Supervised Learning

Georgia Institute of Technology CS 7641: Machine Learning Assignment One

Xi Han

GT ID: xhan306

Introduction

This report explores five supervised learning models in machine learning by implementing models, tuning parameters, investigating the influence of training data size on model performance. Decision Tree, Neural Networks, Boosting, Support Vector Machines (SVM) and k-nearest neighbors (KNN) were used for two different datasets, protein dataset and bank dataset to solve classification problems and were compared in this report.

Dataset Description

1. Protein dataset:
This dataset includes gene name, protein sequence and protein solubility for 3148 proteins. All the proteins are from microbe Escherichia coli. Escherichia coli is a bacterium commonly used in genetic engineering to express recombinant proteins, which is a pivotal process in biotechnology. Producing some proteins in E. coli was, however, inefficient in industrial setting, resulting from low activity of the recombinant proteins. As biocatalysts and therapeutic agents, proteins with high activity have lower unit production cost. Developing biocatalysts with high activity has thus become an important goal for further development of many biocatalytic processes. Some experimental strategies in vivo can improve the expression of recombinant proteins. Such empirical optimizations have, however, been time-consuming and expensive. Moreover, experiments often fail due to opaque reasons. A generic solution is highly desired for enhancing the heterologous protein overexpression and may be ultimately provided by using a computational model that can predict activity of any enzyme accurately from its amino acid sequence and other input information. Developing such model would require a large dataset which contains at least two columns, protein sequence and activity. Since activity and solubility are correlated for some proteins, and there is a relatively large dataset available for solubility as solubility data from different proteins can be pooled together. This dataset is used to predict protein solubility form protein sequence.
At first, some data pre-processing is implemented to convert protein sequence with 20 characters into amino acid composition with numerical values by a descriptor of protr package in R and solubility was converted into 0 and 1 by threshold 0.5. 1 represents this protein is soluble and 0 is not soluble. There are 21 dimensions and last column is our target value protein solubility.

2. Bank dataset:
This dataset was used to build a model predict whether the client would subscribe the term deposit product or not and compare the practicability of several data mining classification using the marketing campaigns records dataset from Portuguese banking institution. Accuracy is adopted as the evaluation metric for the classification models.

There are 30891 observations in the bank. They both have 10 continuous variables such as personal information, social economic index and 10 categorical variables such as individual's job and education level. The target response (y) which is presented as 'Yes' and 'No' is a binary response implying whether the client agree to subscribe to a term deposit or not.
At first, target response "Yes" was converted to 1 and "No" was converted to 0. There are no missing values in the dataset. For model KNN and neural networks, factor was needed to convert into numerical values for prediction.

Methodology

For both dataset, 30% of original data was set as test data and 70% was training data. Various supervised learning models were utilized on training dataset firstly and train accuracy was recorded. Then 10-fold cross validation was used to tune hypeparameters of different models. Best models were selected according to the 10-fold cross validation Accuracy (CV Accuracy). Then best hypeparameters were utilized to explore the relationship between training data size and model accuracy. Finally, the test data which was held out was utilized to compare performance of five supervised learning models and analyse the differences between different datasets. Each model selected is labelled by yellow colour in tables according to the CV accuracy.

Supervised Learning Models

1. Decision trees

Depth of trees were tuned for decision tree model. The depth of the trees, regulated by mincriterion was changed to observe the performance of model for different depth. A split is implemented when the criterion exceeds the value given by mincriterion. For example, when mincriterion equals 0.95, the p-value must be smaller than 0.05 in order to split this node.
To grow large trees, we can set mincriterion to a small value and to prune trees, we can set a large mincriterion. It can be seen from the Table 1, when we tuned mincriterion, there are different nodes in the tree. When mincriterion increases gradually, there are less nodes in the tree and pruning was made. The train accuracy gradually decreases when the pruning was made and CV accuracy increases first and then reduces. When mincriterion is small, overfitting occurs because the training accuracy is high and CV accuracy is low. The model can recognize original data well but cannot predict unseen data accurately. Finally, 0.8 of Mincriterion was selected for further exploration according to the CV accuracy. The structure of decision tree was plotted in Figure 1 with default mincriterion 0.95.
Figure 2 shows the association between training data size with accuracy of decision tree. The training data percentage is the ratio of training data in implementation with original training data. The change of accuracy for train accuracy and CV accuracy is not significant. However, the CV accuracy increases slowly with increasing training data size. It means more data can enable better decision tree model.
For bank dataset, the same process was applied and because the size of bank dataset is about 10 times of protein dataset, the tree will be larger at the same conditions. The dataset can be predicted easier since all the accuracy here is obviously higher than the protein dataset. This may be caused by the characteristics of two datasets. The relationship between features and target response is stronger for bank dataset. In addition, after pruning, the train accuracy is

reducing and CV accuracy has the largest value in the medium mincriterion. Two small trees cannot learn the relationships fully and too large trees may cause overfitting.
In Figure 3, the curve of train accuracy and CV accuracy becomes closer when there are more training data. It is reasonable since it means the model are gradually learning the model fully. The CV accuracy increases when there is more data for training. This is also the reason we commonly want more data for prediction problem.

