KHULNA UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Assignment On Machine Learning

CourseID: CSE 4211

Khandaker Tasnim Huq, Roll: 1307002

12/30/2017

1)Introduction :

> Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it

2) Data Sources and description:

The given dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

3) Machine Learning Methods:

a) Feature Selection:

For getting the best feature set, there are some feature selection methods. In this experiment, two Feature Selection Methods were used:

1) Variance Threshold: It is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples. Suppose, we have a dataset and we want to remove all features that are either one or zero (on or off) in more than 80% of the samples. the variance of such variables is given by

$$\mathrm{Var}[X] = p(1-p),$$

so we can select using the threshold .8 * (1 - .8)

2) Select K best with ANOVA: Select K best selects features according to the k highest scores. ANOVA Compute the F-value for the provided sample.

b) Classifiers:

1) Random Forest: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

2) SVM: There were two methods under SVM were used in this experiment.

a)SVC: Support Vector Classification.The implementation is based on libsvm. For details on the precise mathematical formulation of the provided kernel functions and how gamma, coef0 and degree affect each other.

b)Liner SVC: Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.

3) Neural Network: **Multi-layer Perceptron (MLP)** is a supervised learning algorithm that learns a function:

   $f(\cdot) : R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. The solver for weight optimization that were used in this experiment are:

   a) 'lbfgs' is an optimizer in the family of quasi-Newton methods,
   b) 'sgd' refers to stochastic gradient descent,
   c) 'adam' refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba.

4) Nearest Neighbor: The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning)

5) Gaussian Naïve Bayes:

   Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. The likelihood of the features is assumed to be Gaussian Function.

4)Results:

a) Best Feature Set:
   At first all the classifiers were iteratively run with this feature size site and evaluated separately using Variance Threshold, the candidate threshold are: [0.15, 0.1, 0.01, .001, .0001, .000001, 0.0000000001, 0.0]. It will return the corresponding feature size after doing

the fitting training data and transforming the test data. The candidate size set was: [5L, 6L, 8L, 12L, 17L, 21L, 22L, 22L]. Then These were implemented in SelectKBEst with ANOVA method.

In case of Variance Threshold, the feature variance are:

MDVP:Jitter(Abs) 1.22767883455e-09
MDVP:PPQ 7.30793079709e-06
MDVP:RAP 8.83061541727e-06
MDVP:Jitter(%) 2.32237578189e-05
Jitter:DDP 7.94797951883e-05
Shimmer:APQ3 0.00010178218758
Shimmer:APQ5 0.000143621176916
MDVP:APQ 0.000286204222293
MDVP:Shimmer 0.000350189222418
Shimmer:DDA 0.000916010629611
NHR 0.00166699208499
DFA 0.00303344330242
spread2 0.0072442157036
PPE 0.00770897700641
RPDE 0.010397399212
MDVP:Shimmer(dB) 0.0370637096358
D2 0.142346206109
spread1 1.11559419662
HNR 18.1830606982
MDVP:Fo(Hz) 1610.80375647
MDVP:Flo(Hz) 1660.0546245
MDVP:Fhi(Hz) 9367.50562056

In case of ANOVA, the feature score are:

MDVP:Fhi(Hz) 3.18198860411
NHR 4.89029601206
DFA 5.73650846709
Jitter:DDP 12.4266255123
MDVP:RAP 12.430684748
MDVP:Jitter(%) 12.4885418965
MDVP:PPQ 14.9489909887
MDVP:Flo(Hz) 15.0458948162
MDVP:Fo(Hz) 16.8861034215
MDVP:Jitter(Abs) 17.6073983281
RPDE 18.1123367607
D2 18.817590121
Shimmer:DDA 19.7636863387
Shimmer:APQ3 19.7668416007
MDVP:Shimmer(dB) 20.3277763588
Shimmer:APQ5 20.6266451805
MDVP:APQ 21.6537942131
MDVP:Shimmer 22.4569055805
HNR 28.7132559149
spread2 34.3939556534
PPE 59.7590005826
spread1 69.3815087069

After experimenting, the optimized feature set size is 12. So the best feature set is:

Variance Threshold: ['MDVP:Fhi(Hz)', 'MDVP:Flo(Hz)', 'MDVP:Fo(Hz)', 'HNR', 'spread1', 'D2', 'MDVP:Shimmer(dB)', 'RPDE', 'PPE', 'spread2', 'DFA', 'NHR']

