BUAN 6341 - Machine Learning

Assignment 2

Yung-Kuei Chen

yxc177030

Data Description

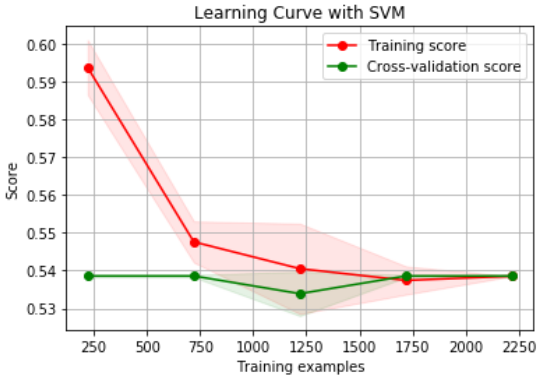
Nowadays, smart phones become an essential part of our life. To have a great phone with reasonable price is the first consideration when people are trying to buy a new phone. Therefore, it is very important to understand the relationship between price and performance about mobile phones to have the highest 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. However, it is not reasonable to put all 60 variables into model. Therefore, I ran the decision tree model to see the feature importance and chose the top 10 important variables to put into the models. Furthermore, to increase the process, I used only 10% of the total data to avoid intense computation.

I will apply Support Vector Machine (SVM), decision tree and boosted decision tree to approach these classification problems.

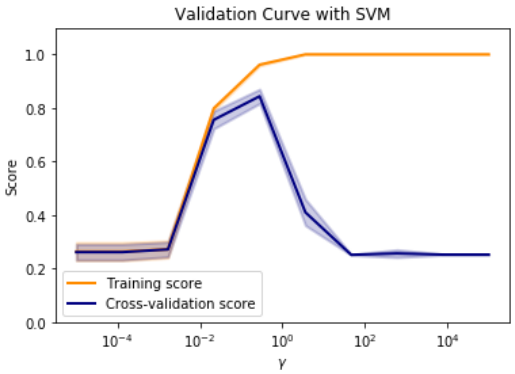
Error rates

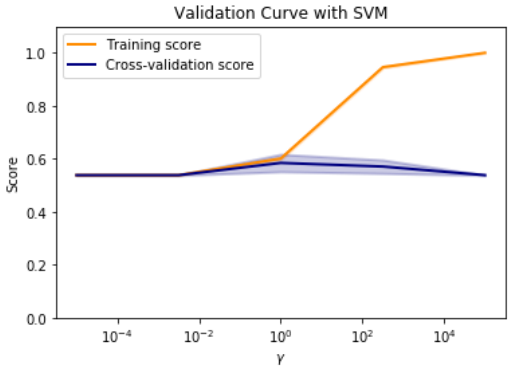
SVM Model



Picture.1 Mobile Price Picture.2 Online News

For this SVM classifier model, I applied the default RBF kernel in sklearn package. The picture.1 and 2 present the learning curve. With more training example, both the cross-validation (CV) score and training score are going higher. That means SVM model will have lower bias when putting more training examples into the model. Moreover, the difference between training score and CV score is getting smaller that means the SVM model has lower variance as the training examples growing. As a result, more training examples is better for SVM model.





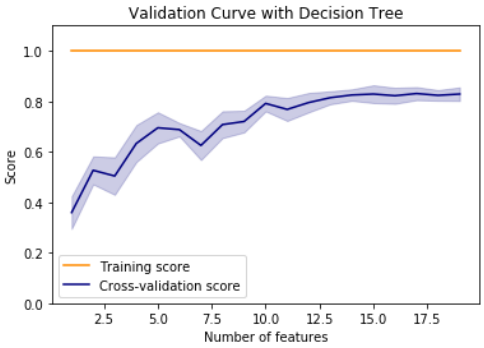
Picture.3 Mobile Price Picture.4 Online News

