Project Report

<Student Intervention>

Jihun Mun Brian

1. Classification V.S. Regression. Which one is more proper to apply in this project?

In this supervised machine learning project, I will use Classifier algorithm. Our data set has two parts, features part and result part. Our result set can be classified into 2 groups, passed or not passed. Also given features set with unknown results, we will predict result passed or not. So It is classification problem.

2. Training and Evaluating Models

Our data set is high dimensions. It has 30 features. And also the number of data set is not that big. So I choose 3 models as following.

- KNearestNeighbors Algorithms
- Decision Tree Algorithms
- SVM - SVC Algorithms

From now on, I will analyze the pros and cons about these 3 models.

1) KNearestNeighbors Algorithm

PROS	CONS
No training involved : We don't need to train our model. New example can be added anytime without time cost!!	Instead, querying time of the predict data is expensive and slow. => O(#examples * #dimensions)
Very powerful, and complexity parameter is simple The only thing what we have to do is decide n_neighbors.	

To sum up, low train cost (O(1)) and expensive run time (O(#examples * #dimensions)). But the result would be not bad. And number of our example is not large, so I choose this model.

2) Decision Tree Algorithm

PROS	CONS
Querying time(Running time) is very fast.	Training time and complexity can be expensive
Normally Decision tree is very interpretable. Other algorithm is black box algorithm, but this is white box algorithm.	It is easy to overfitting. So you should prune when you make decision tree algorithm.

DecisionTree algorithm is very interpretable and once it is trained, the running time would be very fast. But you should be careful about managing complexity(overfitting problem).

3) SVM-SVC Algorithm

PROS	CONS
It is very effective when data set has high dimensional spaces. Also memory effective.	Training time is very long when the data set is large. So normally it it very ineffective with large number of training sets.
It works really well when there are a clear margin space between classification.	It is very weak at noise and overlapping.

SVM-SVC algorithm is very expensive algorithm(training time is very long). But in our case, the data set has high dimensionality and it contains small number examples. So this algorithm is very suitable for our situation.

3. Choosing The Best Model

At this section I will choose the best model among the above 3 algorithms. To do this Let’s compare the three tables below.

KNearestNeighbors Algorithm	Training Set Size		
	100	200	300
Training Time(sec)	0.001	0.001	0.001
Prediction Time(sec)	0.001	0.003	0.006
F1 score for training set	0.794117647059	0.810035842294	0.840262582057
F1 score for test set	0.762589928058	0.791366906475	0.810810810811

DecisionTree Algorithm	Training Set Size		
	100	200	300
Training Time(sec)	0	0	0.001
Prediction Time(sec)	0	0	0
F1 score for training set	0.753846153846	0.779783393502	0.809195402299
F1 score for test set	0.677966101695	0.816901408451	0.816901408451

SVM-SVC Algorithm	Training Set Size		
	100	200	300
Training Time(sec)	0.001	0.003	0.006
Prediction Time(sec)	0.001	0.002	0.005
F1 score for training set	0.847222222222	0.840277777778	0.84188034188
F1 score for test set	0.785714285714	0.814285714286	0.823529411765

Their performance result are all similar. KNN and SVM-SVC algorithms' performance seems to be slightly better than decision tree algorithm. I strongly believe that Support Vector Classifier is the most suitable model for our data set. It shows the best F1 score. And also as I told you, It is very fit on small number of data set with high dimensionality.

Now let's talk about SVM-SVC algorithm more detail in laymen terms. Basically this algorithm classify the data sets. Suppose there are give data sets which are already classified into several groups. This algorithm's purpose is classify new data to previous some groups. This process can be separate to following two steps.

[STEP 1] Training : Based on training set, this algorithm will calculate a boundary line which is called hyperplane. So the hyperplane is the border line between classified groups. For simple example, group1 is above hyperplane and group2 is below hyperplane.
In other word, the algorithm's purpose is finding the optimal hyperplane of given labeled data sets.

[STEP 2] Predicting : Based on trained model, this algorithm classified new data sets into previous groups. So new data will be classified same as the group of point which is belongs to same area divided by hyperplane.