ANOVA: ['spread1', 'PPE', 'spread2', 'HNR', 'MDVP:Shimmer', 'MDVP:APQ', 'Shimmer:APQ5', 'MDVP:Shimmer(dB)', 'Shimmer:APQ3', 'Shimmer:DDA', 'D2', 'RPDE']

The common Features from both of the feature selection methods are then:

set(['RPDE', 'MDVP:Shimmer(dB)', 'PPE', 'spread1', 'spread2', 'HNR', 'D2'])

b) Confusion Matrix:
For Variance Threshold:

Naive Bayes: 0.75

[[10 3]
[7 20]]

Cross - validated scores: [0.83870968 0.67741935 0.77419355 0.67741935 0.87096774] 0.767741935484

SVC: 0.725

[[2 11]
[0 27]]

Cross - validated scores: [0.77419355 0.77419355 0.77419355 0.77419355 0.77419355] 0.774193548387

Neural Net, Solver=adam: 0.675

[[0 13]
[0 27]]

Cross - validated scores: [0.77419355 0.77419355 0.77419355 0.77419355 0.77419355] 0.774193548387

Neural Net, Solver=lbfgs: 0.675

[[0 13]
[0 27]]

Cross - validated scores: [0.77419355 0.77419355 0.77419355 0.77419355 0.77419355] 0.774193548387

Neural Net, Solver=sgd: 0.675

[[0 13]
[0 27]]

Cross - validated scores: [0.77419355 0.77419355 0.77419355 0.77419355 0.77419355] 0.774193548387

Liner SVC: 0.675

[[0 13]
[0 27]]

Cross - validated scores: [0.77419355 0.80645161 0.22580645 0.77419355 0.77419355]
0.670967741935

Random Forest: 0.825

[[9 4]
 [3 24]]

Cross - validated scores: [0.90322581 0.90322581 0.90322581 0.80645161 0.93548387]
0.890322580645

Nearest Neighbor: 0.85

[[10 3]
 [3 24]]

Cross - validated scores: [0.83870968 0.96774194 0.77419355 0.87096774 0.87096774]
0.864516129032

For ANOVA:

Naive Bayes: 0.7

[[12 1]
 [11 16]]

Cross - validated scores: [0.77419355 0.64516129 0.83870968 0.61290323 0.87096774] 0.748387096774

Liner SVC: 0.85

[[7 6]
 [0 27]]

Cross - validated scores: [0.87096774 0.90322581 0.70967742 0.80645161 0.90322581] 0.838709677419

Neural Net, Solver=adam: 0.8

[[8 5]
 [3 24]]

Cross - validated scores: [0.93548387 0.87096774 0.80645161 0.70967742 0.87096774] 0.838709677419

Neural Net, Solver=lbfgs: 0.85

[[9 4]
 [2 25]]

Cross - validated scores: [0.87096774 0.83870968 0.80645161 0.74193548 0.87096774] 0.825806451613

Neural Net, Solver=sgd: 0.725

[[5 8]
 [3 24]]

Cross - validated scores: [0.87096774 0.77419355 0.77419355 0.74193548 0.87096774] 0.806451612903

SVC: 0.725

[[3 10]
 [1 26]]

Cross - validated scores: [0.90322581 0.93548387 0.83870968 0.70967742 0.87096774] 0.851612903226

Random Forest: 0.85

[[8 5]
 [1 26]]

Cross - validated scores: [0.90322581 0.87096774 0.80645161 0.77419355 0.83870968]   0.838709677419