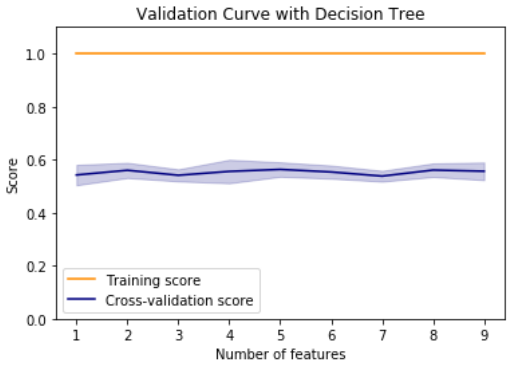
The second approach is the validation curve with the respect of gamma value - the value to control the kernel coefficient. From picture.3, there are no significant difference between the training score and CV score when the gamma is lower than 0. However, when gamma is larger than 0, the CV score decreases dramatically to 0.2 and the training score is still on the way increasing to 1. From picture.4, the difference between training and CV score starts from gamma equal to 1. Therefore, we should choose gamma value more carefully to have more generalized model. We will choose gamma equal to 1 to tune the SVM model for both datasets.

Decision Tree

Picture.5 Mobile price DT with Gini Picture.6 Mobile price DT with Entropy

The decision tree models are built from the sklearn package with 2 criterions – “Gini” (picture.5) and “Entropy” (picture.6) holding other parameters default. Unlike the model with entropy criterion, the CV score in Gini model is stably growing with the number of training examples. Thus, the Gini criterion is the better criteria for decision tree model. The training score in this decision model is always 1 means that the model correctly classifies all the data. However, the highest point for CV score is around 0.85 which is far from 1 means that this model is overfitting. Nevertheless, when putting more training examples into the model, the CV score is growing higher. The bias and the variance of decision tree model are getting lower means the model fits the data more. This proves the point that the more training data the better.

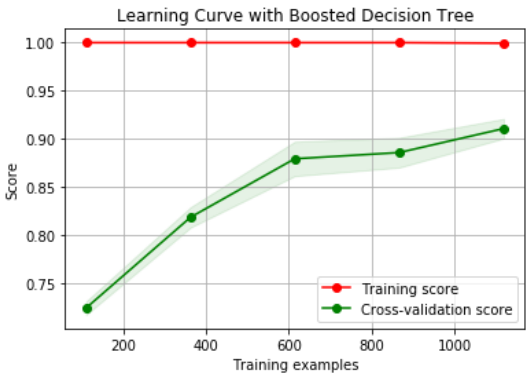


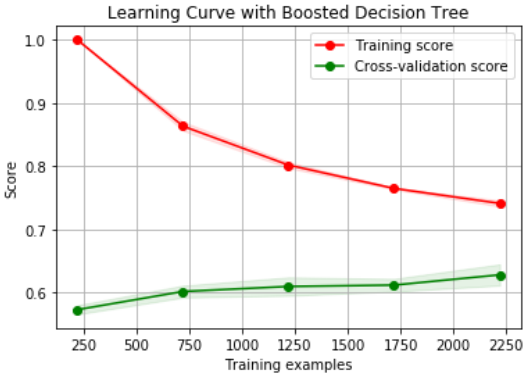


Picture.9 Mobile Price Picture.10 Online News

Another approach to decision tree is the number of features used in the model. From the validation curve in picture.9, the more features we used in the model, the lower bias and variance we can get. This is because that the model performs well when using more dimensions of data to predict the result. However, for picture.10, using less than 10 features does not have significant impact on accuracy in online news dataset. It is because that the feature importance is not significantly different in online news dataset. All the variables are very not important to the model.

