Data pre-processing

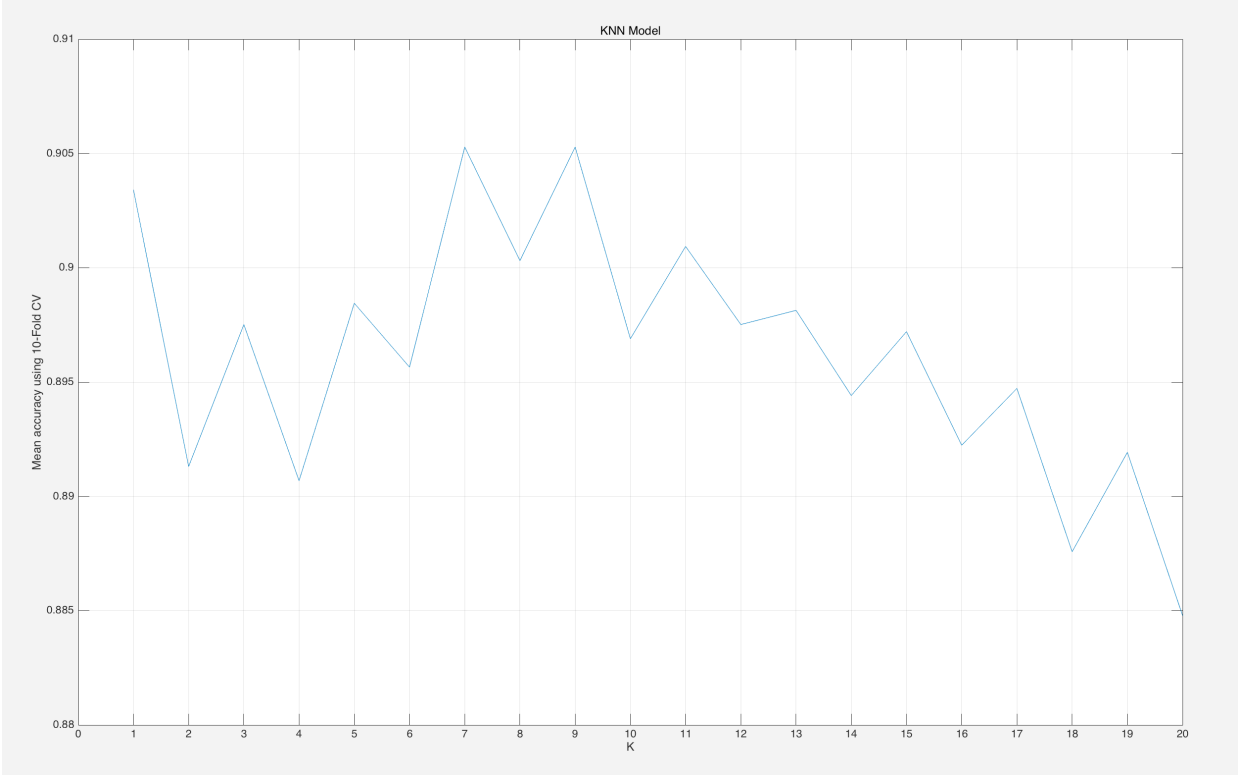
I normalized the data. The reason is most classification algorithms can do better when applied to normalized data. The standardize formula is:

$$\frac{x_i - \text{sample mean}}{\text{sample standard deviation}}$$

KNN Model

KNN model uses the K nearest neighboors to predict the class of unlabeled data.

The result is different when K takes different values. So I tried different K and get the following curve:



We can see when K = 7 or 9, the result is relatively better.

Decision Tree Model

Decision Tree uses InfoGain to build a tree and uses the tree to do the classification.

The accuracy of decision tree is :

0 Fold Train Accuracy:0.999655, Test Accuracy:0.925466 1 Fold Train Accuracy:0.999655, Test Accuracy:0.947205 2 Fold Train Accuracy:1.000000, Test Accuracy:0.925466 3 Fold Train Accuracy:1.000000, Test Accuracy:0.909938 4 Fold Train Accuracy:0.999655, Test Accuracy:0.931677 5 Fold Train Accuracy:0.999655, Test Accuracy:0.934783 6 Fold Train

Accuracy:0.999655, Test Accuracy:0.953416 7 Fold Train Accuracy:0.999655, Test Accuracy:0.903727 8 Fold Train Accuracy:0.999655, Test Accuracy:0.885093 9 Fold Train Accuracy:0.999655, Test Accuracy:0.903727 Decision Tree 10 folds CV with accuracy: 0.922049689441

And the tree is saved in DTree.gv file.

SVM Model

SVM model is based on the biggest gap between classes, and it can generate a hyper plane to separate different classes.

I tried the SVM model and found the result is :

(kernal function is RBF)

0 Fold Train Accuracy:0.948930, Test Accuracy:0.913043 1 Fold Train Accuracy:0.945825, Test Accuracy:0.934783 2 Fold Train Accuracy:0.947205, Test Accuracy:0.928571 3 Fold Train Accuracy:0.945825, Test Accuracy:0.925466 4 Fold Train Accuracy:0.947550, Test Accuracy:0.928571 5 Fold Train Accuracy:0.948930, Test Accuracy:0.928571 6 Fold Train Accuracy:0.947205, Test Accuracy:0.944099 7 Fold Train Accuracy:0.945480, Test Accuracy:0.937888 8 Fold Train Accuracy:0.952381, Test Accuracy:0.897516 9 Fold Train Accuracy:0.946860, Test Accuracy:0.950311 SVM 10 folds CV with accuracy: 0.928881987578

AdaBoost Model

Adaboot method is an ensemble method, it trains many week classifiers in the training phase, and let all the week classifiers vote to get the finial decision in the testing phase.

The weak classifier I use is decision tree classifier. Using 100 weak classifiers and ensamble them together.

The 10 fold cross validation accuracy is :

0 Fold Train Accuracy:0.965493, Test Accuracy:0.947205 1 Fold Train Accuracy:0.963423, Test Accuracy:0.947205 2 Fold Train Accuracy:0.968254, Test Accuracy:0.940994 3 Fold Train Accuracy:0.967909, Test Accuracy:0.950311 4 Fold Train Accuracy:0.967564, Test Accuracy:0.925466 5 Fold Train Accuracy:0.963423, Test Accuracy:0.937888 6 Fold Train Accuracy:0.963768, Test Accuracy:0.959627 7 Fold Train Accuracy:0.967909, Test Accuracy:0.931677 8 Fold Train Accuracy:0.966874, Test Accuracy:0.919255 9 Fold Train Accuracy:0.965493, Test Accuracy:0.934783 AdaBoostClassifier with 100 estimators 10 folds CV with accuracy: 0.939440993789

Summary

Medel	10 Fole Cross Validation Accuracy
KNN	(K=7) 0.905279503106
Decision Tree	0.922049689441
SVM	0.928881987578
Adaboost	0.939440993789

We can see that Adaboost has the highest accuracy because it ensembled lots of weak classifiers together.

KNN has the worsest result because this model is too simple.

The accuracy of Decision Tree and SVM are higher than KNN since these two models are more complex.