Mincriterion	Nodes	Train accuracy	CV accuracy
0.1	249	0.8357532	0.644738791
0.2	189	0.8212341	0.67692102
0.3	147	0.8030853	0.670582065
0.4	141	0.7999093	0.681480872
0.5	121	0.7881125	0.670582065
0.6	97	0.7767695	0.675604689
0.7	59	0.7617967	0.676016043
0.8	45	0.7459165	0.687836281
0.9	39	0.7436479	0.686949815
1	1	0.5494555	0.54951666

Table 1. Decision Tree performance of protein dataset after pruning

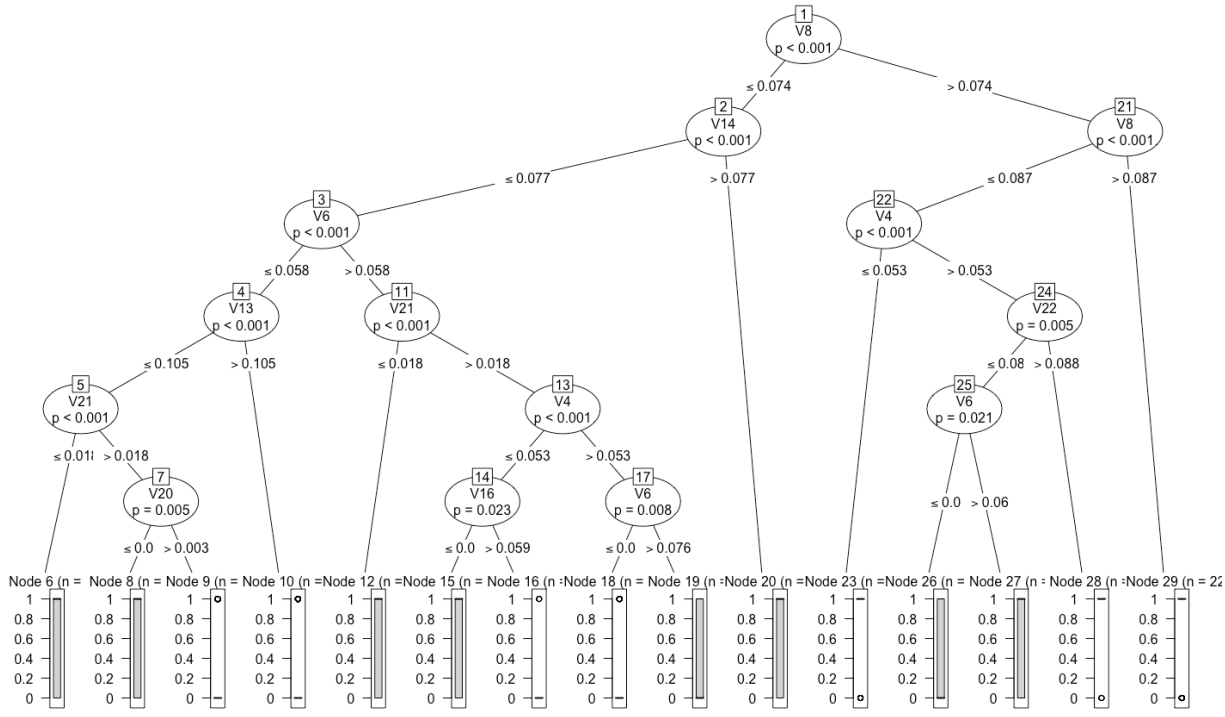

Figure 1. Decision Tree structure of protein dataset

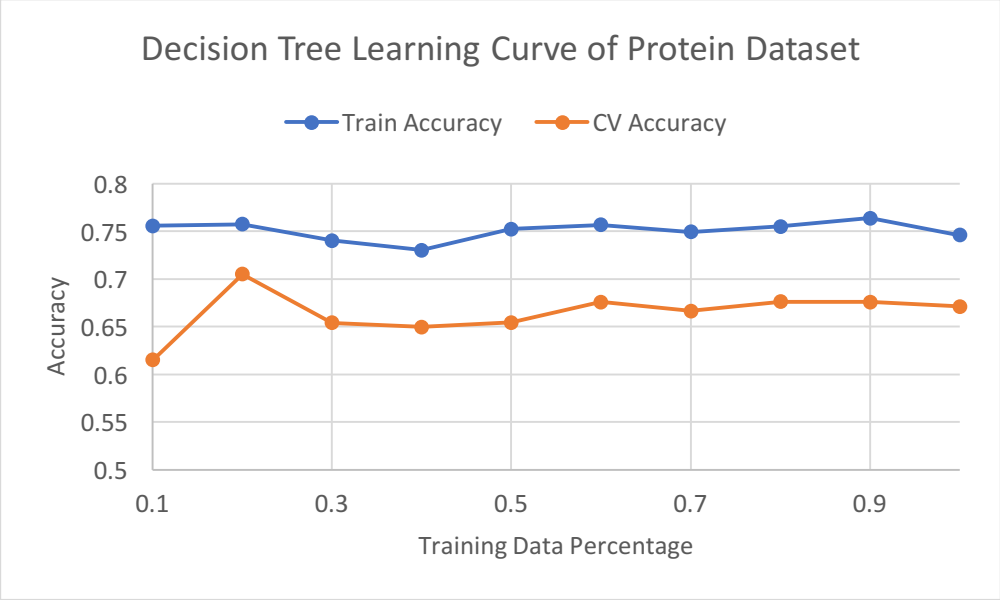

Figure 2. Decision Tree learning curve of protein dataset for different training data percentage

