Aminul Khan
Machine Learning

Assignment # 1
February 2, 2015

Summary

This paper analyses five different machine learning classification algorithms on two different data sets. The specific algorithms analyzed are decision tree, neural networks, boosting, Support Vector Machine (SVM), and K-nearest neighbors. The two classification problems examined are 1) identifying a breast cancer to either be benign or malignant, and 2) evaluates cars to classify their acceptability. The analysis is done in python using scikit-learn library and PyBrain library for neural network algorithm. After through analysis of the data and varying the parameters of the algorithms a specific learning algorithm is identified as the best algorithm for each of the problem.

General Data Pre-Processing Approach

The general steps taken for preprocessing and cleaning the data are as follows:

- Remove any unrelated data fields (e.g. record id)
- Identify any missing value and apply a mean strategy to fill in the missing data.
- Extract the features and convert categorical features into vector fields (e.g. [1 0 0 0]) to ensure all inputs are numbers.
- Normalize the data
- Split data into 2/3 training and 1/3 testing data

Classification Problem One: Wisconsin Breast Cancer Diagnostic

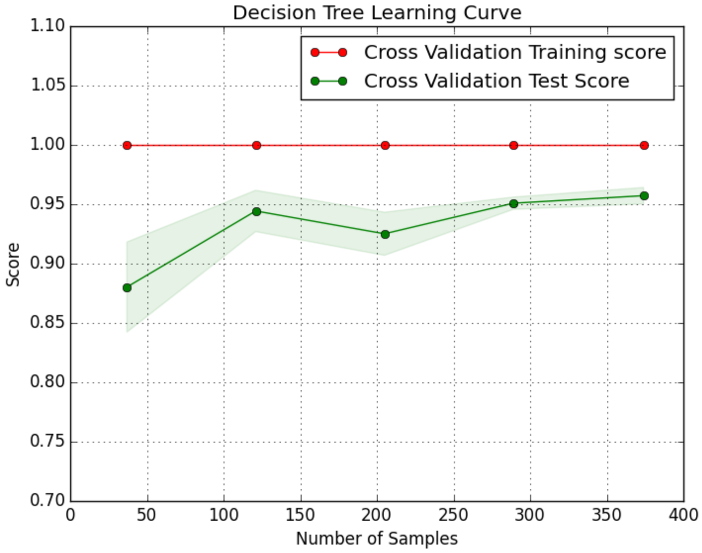
The Wisconsin breast cancer data obtained from UCI Machine Learning Repository contains 699 samples with 10 input attributes. The samples are classified as either benign or malignant. In addition there were 16 missing values within the Bare Nuclei attribute.

In preparing the data I first removed the sample code number, as it is just an ID number that doesn't have any relationship with the cancer data. Next I filled in the 16 missing values of Bare Nuclei using a mean strategy, where the missing values were filled with the mean of all the other data in that column, which turned out to be 4.

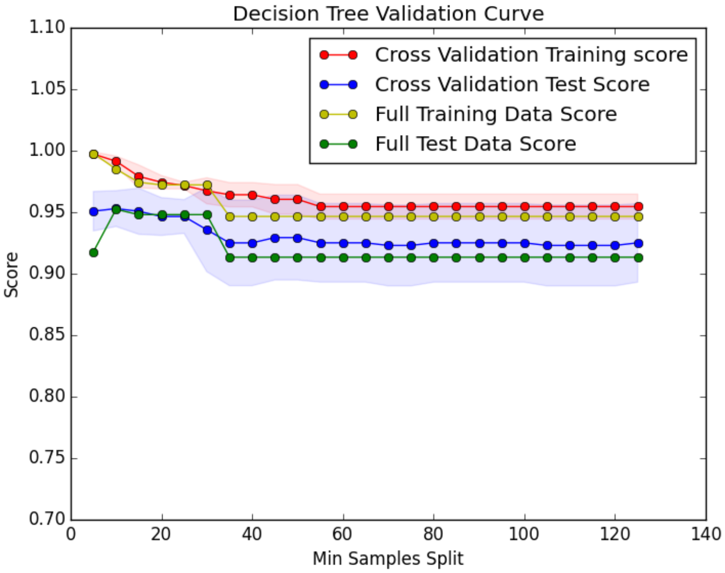
Breast Cancer Learning Algorithm & Data Analysis

Decision Tree

The decision was the first algorithm executed on the data. The default parameters using entropy as the information gain mechanism resulted in a training score of 100% and a testing score of 92.20%. The below graph shows the change in score with varying number of samples. The learning curve shows that as number of samples were increased the score for the cross validation test data also improved to over 95%.



