

514A Programming 2 Report

Name: Zhuoyun Chen, Xuewei Xiao

ID: 499214, 500641

1. Introduction

(a) Description of the problem and the practical impacts of solving it

In this session, we need to handle the binary classification problem to realize the letter recognition based on the given 16 primitive numerical attributes. The values of the attributes are scaled to fit into a range of integer values from 0 through 15. There are in total 20000 samples. Faced with this problem, the factors to be considered in determining a classifier as the “best”. From my point of view, a better classification model should be higher validation test accuracy with relatively low computational complexity, along with suitable model interpretability.

(b) Motivation for multiple classifiers

Since the dataset is though not large in this problem, but the task of letter recognition classification may be hard and time consuming when the size of dataset gets larger, so the computational complexity should be considered. However, there are some letters that are similar to each other that may be confusing to figure out the actual letter, so the validation accuracy is also an important factor. Finally, since the problem may be customized to some specific needs, and when introducing to someone not in the data science or related field, the model should have a suitable interpretability.

(c) Motivation for dimension reduction

I think a good dimensionality reduction method should keep more information of the original data while reducing the dimensionality as much as possible. Dimensionality reduction will cause the loss of original data information, and the data after dimensionality reduction should retain as much information as possible from the original data. Hence the information is an important factor to consider.

(d) Brief description of the dimension reduction method(s) you chose

I used PCA method to reduce the dimension. PCA is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Dimensionality reduction results in a loss of information, in general. PCA-based dimensionality reduction tends to minimize that information loss, under certain signal and noise models.

(e) Speculate on the binary classification problems. Which pair of letters did

you choose for the third problem? Which pair do you predict will be the easiest or hardest to classify?

Binary classification is one of the most common problems in machine learning and it refers to the model of predicting an input sample into one of two classes. I finally choose letter “F” and “Z” as the third pair. From observation, I think the first pair: K and H will be the hardest to classify.

2. Results

2.1 Brief description of the classifier and its general advantages and disadvantages and graphs of cv scores (without dimension reduction)

1) K- nearest Neighbors

KNN algorithm a non-parametric supervised learning method, which makes the classification based on the labels of the k nearest neighbors to the sample and here k is a hyperparameter.

Advantages are:

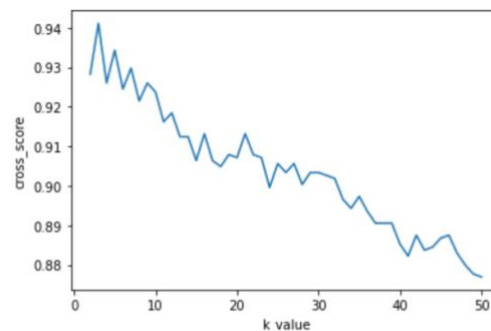
- 1) Easy to realize and to understand.
- 2) Easy to modify the model by changing the value of k.
- 3) One can add some weights to k corresponding to different values of distance.

Disadvantages are:

- 1) The classification results may be influenced by noise data and outliers.
- 2) When handling the unbalanced dataset, the prediction result is inclined to be the one with more samples.
- 3) Time and space consuming since for each sample, the distances from all the rest samples to it need to be calculated to find the k nearest neighbors.

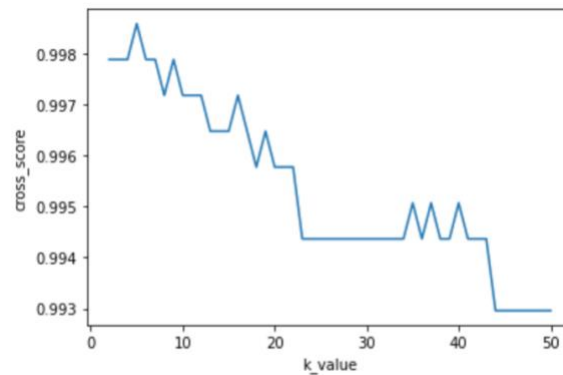
Below is the graph of the cross validation results over the range of hyperparameter k (without dimension reduction):