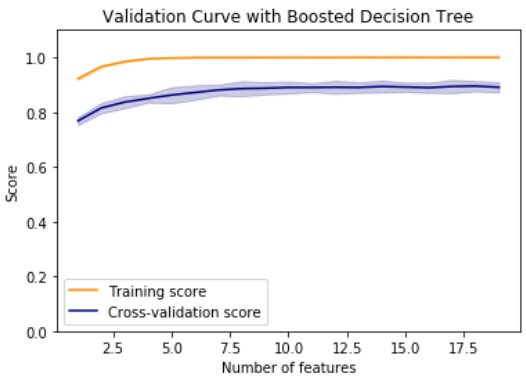
Boosted Decision Tree

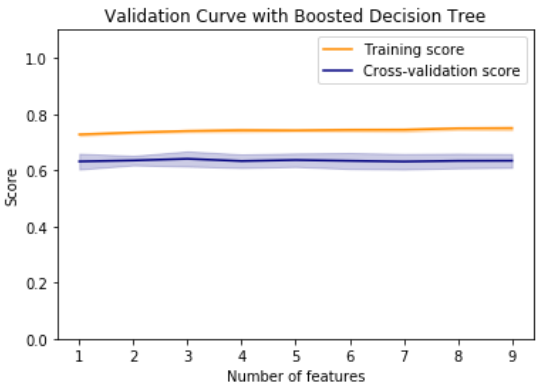




Picture.11 Mobile Price Picture.12 Online News

For the boosted decision tree model, I implemented gradient boosting classifier from sklearn holding all parameters default. Although the learning curve of boosted decision tree is looked similar to the decision tree model, the boosted decision tree has two advantages: higher accuracy and higher efficiency (Picture.11 and 12). Compared to the original decision tree, the boosted tree has higher accuracy. For example, in mobile price case, the accuracy from boosted tree is higher than 0.9 and the accuracy from decision tree is lower than 0.85 in more than 1000 training data examples. For the efficiency, the boosted model converges faster than the original decision tree. With around only 600 training examples, the boosted tree increases 15% accuracy from 75% to almost 90%. On the other hand, the decision tree increases only 10% accuracy from 70% to 80%. In conclusion, the boosted decision tree has a better mechanism to have better fit the data than original decision tree.



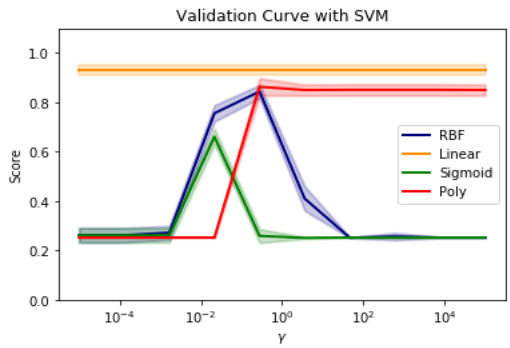


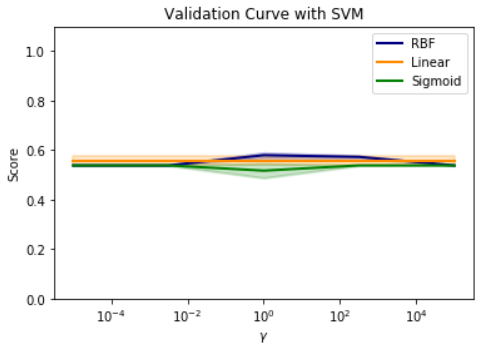
Picture.13 Mobile Price Picture.14 Online News

The second approach is the validation curve. From Picture.13 and 14 the training score and CV score is not changing a lot as more number of features used in the model. That means the boosted decision tree is not that sensitive to how many features used in the model.