Suspecting that the training data was over fitting I decided to map the scores of validation training, validation test, full training, and full test data scores using min samples split as the varying filed. As you can see for vary small number of min sample split the training data was over fitting and the test data was performing least optimally. The best rest was achieved with min sample split of 5 which gave a 99.78% accuracy on training data and 92.20% accuracy on testing data.

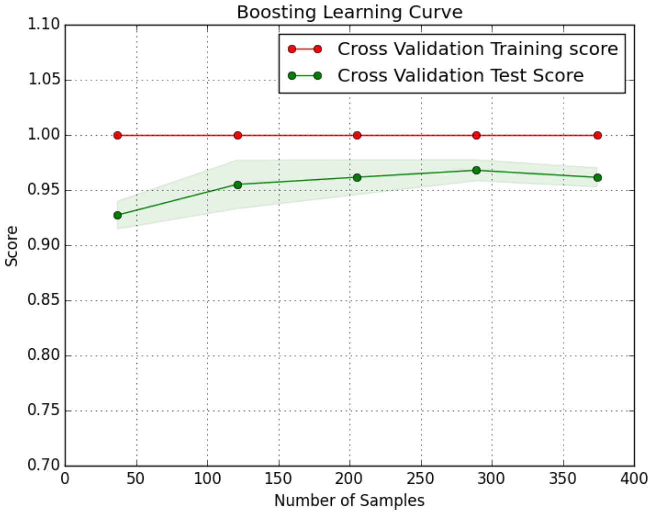


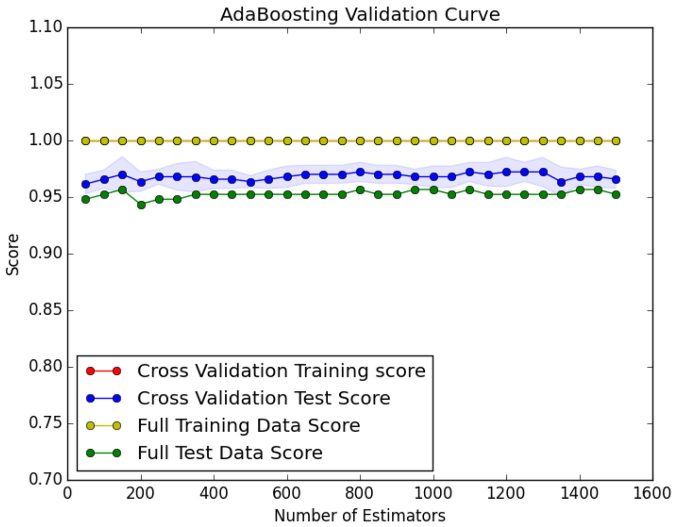
Neural Network

Neural network model was developed using PyBrain and it performed the worst. The accuracy on the training data was 66.23% and the accuracy on the testing data was 64.06%. Which remained around the same vicinity even after performing Kfold cross validation on the data.

Ada Boosting

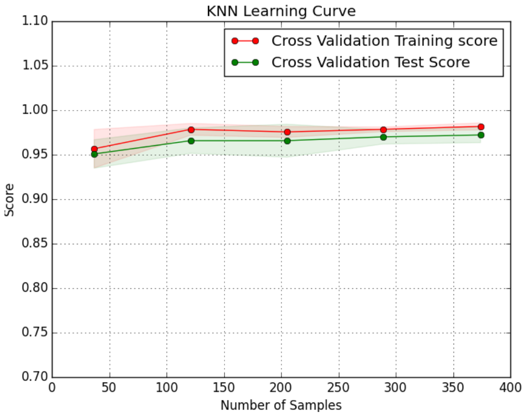
Next analysis was done on the same breast cancer data using Ada Boosting. Initial Ada Boosting algorithm with default values presented a score of 100% accuracy for training data and 94.80% with the testing data. Initial inclination makes it seem that boosting is also over fitting. However, using Kfold cross validation of 5 I graphed the validation curve (second graph below) for varying number of estimators against the accuracy scores. This can be seen on the second graph below. As expected you can see the accuracy score on the testing value increases with number of estimators. But the Kfold cross validation on the training data still kept the accuracy score at 100% or near 100% using this algorithm. The weak learner using in this analysis is decision tree. As we learned in the lectures boosting usually doesn't tend to over fit unless the weak learner is over fitted. This doesn't seem to be the case here as the model performs very well against unseen data. Using gradient search the best accuracy score achieved using 650 estimators, which resulted in a 95.23% accuracy on the testing data.