Pair1 (H and K):



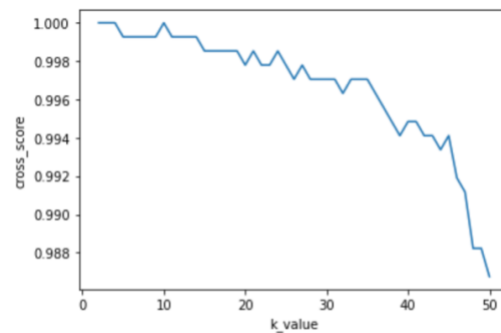
The best value of k is: 3
Running time is: 0:00:02.623409

Pair2 (M and Y):



The best value of k is: 5
Running time is: 0:00:02.775636

Pair3 (F and Z):



The best value of k is: 2
Running time is: 0:00:02.616591

2) Decision Tree

Decision tree is an effective tool that can help decision makers to conduct sequence decision analysis. The method is to express the relevant strategy, natural state, probability, and profit value in the problem through lines and graphs in a form similar to a tree.

Advantages are:

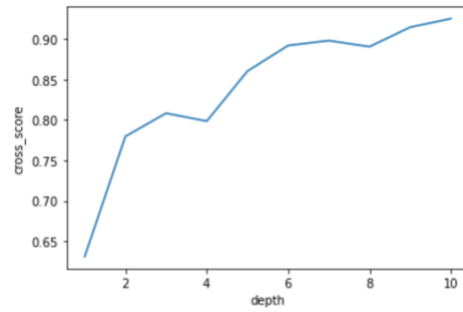
- 1) Shallow decision trees are visually very intuitive and easy to interpret.
- 2) There is no need to make any assumptions about the structure and distribution of the data.
- 3) The decision tree can capture the interaction between variables.

Disadvantages are:

- 1) Deep decision trees are visually and interpretatively difficult.
- 2) The decision tree has a relatively large demand for sample size.
- 3) Its function of dealing with missing values is very limited.

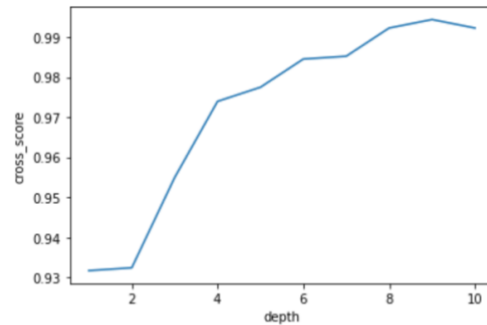
Below is the graph of the cross validation results over the range of the hyperparameter *max_depth* (without dimension reduction):

Pair1 (H and K):



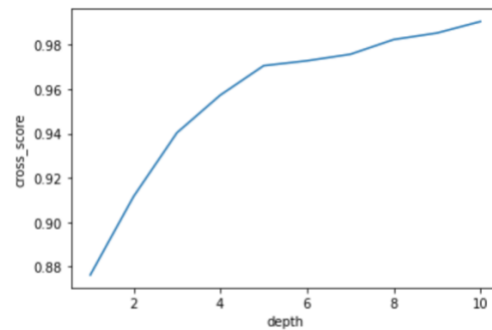
The best value of depth is: 10
Running time is: 0:00:00.152343

Pair2 (M and Y):



The best value of depth is: 9
Running time is: 0:00:00.149402

Pair3 (F and Z):



The best value of depth is: 10
Running time is: 0:00:00.140282

3) Random Forest

A random forest is a classifier that contains multiple decision trees, and its output classes are determined by the mode of the classes output by the individual trees.

