Machine Learning
Report of Assignment - 3
George Koshy - gxk140830

Question:

Choose any five datasets from the UCI data set website and compare the performance of classifiers the classifiers:
- Decision Trees
- Perceptron (Single Linear Classifier)
- Neural Net
- Support Vector Machines
- Naïve Bayes Classifiers

Language used:
Python

Libraries Used:
Latest dev version of sklearn (Latest dev version 0.18DEV is important as it only has the implementation of Neural Networks as of 03/06/2016)

IDE used:
Jupyter

Analysis of running all the algorithms on the chosen datasets twice:

Dataset	No.instances	No.attributes	Train/test split	Decision Tree Accuracy	Perceptron Accuracy	Neural Net Accuracy	SVM Accuracy	Naive Bayes Accuracy	Best accuracy
Phishing-1	11055	30	80/20	0.966983	0.822705	0.932610	0.951606	0.599276	Decision Tree
Phishing-2	11055	30	80/20	0.958390	0.910900	0.922659	0.944369	0.582994	Decision Tree
Breast Cancer-1	569	31	80/20	0.921053	0.333333	0.666667	0.666667	0.666667	Decision Tree
Breast Cancer-2	569	31	80/20	0.929825	0.377193	0.622807	0.622807	0.614035	Decision Tree

Dataset	No.instances	No.attributes	Train/ test split	Decision Tree Accuracy	Perceptron Accuracy	Neural Net Accuracy	SVM Accuracy	Naive Bayes Accuracy	Best accuracy
Credit Card Approval-1	30000	23	80/20	0.724833	0.770333	0.781000	0.781333	0.386333	SVM
Credit Card Approval-2	30000	23	80/20	0.730333	0.304333	0.781167	0.780667	0.381333	Neural Net
Transfusion-1	748	4	80/20	0.693333	0.806667	0.806667	0.786667	0.746667	Perceptron/ NN
Transfusion-2	748	4	80/20	0.653333	0.753333	0.753333	0.753333	0.753333	Perceptron/NN/ SVM/NB
Ionosphere -1	351	34	80/20	0.845070	0.774648	0.676056	0.943662	0.901408	SVM
Ionosphere -1	351	34	80/20	0.873239	0.774648	0.633803	0.887324	0.873239	SVM

Above is the table elucidating the accuracies of all the datasets.

Observations:

- As seen above various datasets on running multiple times gives slightly varying accuracies.
- Some algorithms perform very well for some datasets and some perform even bad than assigning classes randomly(ideally 50%)
- **The most consistent algorithm observed across datasets was Support Vector Machines**. No matter what the dataset was, SVM gave a pretty decent result(Not the best all the time) SVMs work well when a generally smooth function (or equivalently, mostly varying in a lower-dimensional manifold) gives a good performance. The fact is that is that the generalization properties of an SVM do not depend on the dimensionality of the space. So it performs significantly well for all data sets.
- Neural networks also performed pretty well across datasets.
- **The worst performing algorithms turned out to be Perceptron and Naive Bayes.**
- Perceptron doesn't perform well for some algorithms as the data might not be linearly separable.
- The basic assumption for Naive Bayes algorithm is that the attributes have to be independent. Looks like some of the datasets violated this principle and this resulted in giving very poor results.
- Decision Trees also perform consistent across datasets and is a safe choice irrespective of the dataset.

Conclusion:
Ranking the algorithm according to their usefulness if correlation between the attributes in the dataset is not known.

1. SVM
2. Neural Networks
3. Decision Trees
4. Perceptron
5. Naive Bayes
6. Random Assignment of class variables