Mincriterion	Nodes	Train accuracy	CV accuracy
0.1	423	0.9279967	0.908482155
0.2	311	0.9249445	0.912596768
0.3	247	0.9236959	0.91208798
0.4	215	0.922771	0.913567877
0.5	197	0.9221698	0.912966881
0.6	173	0.9217074	0.911625808
0.7	153	0.9213836	0.914677874
0.8	123	0.9173603	0.913707321
0.9	109	0.9172216	0.912967458
1	1	0.8897521	0.889752772

Table 2. Decision Tree performance of bank dataset after pruning

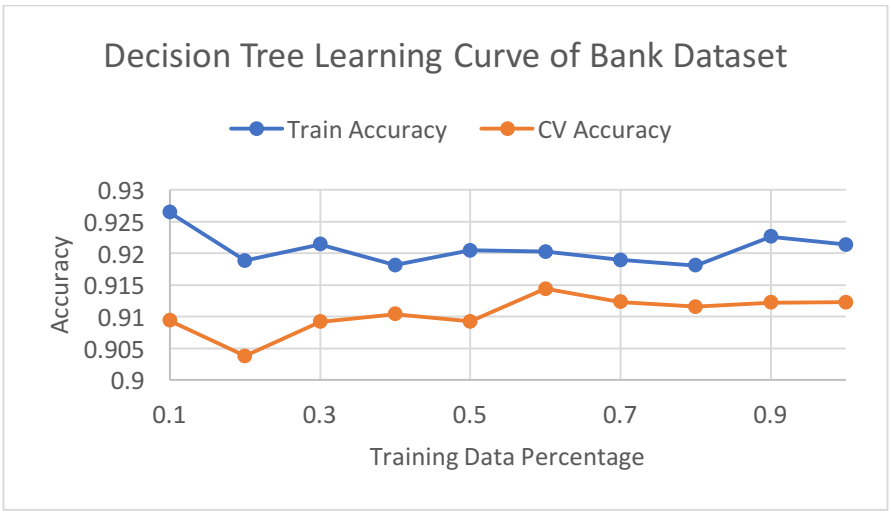

Figure 3. Decision Tree learning curve of bank dataset for different training data percentage

2. Neural networks

The training time of neural networks is long, and especially for CV accuracy. Two parameters of neural networks were explored here. In Table 3, different hidden layers which do not include input and out layers were investigated. When the hidden layers are 0 and there are 20 neurons, the CV accuracy is the best and taken as parameters for next exploration. It shows complex architecture of neural networks is not necessary for some problems and may cause overfitting sometimes. In Figure 5, when we increased the training data size, the accuracy increased for both train accuracy and CV accuracy. More data helps the model understand the data better.
For bank dataset, categorical values were converted into one-hot data and numerical values were normalized into 0-1. Then same parameter tuning process was conducted and 3 hidden layers and 10 neurons were hypeparameters which achieved the best CV accuracy. In Figure 6, it can be observed that overfitting was very significant when the training data percentage was 0.2. After more data was trained, two curves of accuracy became closer.
For Figure 4, model performance of different epochs was recorded in those two figures for to datasets. At epoch 2, the accuracy suddenly increased and then accuracy increased slowly for training data and test data for both datasets.

Hidden Layers	Neurons	Train Accuracy	CV Accuracy
0	10	0.7659	0.7155
	20	0.7882	0.7364
	50	0.7827	0.7314
1	10	0.7977	0.7168
	20	0.8245	0.7227
	50	0.885	0.7114
2	10	0.7923	0.7236
	20	0.8514	0.7186
	50	0.9423	0.6814
3	10	0.8072	0.7154
	20	0.8564	0.7027
	50	0.9414	0.6850

Table 3. Neural networks performance of protein dataset for different hidden layers and neurons

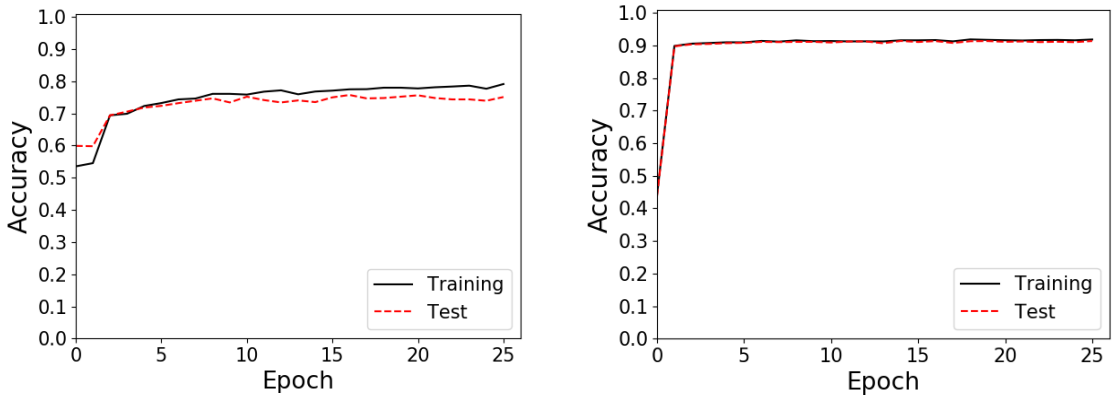

Figure 4. Model performance of protein dataset(left) and bank dataset(right) for different epochs