Advantages are:

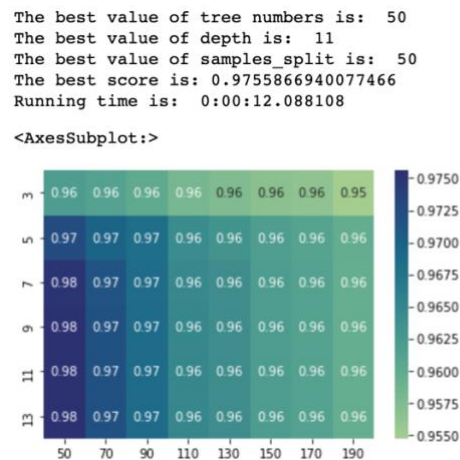
- 1) For many kinds of data, it can produce high accuracy classifiers.
- 2) It can handle many input variables.
- 3) It can evaluate the importance of variables when deciding the category.
- 4) When building a forest, it can internally generate unbiased estimates of the generalized error.
- 5) For imbalanced classification datasets, it can balance the error.

Disadvantages are:

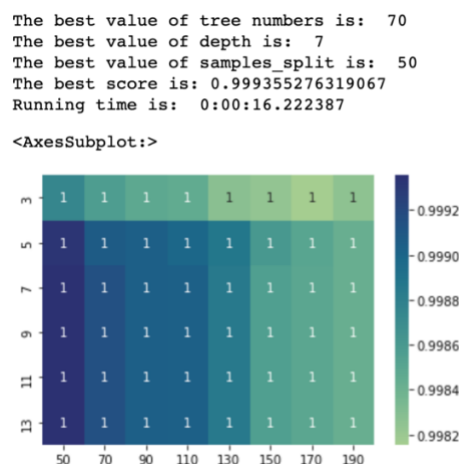
- 1) When the number of decision trees in the random forest is large, the space and time required for training will be very large.
- 2) It may be easier to cause overfitting on the training dataset.

Below is the graph of the cross validation results over the range of the hyperparameter $n_estimators$, $depth$ and $sample_splits$ (without dimension reduction):

Pair1 (H and K):



Pair2 (M and Y):



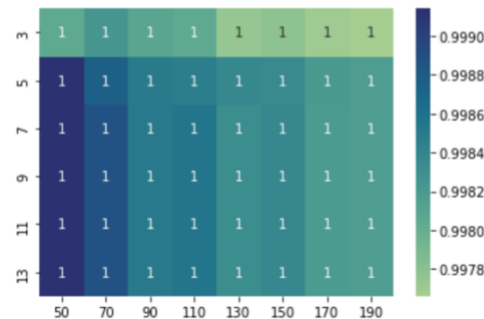
Pair3 (F and Z):

```

The best value of tree numbers is: 70
The best value of depth is: 5
The best value of samples_split is: 50
The best score is: 0.9991404065371435
Running time is: 0:00:15.954283

```

<AxesSubplot:>



4) SVM

Support Vector Machine (SVM) is a kind of generalized linear classifier that performs binary classification on data according to supervised learning, and its decision boundary is the maximum margin for solving the learning sample.

Advantages are:

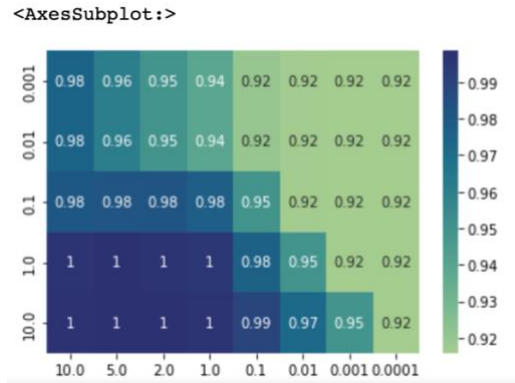
- 1) Machine learning in the case of small samples is addressed.
- 2) Since the problem of the curse of dimensionality and nonlinear separability is overcome by using the kernel function method, the computational complexity is not increased when mapping to a high-dimensional space. (Since the final decision function of the support vector machine algorithm is only determined by a small number of support vectors, the computational complexity depends on the number of support vectors, not the dimension of the entire sample space)

Disadvantages are:

- 1) The support vector machine algorithm is difficult to implement for large-scale training samples, because the support vector algorithm uses quadratic programming to solve the support vector, which will design the calculation of the m-order matrix, so when the matrix order is large, it will consume a lot of machine memory and operation time.
- 2) The classic SVM only gives the algorithm of two classifications, and in data mining, it is generally necessary to solve the classification problem of multi-classification, and the support vector machine is not ideal for solving the multi-classification problem.