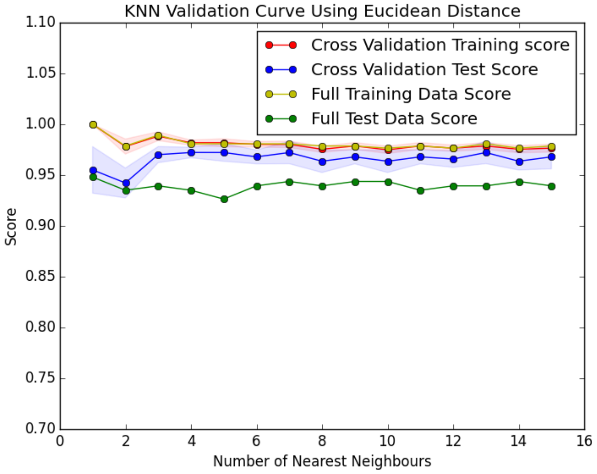


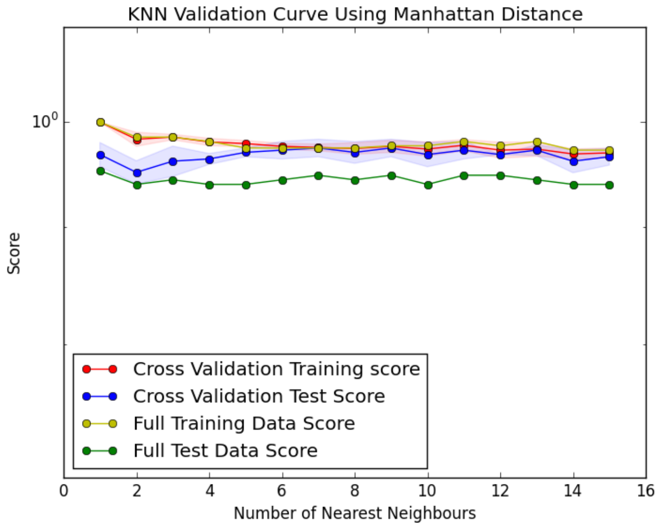


KNN

The below three graphs outline the KNN algorithm on the breast cancer data. The first is the learning curve with varying number of samples. Though their were slight improvement on the score with varying number of samples it did not vary the scores much. This is expected as the algorithm is doesn't really train anything. But rather during the prediction it evaluates KNN to find the value for unseen data. The next two graphs show the accuracy scores with varying number of nearest neighbors. The first using Eucidean distance and the second using Manhattan distance calculation. It can be inferred from the graph that the Eucidean distance calculation did the best.







Support Vector Machines

The final algorithm tested on the breast cancer data was the support vector machine (SVM). Since SVM is trying to maximize the distance between two class of data it does significantly well as you provide it more and more data. The mean accuracy of SVM from Kfold cross validation was 92.94%. The below graph shows the drastic change in accuracy scores of SVM as the number of samples was increased from 50 to around 400. It is fair to say, based on this analysis, if we have a lot of data SVM maybe the model of choice.