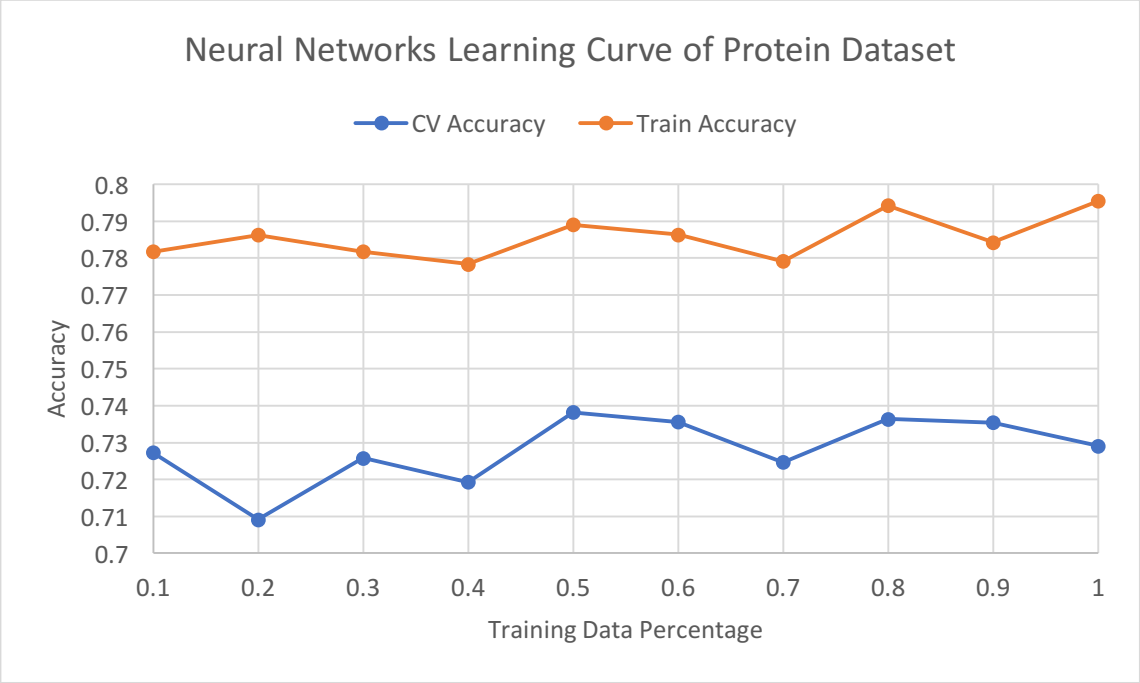

Figure 5. Neural Networks learning curve of protein dataset for different training data percentage

Hidden Layers	Neurons	Train Accuracy	CV Accuracy
0	10	0.9205	0.9105
	20	0.9245	0.9104
	50	0.9275	0.9081
1	10	0.9197	0.9094
	20	0.9268	0.907
	50	0.9429	0.9025
2	10	0.919	0.9104
	20	0.9287	0.9061
	50	0.9521	0.8992
3	10	0.9173	0.9107
	20	0.9256	0.9068
	50	0.9509	0.8999

Table 4. Neural networks performance of bank dataset for different hidden layers and neurons

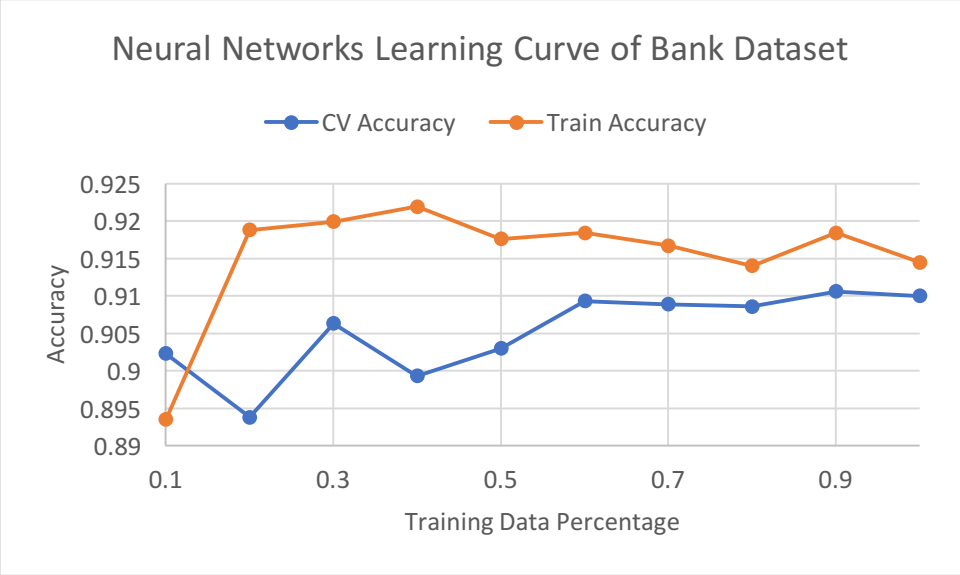

Figure 6. Neural Networks learning curve of bank dataset for different training data percentage