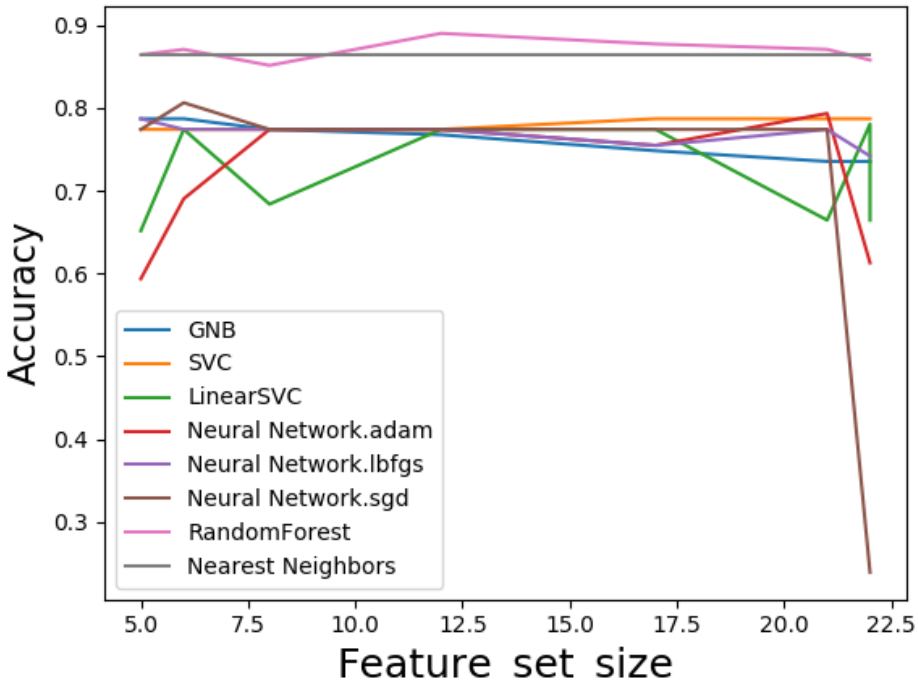
Nearest Neighbor: 0.8

[[7 6]
 [2 25]]

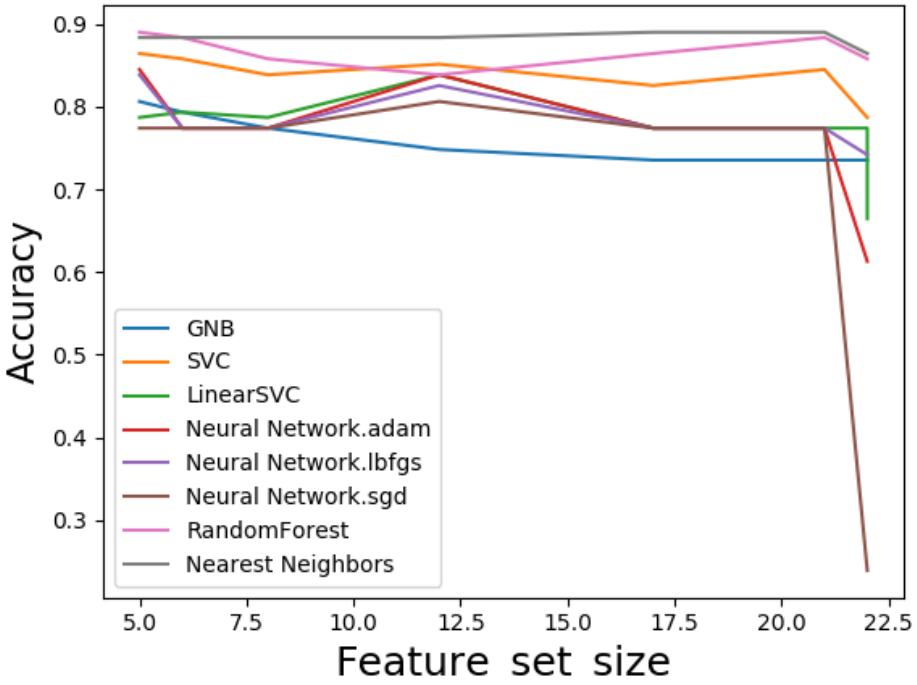
Cross - validated scores: [0.93548387 0.90322581 0.90322581 0.77419355 0.90322581]   0.883870967742

c) Accuracy Curve for different feature set. X axis: Feature Set, Y axis: Accuracy:

For Variance Threshold:



For Feature Selection with ANOVA:



5)Conclusion:

This experiment gave us a great insight about the aspects of Machine Learning and its application.

ANOVA gave better results than Threshold Variance in case of feature selection technique. In both cases, Random Forest performs best.

As a subfield of Artificial Intelligence, its goal is to enable computers to learn on their own. A machine's learning algorithm enables it to identify patterns in observed data, build models that explain the world, and predict things without having explicit pre-programmed rules and models.

6)References:

a) Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

b) https://matplotlib.org/

c) https://www.coursera.org/learn/machine-learning