Performance comparisons

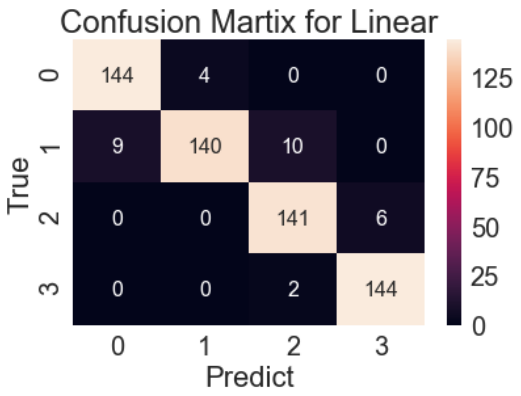
SVM

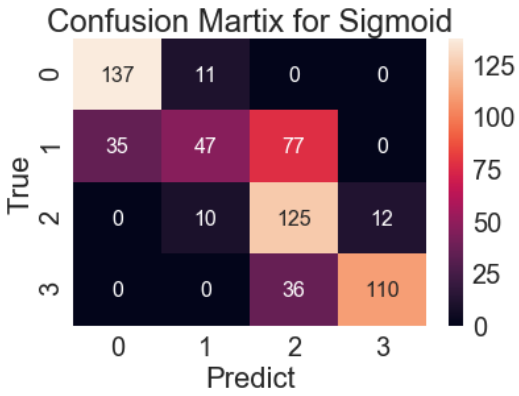




Picture.15 Mobile Price Picture.16 Online News

From previous result, we noticed that the more training example we used in training SVM model, the more accuracy we will get. Therefore, in the case of choosing the kernel in SVM, we only focused on the gamma value in kernel function. From picture.15 and 16, we chose different gamma for different kernel and plotted the confusion matrix.

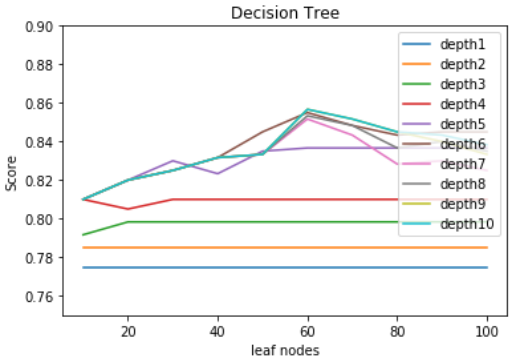


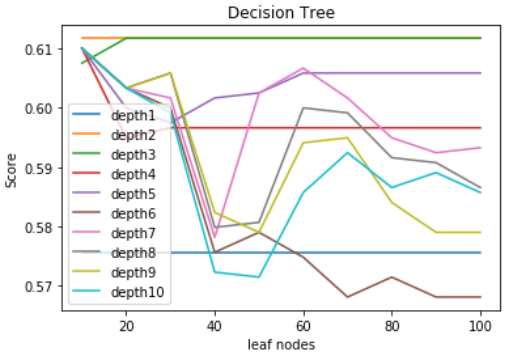


Picture.17 Mobile Price SVM linear Picture.18 Mobile Price SVM sigmoid

In mobile price case, for instance, the linear kernel did a decent job with around 95% accuracy (picture.17) on test dataset (linear kernel is not affected by gamma). For RBF and Poly kernel, the highest scores of these two kernels hit around 85% accuracy with parameter selection. However, for the sigmoid kernel, it only got around 70% accuracy even with the selection of gamma (picture.18). As a result, it is very important to choose the right kernel and the correct parameters to get a generalized model. However, in the online news case, there is no significant different between kernels. Hence, we can say that it is not efficient for us to use support vector machine model to classify the amounts of shares in online news dataset.