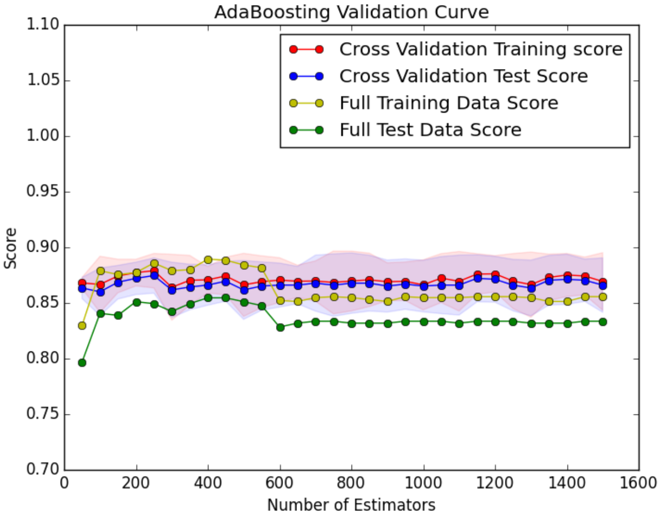
Summary of Breast Cancer Analysis Result:

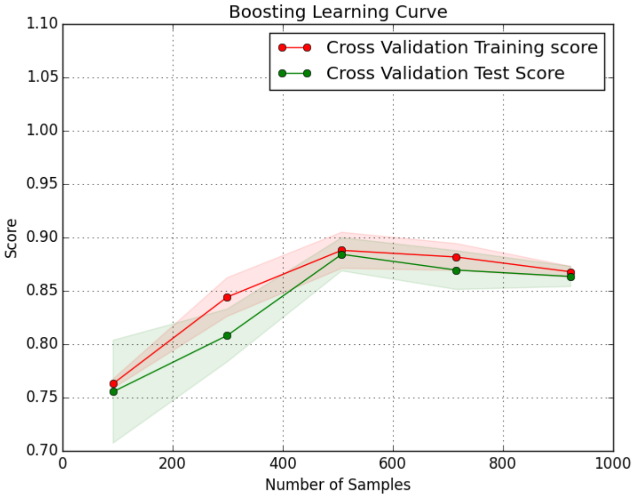
The algorithm of choice for the breast cancer data is boosting with number of estimators as high as 450. This aligns with are lectures as boosting does well with more estimators. Wit the summation of the results of the weak learner boosting has created a better model then the other algorithms.

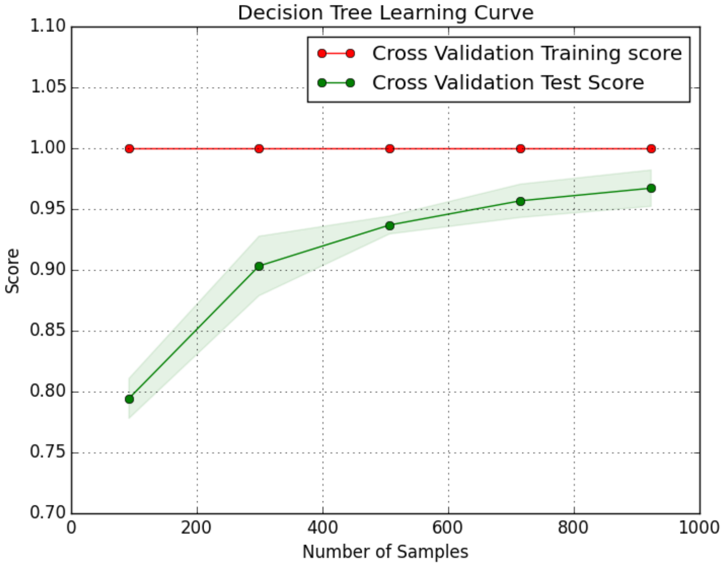
Classification Problem Two: Car Quality Check

