Chinonso Maduakolam

Project 2: Student intervention system

1. **classification** - Because the students are separated into two bucket. student who will graduate and those who are at risk of not graduating. So you are mapping from an input space to a strict set of choices unlike regression where you have mapping from an input space to a real number out of an infinite set of real numbers.

SVC	Training Set Size		
	100	200	300
Training time (secs)	0.001	0.002	0.003
Prediction time (secs)	0.000	0.001	0.003
F1 score for training set	0.833	0.842	0.847
F1 score for test set	0.825	0.846	0.855

Decision Tree	Training Set Size		
	100	200	300
Training time (secs)	0.000	0.001	0.001
Prediction time (secs)	0.000	0.000	0.001
F1 score for training set	1.0	1.0	1.0
F1 score for test set	0.692	0.698	0.785

KNN	Training Set Size		
	100	200	300
Training time (secs)	0.000	0.001	0.000
Prediction time (secs)	0.001	0.001	0.004
F1 score for training set	0.850	0.785	0.855
F1 score for test set	0.838	0.811	0.814

What are the general applications of this model ?

The general application of these models(SVC,Decision Tree, KNN) are for supervised learning which can be split into two types one being classification and the other regression. In this specific application we are looking at a set of students and we are analyzing features pertaining to each of these students and predicting whether or not said student will graduate. Hence this is a binary classification of data a clear decision boundary between two possible outcomes.

Strengths of SVC ?

1. SVM maximizes margin, so the model is slightly more robust .
2. SVM support the use of kernel trick, so you can model even non-linear relations .
3. Not trapped in local minima.
4. Works well with smaller training sample (number of support vectors do not matter much).

Weakness of SVC ?

1. Very long training time.
2. Picking/finding the right kernel is a challenge.
3.No standardized way for dealing with multi-class problems;.

Why did you choose to apply this model(s) ?

The reason why i choose these models after looking at the data is because we classifying labeled data that has less than 1000 samples. SVM's work well with small data set and has a number of real world applications.

Strengths of Decision Trees ?

1. Easy to interpret and explain .
2. Nonlinear relationships between parameters do not affect tree performance
3. Requires relatively little effort from user for data preparation

Weakness of Decision Trees ?

1.Very prone to over-fitting data.
2. Can become quite large and a result, pruning the tree is needed.
3. Classifies by rectangular partitioning (so does not handle correlated features very well)

Why did you choose to apply this model(s) ?

The reason why i choose these models after looking at the data is because we are predicting a category of labeled data , it has lots of uses from business to control systems to biomedical engineering to software development etc. This can be attributed to the fact that is very easy model of explain and interpret.

Strengths of KNN ?

1. Effective in the training data is large.
2. Robust to noisy training data.

3. Provides good generalization accuracy on many domains
4. It learns very quickly.

Weakness of KNN ?

1. Take up a lot of memory to run (storing all the instances).
2. Distance based learning is not clear about which type of distance (Manhattan ,Foot rule, Euclidean etc) formula and which attributes to use to produce the best results.
3. Computation cost is quite high because we need to compute distance of each query instance to all training samples.
4. Work well for a small number of dimensions, but not a high number of dimensions

Why did you choose to apply this model(s) ?

The reason why i choose these models after looking at the data is because we are predicting a category of labeled data and the size of data is quite small. Also KNN is easy to understand and has many real world application.

Choosing the best Model.

Based on the experiments performed earlier, it is clear that for this particular application, KNN model is the best model. Because unlike SVC which had higher f1 scores for both training and prediction, it came at the cost of higher training and prediction times and unlike Decision Trees which had much lower training and prediction time, it came at the cost of dismal f1 scores during prediction. KNN offers the best of the other 2 models by having the high f1 score for training and prediction set and a lower train time, though slightly higher prediction time.

How KNN works.

KNN meaning k- nearest neighbors is a model that makes a prediction on the basis of its neighbors. KNN identifies the k-nearest neighbor of C given training set with multiple features . C is another set whose class we want to estimate. It works by classifying C based on similarity measure (e.g., distance function) to find the nearest cases to a new case.
For example if we have k = 5 and classes 'graduate' and 'won't graduate', find the class of C.
Since k=5, we have to find the five nearest neighbors of C. So if 3 out of the five neighbors of have a label 'graduate' and the other 2 have a label 'won't graduate', then the algorithm will choose C as graduate.

The final f1 score = 0.828

Sources:

Dikran Marsupial. "Re: Advantages and disadvantages of SVM?"*Stack Exchnage*. Cross Validated,

02 May. 2012. Web: http://stats.stackexchange.com/questions/24437/advantages-and-disadvantages-of-svm

Kardi Teknomo, Phd.Strength and Weakness of k-Nearest Neighbor Algorithm, 02 May. 2012. Web: http://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm

Bala Desgpande,06.12.2011, 4 key advantages of using decision tree for predictive analytics Retrieved from: http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics

William Chen. "Re: What are the disadvantages of using a decision tree for classifaication ?"*Quora*. Cross Validated, 26 Sep. 2015. Web: https://www.quora.com/What-are-the-disadvantages-of-using-a-decision-tree-for-classification