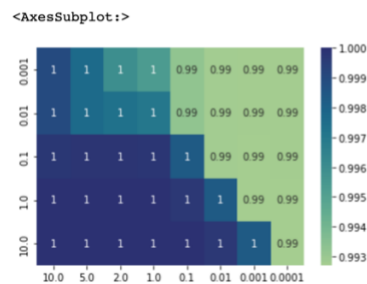
Below is the graph of the cross validation results over the range of the hyperparameter kernel(without dimension reduction), we choose kernel function 'rbf' and tune parameter the penalty value C and gamma:

Pair1 (H and K):



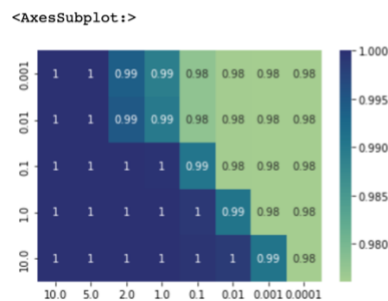
Pair2 (M and Y):

Fitting 5 folds for each of 40 candidates, totalling 200 fits
The best value of C is: 1
The best value of gamma is: 0.1
The best score is: 1.0
Running time is: 0:00:09.902449



Pair3 (F and Z):

Fitting 5 folds for each of 40 candidates, totalling 200 fits
The best value of C is: 0.001
The best value of gamma is: 0.1
The best score is: 1.0
Running time is: 0:00:09.566362



5) ANN

Artificial neural network is an operational model that consists of interconnected connections between a large number of nodes (or neurons). Each node represents a specific output function, called the activation function. The connection between each two nodes represents a weighted value for the signal passing through the connection, called the weight, which is equivalent to the memory of the artificial neural network. The output of the network varies

according to the connection method of the network, the weight value and the excitation function.

Advantages are:

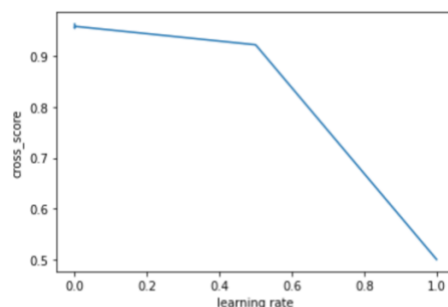
- 1) It has self-learning function. For example, when realizing image recognition, only many different image templates and corresponding results to be recognized are input into the artificial neural network, and the network will gradually learn to recognize similar images through the self-learning function. The self-learning function is particularly important for prediction.
- 2) It can find optimal solutions at high speed. Finding an optimal solution to a complex problem often requires a large amount of computation. Using a feedback artificial neural network designed for a certain problem and using the high-speed computing power of the computer, it is possible to quickly find the optimal solution.

Disadvantages are:

- 1) The most serious problem is the inability to explain their own reasoning process and reasoning basis.
- 2) The necessary query cannot be asked of the user, and when the data is insufficient, the neural network cannot work.
- 3) Since it changes the characteristics of all problems into numbers and turning all reasoning into numerical calculations will inevitably result in loss of information.

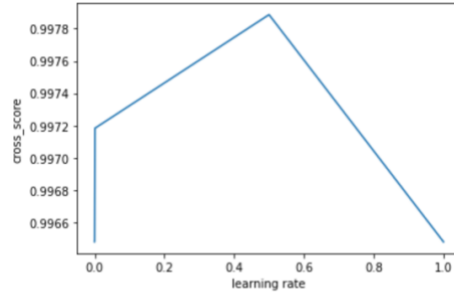
Below is the graph of the cross validation results over the range of the hyperparameter learning rate alpha (without dimension reduction), we choose the value of from the list [1.0,0.5,0.001,1e-5,1e-8]:

Pair1 (H and K):