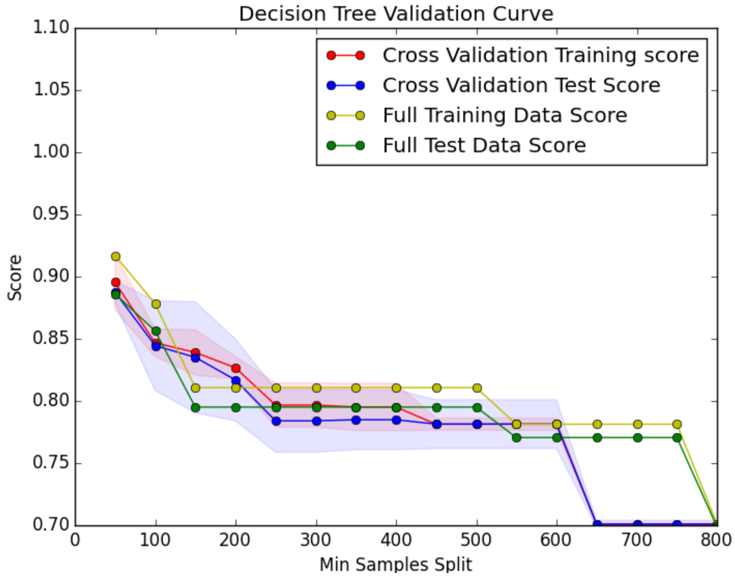
The car quality check data set also obtained from UCI Machine Learning Database had a sample size of 1728 with 6 attributes. A similar analysis as the breast cancer data was performed on the car data as well. Below graphs show the results of the various algorithm executed on the call data.

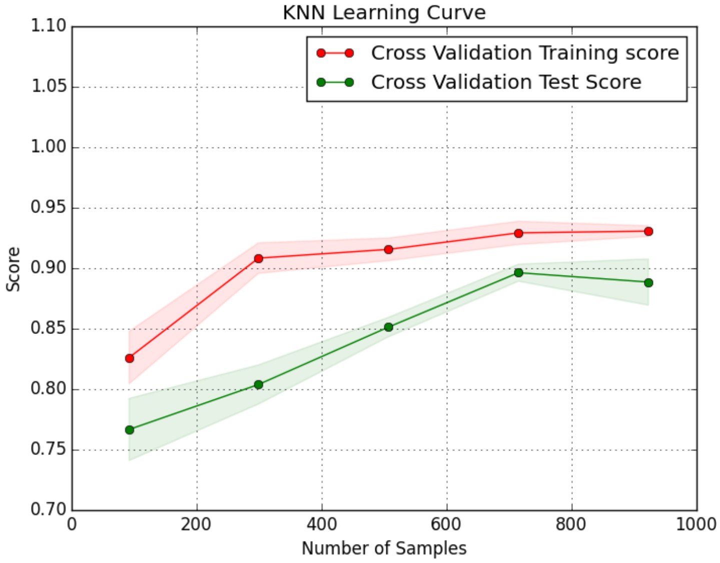
The best performing algorithm for the car data was KNN. I assume this makes sense because cars that are similar in several of their features will be in the same quality. So using nearest neighbor the algorithm was to find the most look alike car.

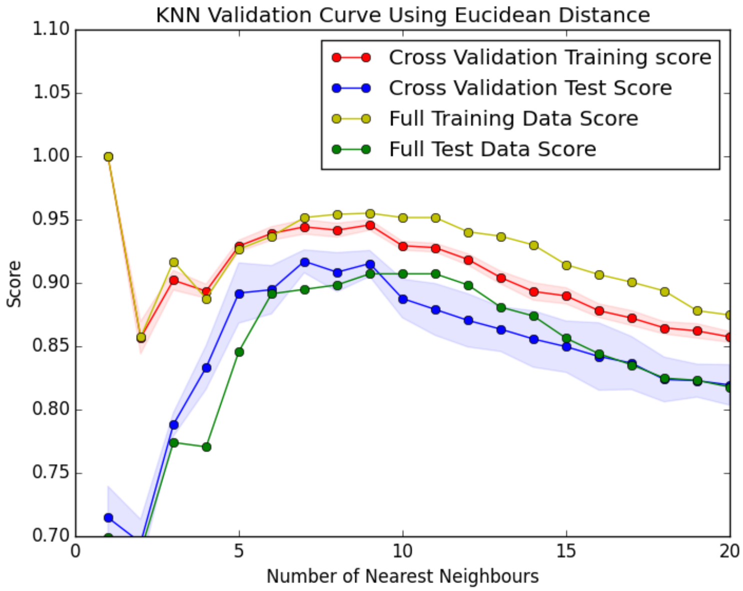

AdaBoosting Validation Curve
Cross Validation Training score
Cross Validation Test Score
Full Training Data Score
Full Test Data Score
Score
Number of Estimators
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
200
400
600
800
1000
1200
1400
1600


Boosting Learning Curve
Cross Validation Training score
Cross Validation Test Score
Score
Number of Samples
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
200
400
600
800
1000

