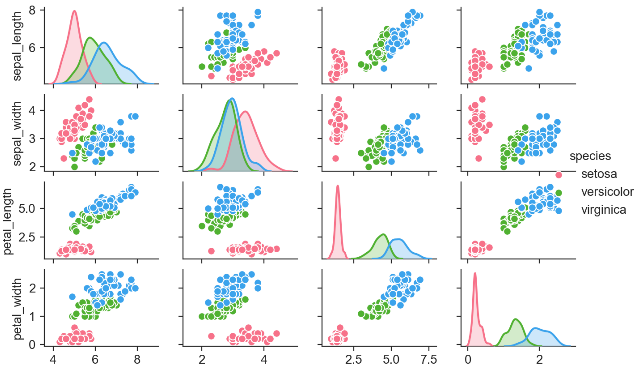
Assignment 1 ECE 657A

He Bing 20848700 b29he@uwaterloo.ca

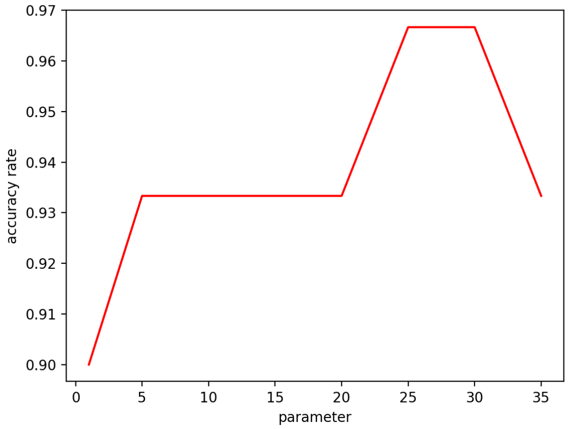
January 2020

1 Question 1



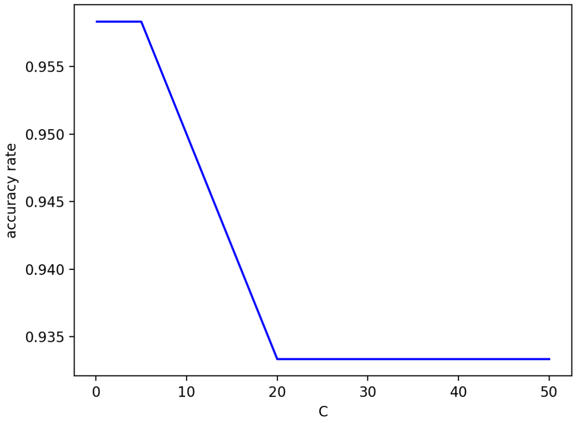
For most of the graphs, setosa can be seen clearly from the other two kinds of iris. In the mean time, the bound of versicolor and virginica is blurry. In the graph of 'petal_length-sepal_length', 'petal_length-sepal_width','petal_width-sepal_length' and 'petal_width-sepal_width', the dots of setosa are under those of versicolor and virginica, virginica is on the top of versicolor.
For the plots of 'sepal_width-setal_length','petal_width-petal_length','petal_length-petal_width','sepal_width-petal_width','sepal_length-petal_length' and 'sepal_length-petal_width', the dots of setosa are on the left of others, and virginica is on the right of versicolor.
As for the graph of 'sepal_width-sepal_length', setosa is on the top left of others. And in 'sepal_length', setosa is on the bottom right.

2 Question 2



The best value of parameter is 25, the accuracy rate is 100%.

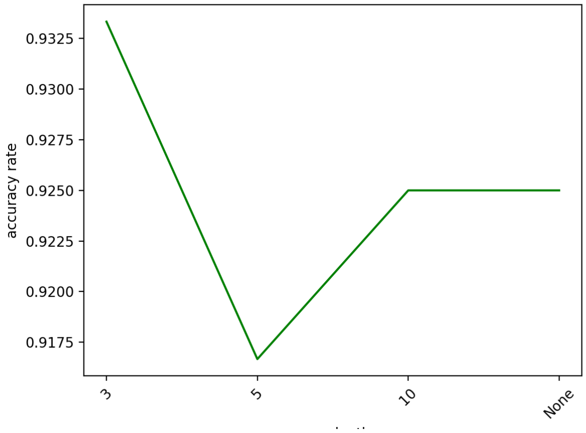
3 Question 3



The best value of c is 0.1, the accuracy rate is 96.67%.