Boosting

Two parameters were tried for boosting, which is an ensemble method by combining several weak learners together. In the code implementing process, the running time for boosting is obvious longer compared with other models. Mfinal means the number of iterations for which boosting is run or the number of trees to use. Maxdepth represents the max depth of trees of this model. As shown in Table 5, pruning was used when the max depth is small. For the same max depth, more iterations or more trees used can improve the performance of training dataset. But CV accuracy achieved the best performance when the max depth was 8 and iterations were 10.

In Figure 7, curves of train accuracy and CV accuracy became closer when the training data size increased. It means there are overfitting when the training data size is very small, which shows high train accuracy and low CV accuracy here. After adding more data, this problem was solved. For bank dataset, the same max depth and iteration were selected according to Table 6. More iteration or more trees improves train accuracy rather than CV accuracy. Complex model is not always necessary for some datasets. For bank dataset, the trend that two curves are closing with increased training data size becomes more obvious in Figure 8. It shows overfitting is more significant for bank dataset when the training data is small.

Maxdepth	Mfinal	Train Accuracy	CV Accuracy
4	5	0.7513612	0.7200555
5	5	0.7749546	0.70645
6	5	0.7885662	0.7027849
8	5	0.7912886	0.7164562
10	5	0.7899274	0.7037392
8	10	0.8203267	0.7286508
8	50	0.9314882	0.7173427
8	100	0.9909256	0.7150494

Table 5. Boosting performance of protein dataset for different iterations

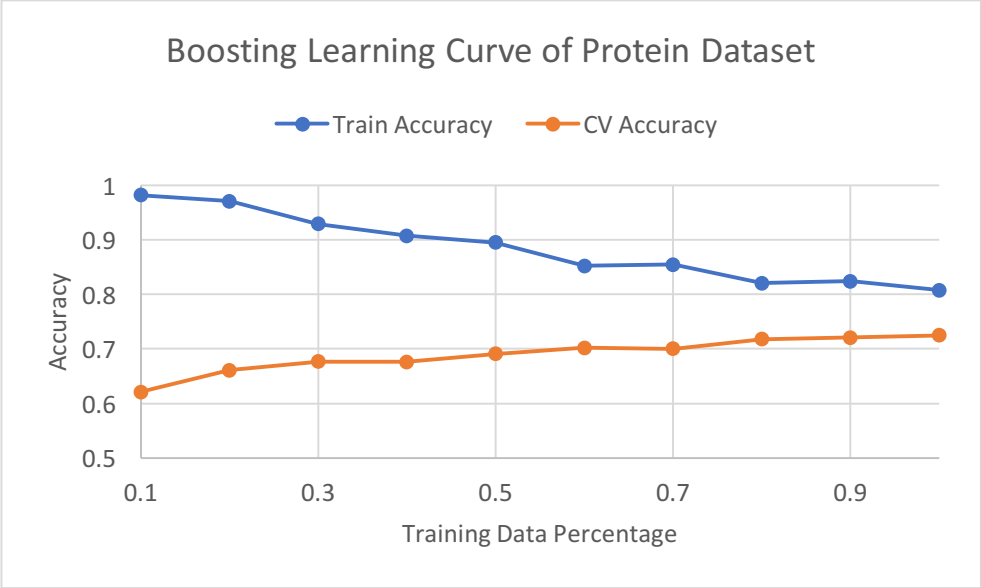

Figure 7. Boosting performance of protein dataset for different training data percentage

Maxdepth	Mfinal	Train Accuracy	CV Accuracy
4	5	0.7513612	0.7200555
5	5	0.7749546	0.70645
6	5	0.7885662	0.7027849
8	5	0.7912886	0.7164562
10	5	0.7899274	0.7037392
8	10	0.8203267	0.7286508
8	50	0.9314882	0.7173427
8	100	0.9909256	0.7150494