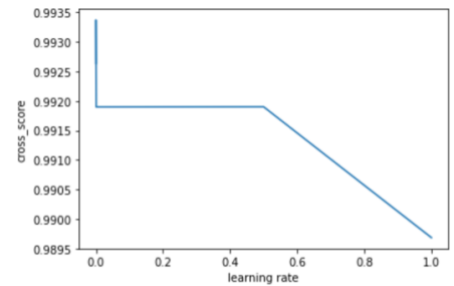
The best learning rate is: 1e-08
Running time is: 0:01:01.545736

Pair2 (M and Y):



The best learning rate is: 0.5
Running time is: 0:01:29.529569

Pair3 (F and Z):



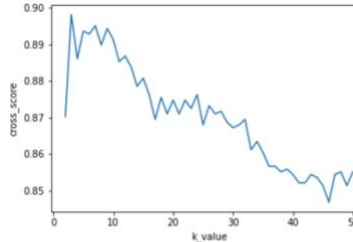
The best learning rate is: 1e-05
Running time is: 0:01:14.656002

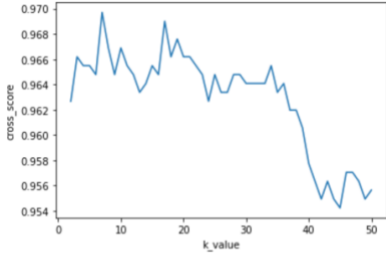
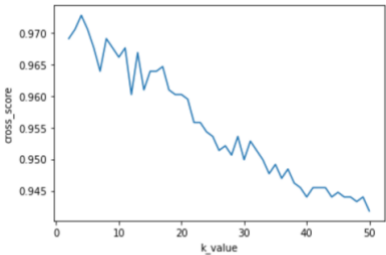
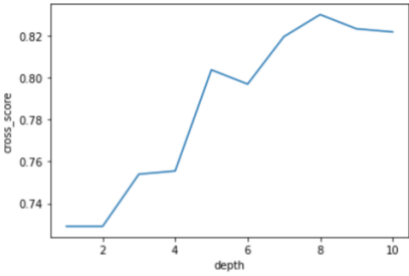
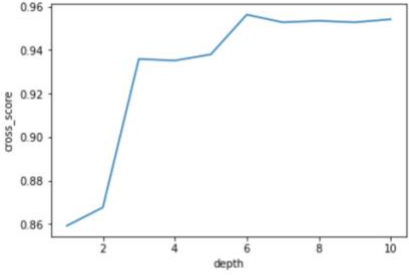
2.2 Dimension Reduction

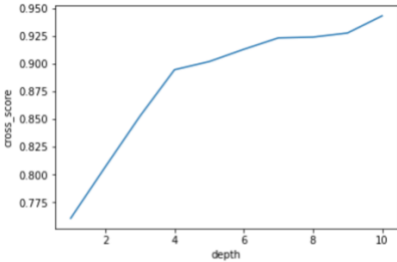
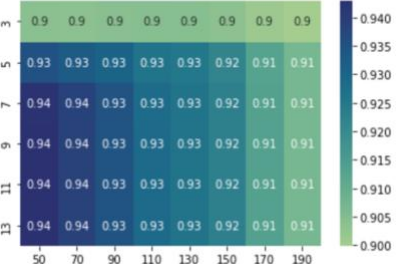
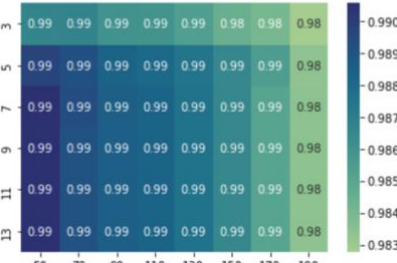
In this problem, we then apply PCA method to realize the dimension reduction.