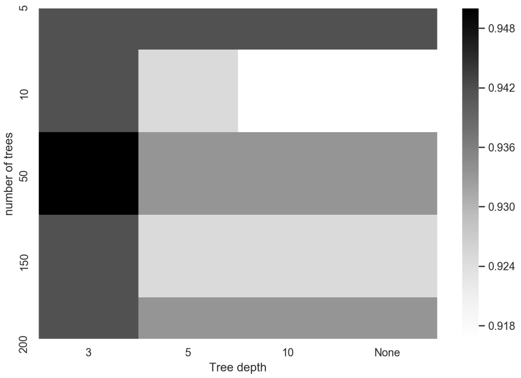
4 Question 4

4.1



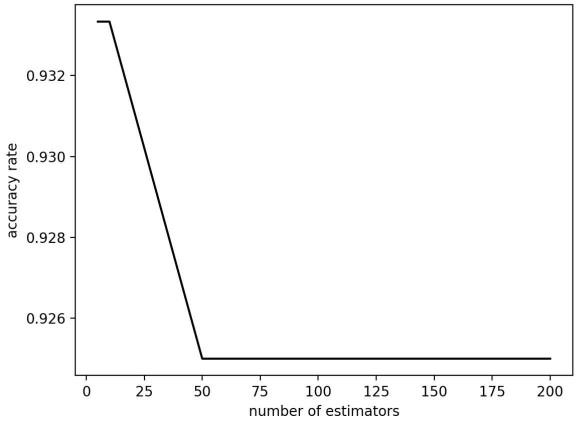
The best value of maximum depth is 3, the accuracy rate is 100%.

4.2



The best number of trees is 50, the best max depth is 3, the accuracy rate is 100%.

4.3



The best number of estimators is 5, the accuracy rate is 100%.

5 Question 5

5.1 Explain why you had to split the dataset into train and test sets?

After the training of dataset, we should know the accuracy so that we can know whether the model can be used in the future predicting. So we should reserve some data as test set to know how accurate exactly is this model.

5.2 Explain why when finding the best parameters for KNN you didn't evaluate directly on the test set and had to use a validation test.

The two sets are for the totally different two purposes. For validation set, is to detect over-fitting and help with best parameter search. For the test set, it is used to measure the performance of the model. If we directly use the test set, it will definitely have an influence on the final result.

5.3 What was the effect of changing k for KNN. Was the accuracy always affected the same way with an increase of k? Why do you think this happened?

Before 25, the accuracy rate grows as k getting greater. But after 30, the accuracy rate goes down as k increasing. So definitely the accuracy is not always affected the same way.
KNN is an algorithm to find the n-nearest neighbors of the case to predict it.

So there will be a best value of k, after the best value, more 'wrong' cases will be included into the neighbor set. So the accuracy will decrease as k goes up after the best k.

5.4 What was the relative effect of changing the max depths for decision tree, random forests, and gradient boosting? Explain the reason for this.

Generally speaking, the larger the depth, the better the fitting effect, but at the same time, it will increase the computational complexity, slow down the calculation speed, and over-fitting will also occur, which would lead to the decrease of accuracy. In general, this value can be ignored when there is little value of data or features. If the sample size and features are large, this value should be limited. The specific value depends on the distribution of the data.

5.5 Comment on the effect of the number of estimators for Gradient Tree Boosting and what was the relative effect performance of gradient boosting compared with random forest. Explain the reason for this.

Generally speaking, as the number of estimators grows, the higher accuracy rate is. But it can be seen that when the number of estimators exceeds a certain value, the error rate of the model converges. When the number of trees increases, the effect will not improve, but it will only slow down the learning process.

5.6 What does the parameter C define in the SVM classifier? What effect did you observe and why do you think this happened?

C is a regularization parameter that controls the trade off between the achieving a low training error and a low testing error that is the ability to generalize your classifier to unseen data.
C is a parameter which to make a balance between training error and testing error, so that the model can be widely used into predicting. In other words, c defines a domain that how much you want to avoid misclassifying each training example. The accuracy drop sharply after 5, then it becomes stable after 20.
If C is too small, it may lead to a result of underfitting. if C is too large, the model will try to classify each training example correctly.Doing this will lead to loss in generalization properties of the classifier.