Table 6. Boosting performance of bank dataset for different iterations

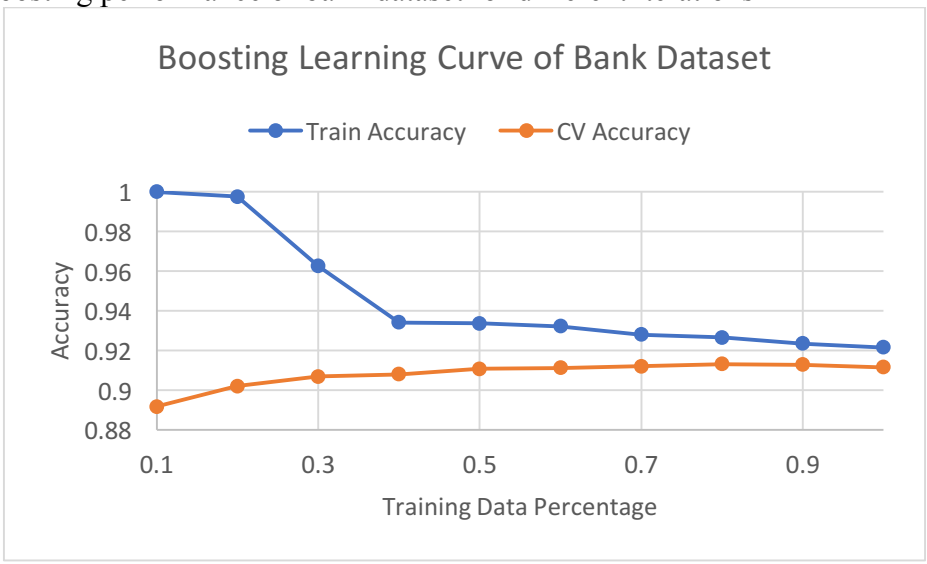

Figure 8. Boosting performance of bank dataset for different training data percentage

Support Vector Machines

SVM is a simple learner and model training is fast. Different kernels of SVM were utilized and corresponding accuracy was recorded in Table 7 and 8 for two datasets. “Radial” is the best kernel for both datasets. In Figure 9, train accuracy decreases and CV accuracy increases when the training data increases and the trend of two curves becoming closer can be seen. For bank dataset, no such trend can be found in Figure 10 and no overfitting exists for this dataset. CV accuracy changes in the same direction with train accuracy.

Kernel	Train Accuracy	CV Accuracy
radial	0.8393829	0.7363739
linear	0.7068966	0.6996154
sigmoid	0.4664247	0.5131201
polynomial	0.7654265	0.647427

Table 7. SVM performance of protein dataset for different kernels

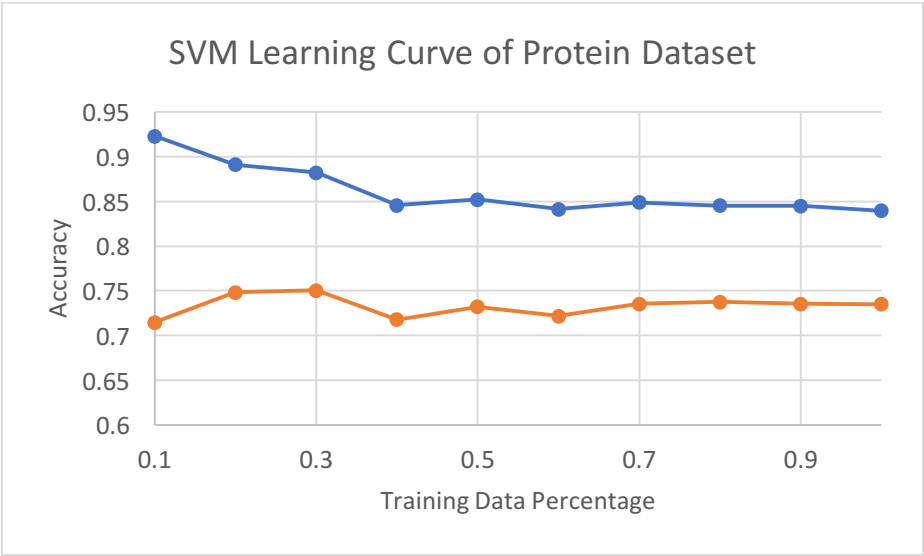

Figure 9. SVM learning curve of protein dataset for different training data percentage

Kernel	Train Accuracy	CV Accuracy
radial	0.9072789	0.9057521
linear	0.8993711	0.8991396
sigmoid	0.8061413	0.810627
polynomial	0.9034406	0.9017752

Table 8. SVM Performance of bank dataset for different kernels

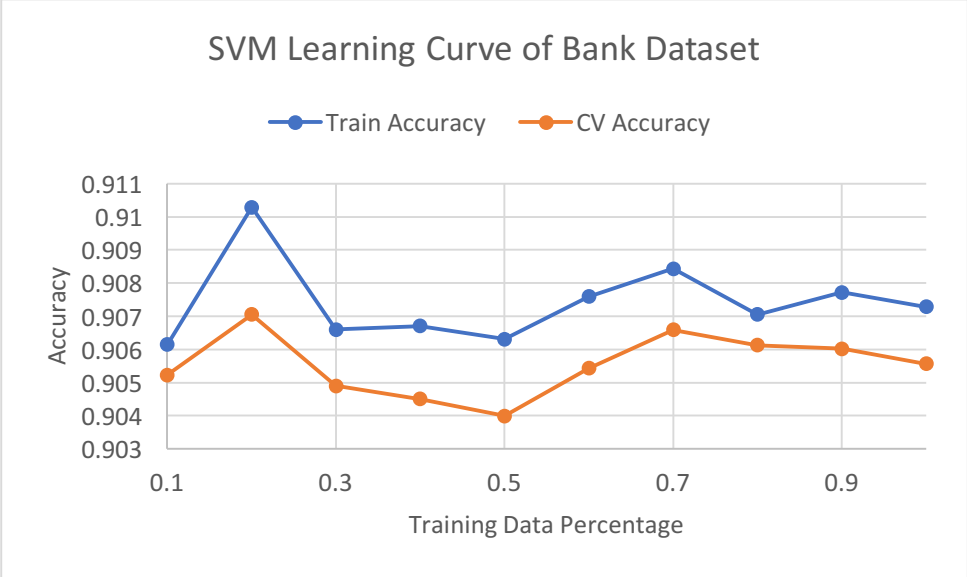

Figure 10. SVM learning curve of bank dataset for different training data percentage