Decision Tree

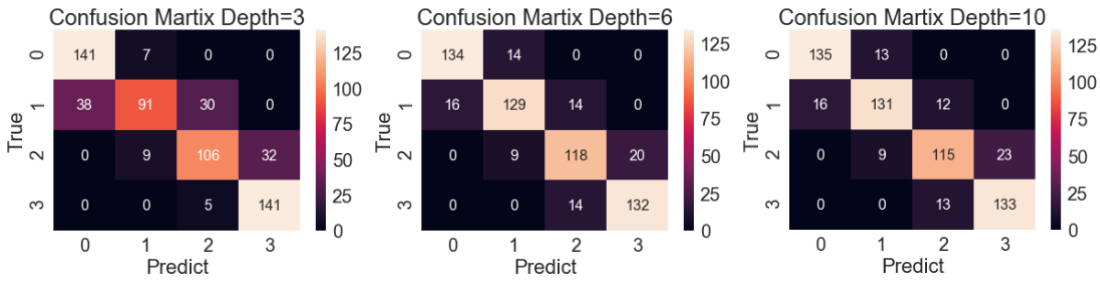




Picture.19 Mobile Price

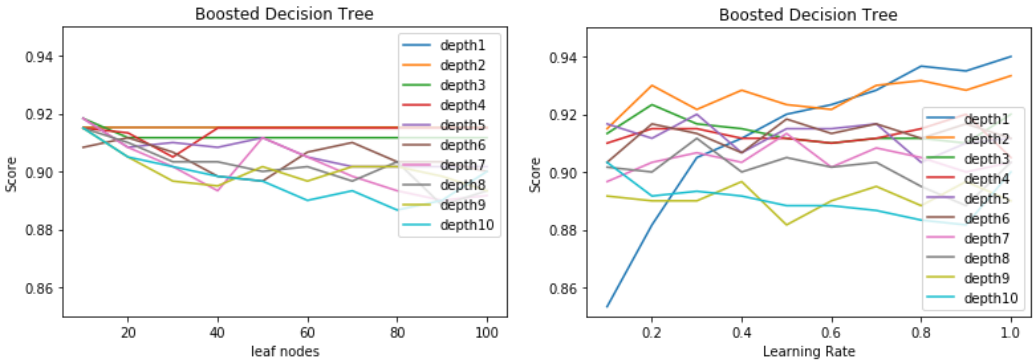
Picture.20 Online News

Experience with different depth and node to understand what is the suitable parameters for this dataset. Take mobile price data for example, from picture.19, we noticed that the decision tree generally performs not good when depth less than 5 means the model cannot efficiently capture the key features to make a correct decision. As the depth more than 5, the curves of accuracy are very similar (picture.19). The peak of accuracy is at the number of leaf nodes equal to 60. The reason why the peak is not at the highest number of leaf nodes is because of the overfitting problem. In conclusion, the parameter for this decision model should be depth no less than 5 and the total number of leaf nodes should be 60.



Picture.21 Mobile Price Data DT confusion matrix

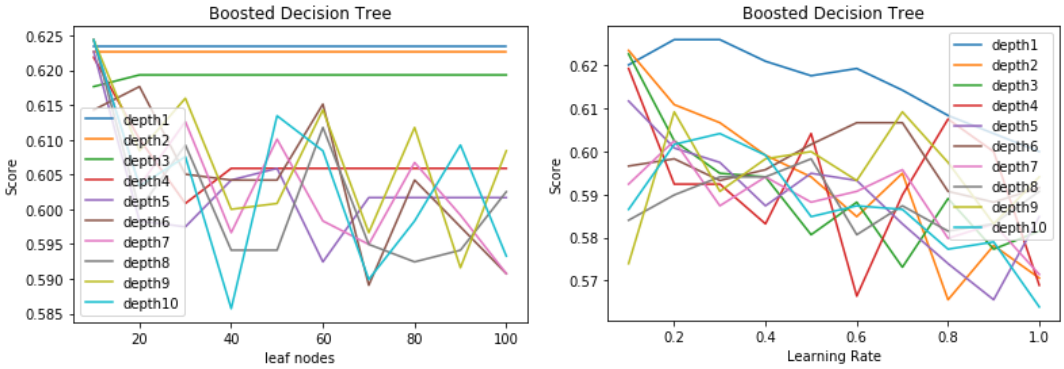
Boosted Decision Tree



Picture.22 Mobile Price BDT pruning

In the case of boosted decision tree, I changed the max depth of tree and the max number of leaf nodes holding other parameters constant. Different from the decision tree, the boosted tree tends to have less depth and less nodes. Even with only a few nodes in the tree, the boosted model still got higher accuracy than original decision tree model.

Second, using learning rate to prune the boosted decision tree model. For the model with fewer number of depth, the accuracy increases when learning rate increases. On the other hand, for the model with larger number of depth, the accuracy decreases when the learning rate increases. Therefore, we should find the best learning rate with the respect to the number of depth in the model.

