dataset to make the prediction more accurate.

Online News Popularity Dataset

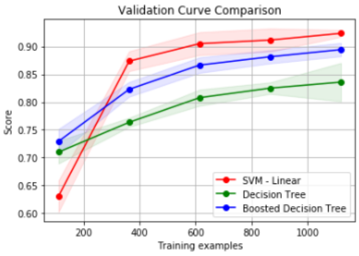


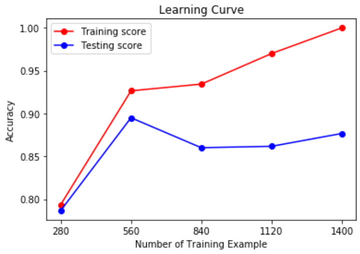


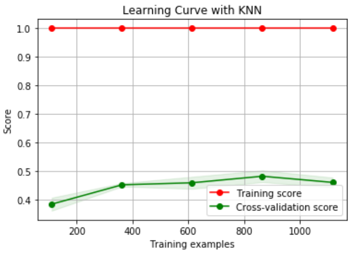
The changes of testing score in both models are small with respect to different training example. Overall, the ANN model has higher accuracy, but longer training time. Therefore, I should choose between two criteria - the accuracy and the training time.

Ranking of the algorithms

Mobile Price Dataset

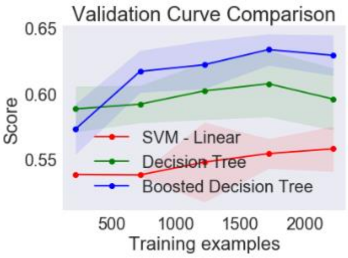


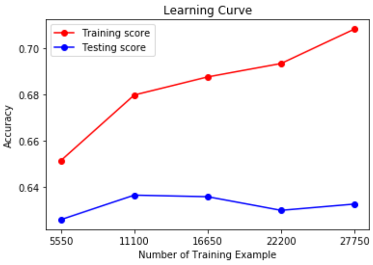


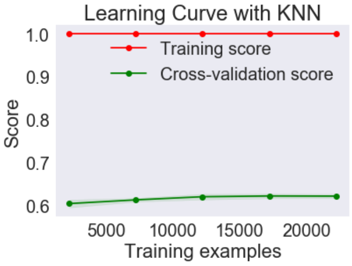


Among these 5 algorithms, the SVM model reaches the highest accuracy – over 90% accurate. Although ANN model performs almost as same as the SVM model, it takes a lot of time to train the model to reach such a high accuracy. Therefore, the SVM model is still the most suitable model for this dataset.

Online News Popularity Dataset







The ANN model is not the best solution for this dataset. The ANN model has low testing score and long training time which is very inefficient. The KNN model is still slightly increasing the accuracy when adding more training examples into the model. However, the accuracy of the KNN model is still lower than Boosted Decision Tree model. Therefore, the Boosted Decision Tree model is still the best choice for online news dataset.

Conclusion

After running 2 more algorithms on two different datasets, it is clear the priority of the model is the efficiency. Although the ANN model always outperforms in most of different dataset, it takes a lot of time to train the model to reach such a high accuracy. Thus, one cannot decide the model by only judging the accuracy but efficiency of that model.