k-nearest neighbors

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It does not need time to train and when new unseen points come in, KNN labels them according to the similarity measure. We tune parameter K to observe model performance. For protein dataset, when 10 nearest neighbours were a cluster, CV accuracy is the best. Small K means this algorithm groups points in details and 1 means each point is a cluster. In that case, the train accuracy will be very high because this algorithm can remember each point. However, for unseen points, small K is not necessary for good performance. In Figure 11, it can be demonstrated that larger training dataset does not improve the model performance. For bank dataset, when K is 50, best accuracy was achieved and it means different dataset has different characteristics and is suitable for different hypeparameters. In Figure 12, larger training dataset improves the performance for both train accuracy and CV accuracy. Those two curves have the same trend in the plot.

K	Train Accuracy	CV Accuracy
100	0.6515426	0.6465405
50	0.6705989	0.6565076
30	0.6873866	0.6601687
10	0.7295826	0.6660736
5	0.7717786	0.6655944
1	0.996824	0.62297

Table 9. KNN Performance of protein dataset for different k

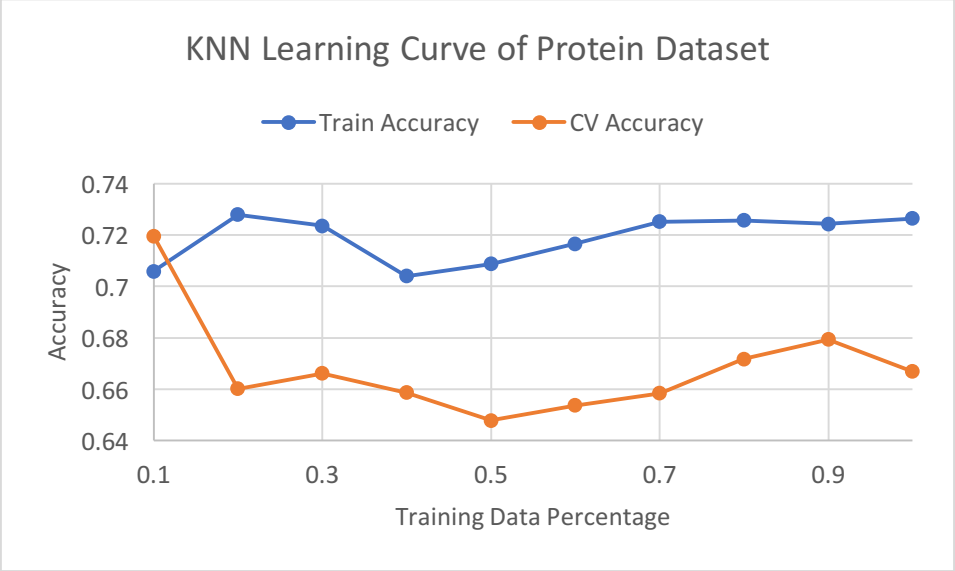

Figure 11. KNN learning curve of protein dataset for different training data percentage

K	Train Accuracy	CV Accuracy
100	0.9132908	0.9117189
50	0.9143082	0.9119495
30	0.916343	0.9113026
10	0.922401	0.9083892
5	0.9304477	0.9033943
1	1	0.8877167

Table 10. KNN Performance of bank dataset for different k

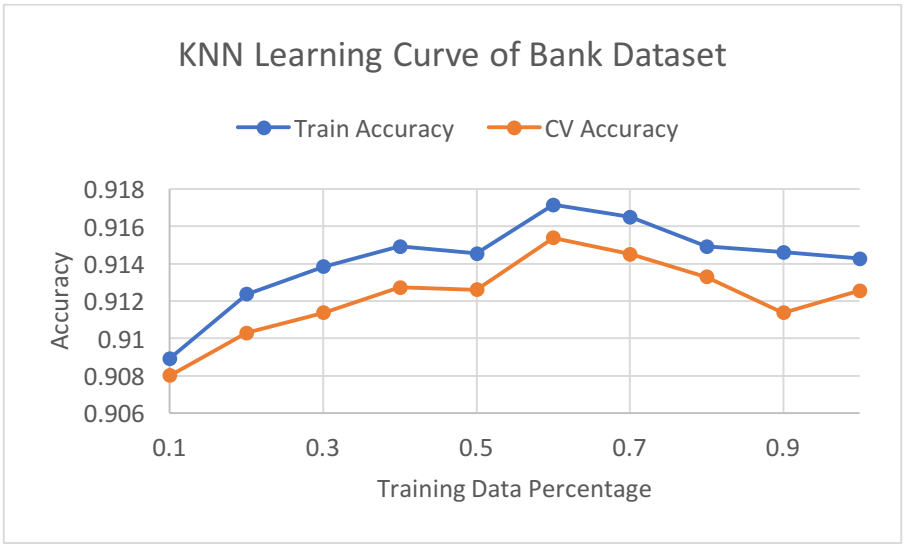

Figure 12. KNN learning curve of bank dataset for different training data percentage