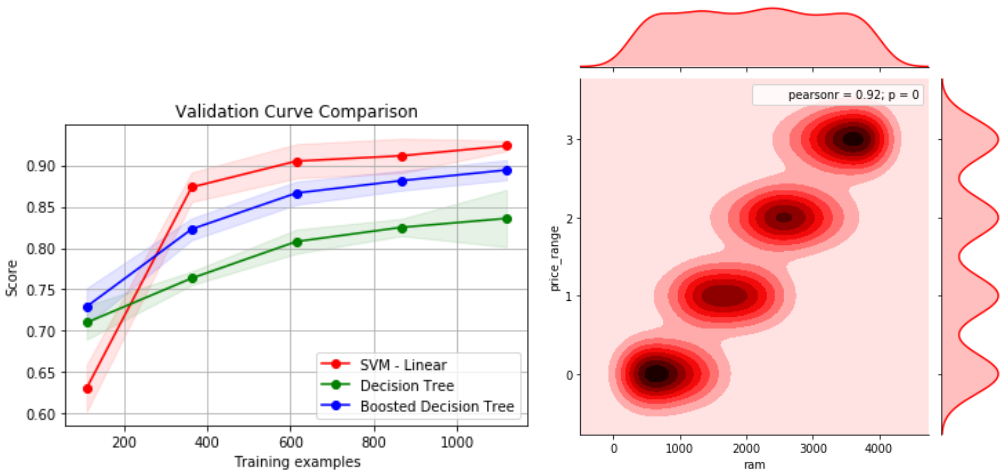


Picture.23 Online News data BDT pruning

For the online news model, the boosted decision tree also tends to have lower depth in the model. However, as the learning rate goes up, the accuracy from all the depths are going down which means that we should use smaller learning rate to slowly get to the optimal solution.

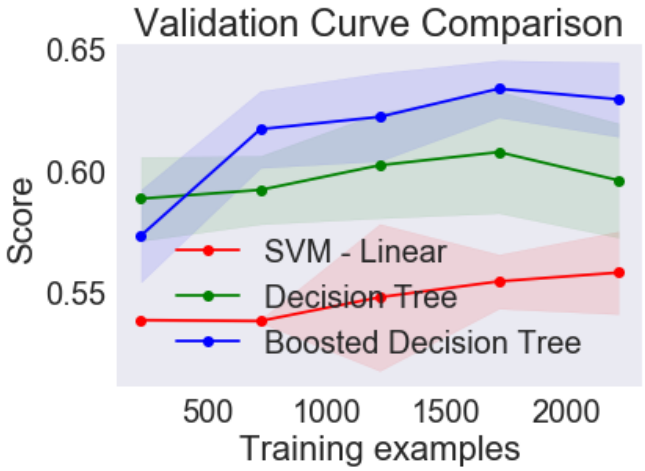
Comparisons of the three learning algorithms

Mobile Price Dataset



For these three algorithms, the more training examples used in training leads to the more accurate model. In this case, for small training dataset, the first choice is the boosted decision tree model. However, when it comes to larger training dataset, we will prefer to choose the SVM model because of the highest accuracy. It is because that the data in mobile price dataset is a clean dataset which the category is significantly different form other groups. That makes SVM model works pretty well. As a result, if we have enough data point to train the model, it is a great idea to use SVM in this dataset.

Online News Popularity Dataset



For the online news data, the SVM is not a good model to fit the data. The reason behind this is because that the data in news dataset has so many dimensions and it is hard to project the data point into even higher dimension to separate each point. The boosted decision tree performs better among these three algorithms. Although, in the small training example, the decision tree out performs than boosted tree, the boosted tree corrects the model very quickly to get the highest accuracy.

Conclusion

After running support vector machine, decision tree and boosted decision tree models on two different datasets, it is clear to know that for every different dataset we need to find a suitable model and try to find the best parameter to make model more fit the data. The SVM works outperform on mobile price data. The boosted decision tree works better in the online news data. For every different situation, choosing the right model will save lots of computational power and will make the model more accurate.