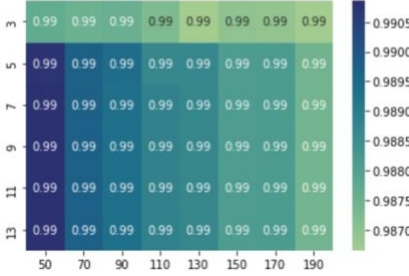
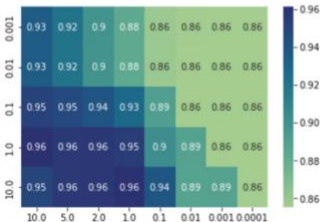
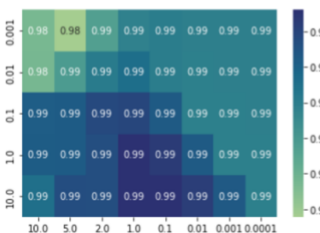
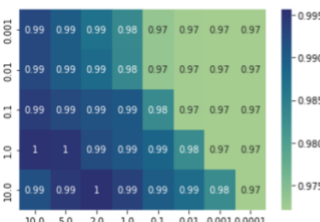
PCA Method: Principal Component Analysis PCA is a technique for simplifying datasets. It is a linear transformation. This transformation transforms the data into a new coordinate system such that the first largest variance of any data projection is in the first coordinate (called the first principal component), and the second largest variance is in the second coordinate (the second principal component). ingredients), and so on.

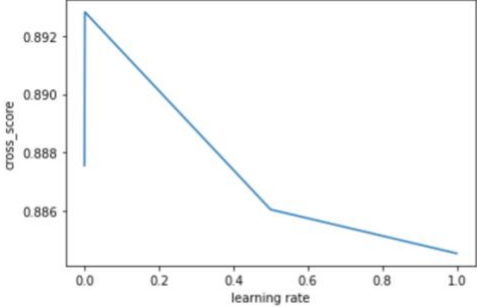
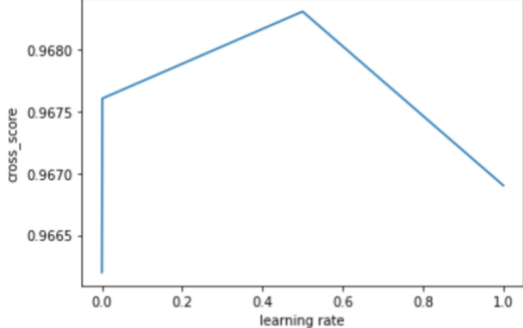
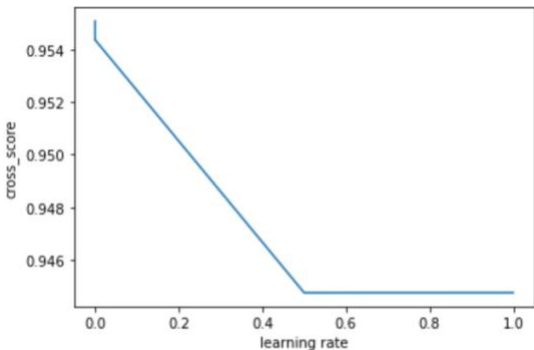
Results of classification methods after dimension reduction.

Pair	Method	Running Time(s)	Graph	Accuracy
1	KNN	1.7301	<div><p>The best value of k is: 3 Running time is: 0:00:01.730136</p></div>	0.9054

2	KNN	1.8358	 <p>The best value of k is: 7 Running time is: 0:00:01.835780</p>	0.9937
3	KNN	1.7496	 <p>The best value of k is: 4 Running time is: 0:00:01.749571</p>	0.9669
1	Decision Tree	0.1630	 <p>The best value of depth is: 8 Running time is: 0:00:00.163037</p>	0.9527
2	Decision Tree	0.1600	 <p>The best value of depth is: 6 Running time is: 0:00:00.160064</p>	0.9937

3	Decision Tree	0.1603	 <p>The best value of depth is: 10 Running time is: 0:00:00.160279</p>	1.0
1	Random Forest	14.4448	<p>The best value of tree numbers is: 70 The best value of depth is: 11 The best value of samples_split is: 50 The best score is: 0.9428343586238324 Running time is: 0:00:14.444788</p> <p><AxesSubplot:></p> 	0.9122
2	Random Forest	7.1044	<p>The best value of tree numbers is: 20 The best value of depth is: 9 The best value of samples_split is: 50 The best score is: 0.9905572780608131 Running time is: 0:00:07.104425</p> <p><AxesSubplot:></p> 	0.9177