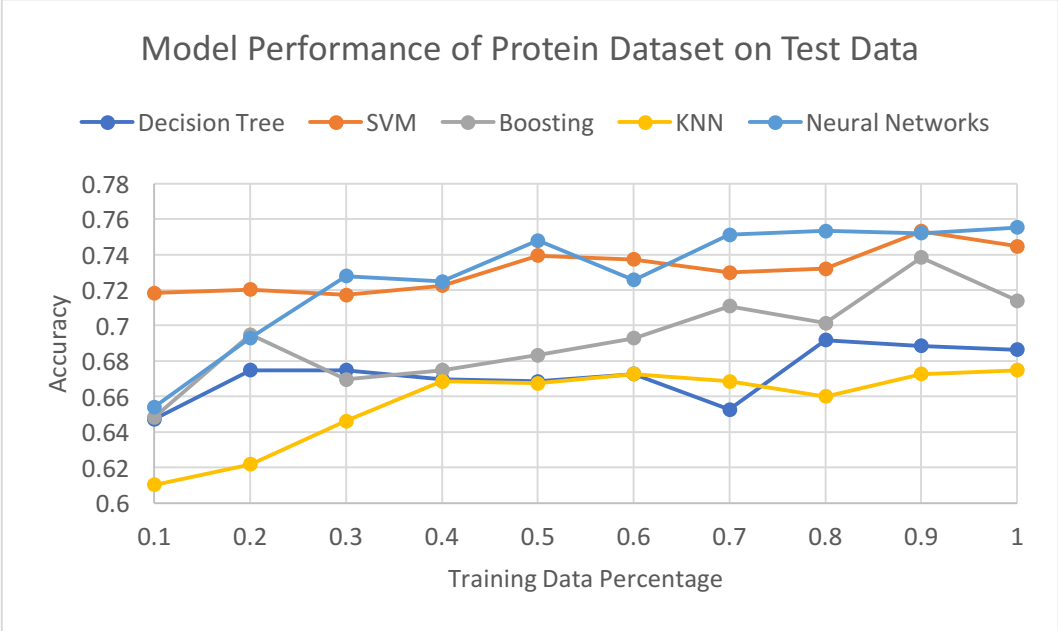

Figure 13. Comparison of performance of 5 models of protein dataset

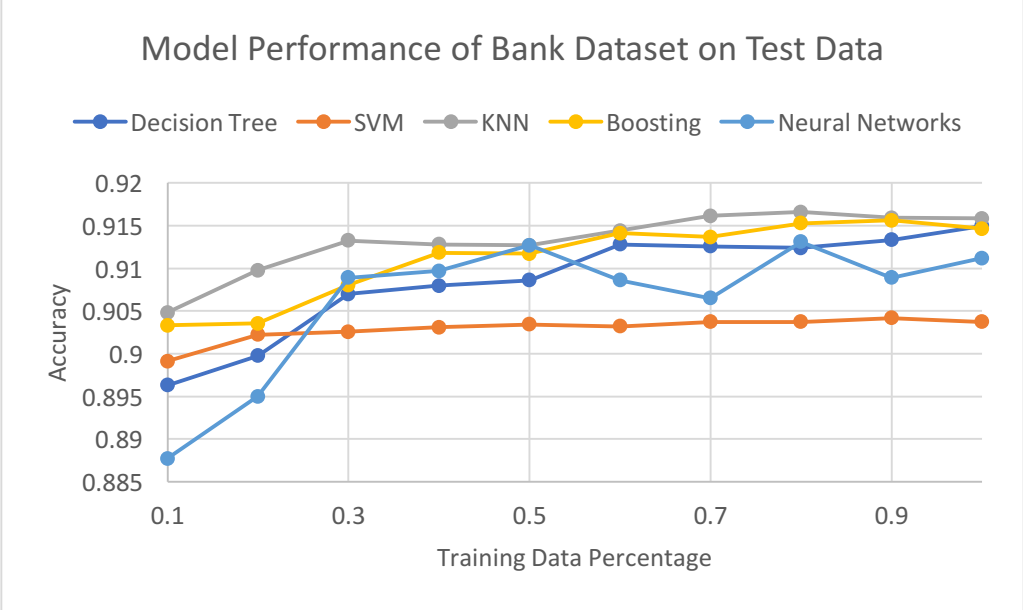

Figure 14. Comparison of performance of 5 models of bank dataset

For protein dataset, different model performance on test data which was held out during tuning process was compared in Figure 13. With increasing training data size, accuracy of each model increases. For all five models, SVM and neural networks perform best in predicting protein solubility from amino acid sequence. When the training data percentage is 1, neural network achieved the best accuracy about 75.5%. For bank dataset in Figure 14, performance of all models improves when the training data size becomes large. And KNN seems has the best performance for predicting whether the client would subscribe the term deposit product from their information. KNN achieved accuracy about 91.6% when training data percentage is 1. For two datasets, bank dataset always has better performance compared with protein dataset. It may be caused by the dataset size or the characteristics of two datasets. Strong relationship between features and target response or larger data size may result in higher performance for all models.