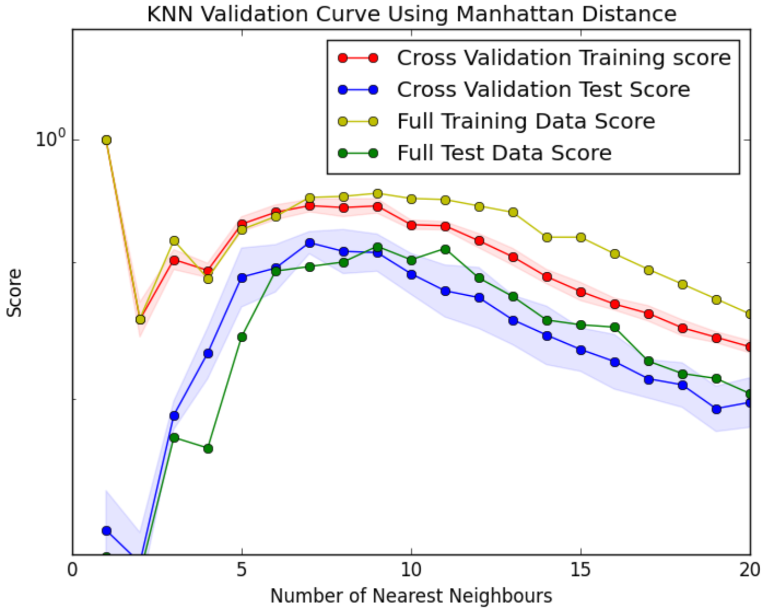

Decision Tree Learning Curve
Cross Validation Training score
Cross Validation Test Score
Score
Number of Samples
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
200
400
600
800
1000


Decision Tree Validation Curve
Cross Validation Training score
Cross Validation Test Score
Full Training Data Score
Full Test Data Score
Score
Min Samples Split
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
100
200
300
400
500
600
700
800

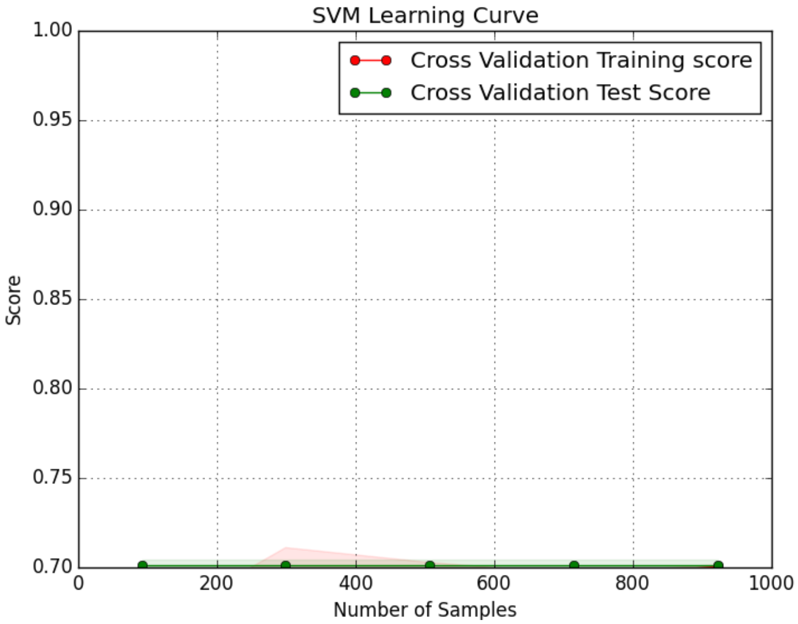

KNN Learning Curve
Cross Validation Training score
Cross Validation Test Score
Score
Number of Samples
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
200
400
600
800
1000



KNN Validation Curve Using Eucidean Distance



KNN Validation Curve Using Manhattan Distance


SVM Learning Curve
Cross Validation Training score
Cross Validation Test Score
Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
200
400
600
800
1000
Number of Samples

Bibliography

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30. 4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).
4. M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.
5. B. Zupan, M. Bohanec, I. Bratko, J. Demsar: Machine learning by function decomposition. ICML-97, Nashville, TN. 1997 (to appear)