3	Random Forest	18.8940	<p>The best value of tree numbers is: 70 The best value of depth is: 7 The best value of samples_split is: 50 The best score is: 0.9908564812092792 Running time is: 0:00:18.894003</p> <p><AxesSubplot:></p> 	0.9603
1	SVM	10.8695	<p>Fitting 5 folds for each of 40 candidates, totalling 200 fits The best value of C is: 1 The best value of gamma is: 10 The best score is: 0.9610845295055821 Running time is: 0:00:10.869489</p> <p><AxesSubplot:></p> 	0.9459
2	SVM	8.2809	<p>Fitting 5 folds for each of 40 candidates, totalling 200 fits The best value of C is: 10 The best value of gamma is: 0.1 The best score is: 0.99319564789182 Running time is: 0:00:08.280945</p> <p><AxesSubplot:></p> 	0.9873
3	SVM	7.8270	<p>Fitting 5 folds for each of 40 candidates, totalling 200 fits The best value of C is: 10 The best value of gamma is: 2 The best score is: 0.9955934565848465 Running time is: 0:00:07.826988</p> <p><AxesSubplot:></p> 	0.9934

1	ANN	49.9765	 <p>The best learning rate is: 0.001 Running time is: 0:00:49.976485</p>	0.9392
2	ANN	49.0320	 <p>The best learning rate is: 0.5 Running time is: 0:00:49.032000</p>	0.9937
3	ANN	60.7895	 <p>The best learning rate is: 1e-08 Running time is: 0:01:00.789454</p>	0.9934

3. Results

3.1 Performance and runtime of different classifiers

3.1.1 Performance table without dimension reduction:

Pair	Method	Running Time(s)	Accuracy
1	KNN	2.6234	0.9392
2		2.7756	1.0
3		2.6166	0.9813

1	Decision Tree	0.1523	1.0
2		0.1494	1.0
3		0.1402	1.0
1	Random Forest	12.0881	0.8649
2		16.2224	0.9984
3		15.9543	0.9536
1	SVM	11.7055	1.0
2		9.9024	1.0
3		9.5664	0.5430
1	ANN	61.5457	0.5135
2		89.5396	0.4747
3		74.6560	0.4570

Conclusion:

When no dimension reduction is implemented, we can see that decision tree can provide a very good value with less time cost in all pairs. Meanwhile, SVM and ANN are time-costing and with less accuracy.

3.1.2 Performance table with dimension reduction:

Pair	Method	Running Time(s)	Accuracy
1	KNN	1.7301	0.9054
2		1.8358	0.9937
3		1.7496	0.9669
1	Decision Tree	0.1630	0.9527
2		0.1600	0.9937
3		0.1603	1.0
1	Random Forest	14.4448	0.9122
2		7.1044	0.9177
3		18.8940	0.9603
1	SVM	10.8695	0.9459
2		8.2809	0.9873
3		7.8270	0.9934
1	ANN	49.9765	0.9392
2		49.0320	0.9937
3		60.7895	0.9934

Conclusion:

When dimension reduction is implemented, we can see that for all methods are less time-costed but with high accuracy. In this scenario, decision tree can still provide a very good value with less time cost in all pairs. However, SVM and ANN are more accurate.

3.2 Lessons Learned

(a)

I will choose decision tree as the classification model since they consume the least time and have the good performance, even after the dimension reduction, the average accuracy is above 0.90.

(b)

After dimension reduction, I think both my runtime and accuracy gets better. It is reasonable because we reduce the features to 4 after PCA. It is more focused with less noise.

(c)

Giving the same task with a new dataset, I may try oversampling to give more data to train the model and I will choose some better way to do the dimension reduction. Hence, the accuracy may increase.