Name: Aditya Borde

-----Support Vector Machine-----

Data Set: [Promoters](http://www.hlt.utdallas.edu/~vgogate/ml/2015s/homeworks/promoters_data.tar.gz)

**Accuracy** **Using LIBSVM:**

|  |  |
| --- | --- |
| Kernels | Accuracy |
|
| Linear Kernel (0) | 85.7143% |
| Polynomial Kernel (1) | 74.2857% |
| Exponential Kernel (2) | 77.1429% |
| Sigmoid Kernel (3) | 45.7143% |

**Perceptron:**

Data Set: [Promoters](http://www.hlt.utdallas.edu/~vgogate/ml/2015s/homeworks/promoters_data.tar.gz):

**Accuracy** **Using WEKA:** 85.7143%

**LIBSVM Commands:**

1) To create the model from the Training Set predefining the number of kernels:

**F:\libsvm-3.20\windows> svm-train.exe –t <#Kernels> <Path\_TrainingSet>**

2) To give the prediction using generated training set model on validation set:

**F:\libsvm-3.20\windows> svm-predict.exe <Path\_ValidationSet> <Path\_TrainingModel> <Path\_OutputFile>**

E.g.:

1) To create Linear Kernel Model from training set

**F:\libsvm-3.20\windows> svm-train.exe –t 0 training.new**

2) Accuracy using above training model on Validation set

**F:\libsvm-3.20\windows>** **svm-predict.exe validation. New training.new. Model svmValidation0.out**

**Conclusions:**

From the LIBSVM – We can see that there is decrease in the accuracy value as the number of kernels are increasing on given data set. Therefore, increase in dimensionality will decreases the accuracy in Support Vector Machine.

From the Perceptron using WEKA – Perceptron gives the same accuracy as that of the LIBSVM using Linear Kernel.

**Reasoning:**

From the LIBSVM accuracy using different kernel, linear kernel gives most accurate result. It shows, that data is linearly separable. Hence Perceptron shows same result as that of Linear Kernel in LIBSVM as data is linearly separable.

-----Boosting-----

**Classifiers Used:** Decision Stump, Naïve Bayes and J48.

All tables contain Error rates: (Incorrectly classified Instances in %)

**Data Set:** [Chess (king-rook-vs-king-pawns](http://www.hlt.utdallas.edu/~vgogate/ml/2015s/homeworks/promoters_data.tar.gz))

**Iterations = 30**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.9487 % | 31.95% | 5.7259 % |
| Naïve Bayes | 12.1402 % | 12.234% | 4.6308 % |
| J48 | 0.5632 % | 0.5632 % | 0.4068 % |

**Iterations = 100**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.9487 % | 31.95% | 4.6934 % |
| Naïve Bayes | 12.1402 % | 12.1089% | 4.5995 % |
| J48 | 0.5632 % | 0.5006 % | 0.4693 % |

**Iterations = 150**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.9487 % | 30.45% | 4.5995 % |
| Naïve Bayes | 12.1402 % | 12.076% | 4.5995 % |
| J48 | 0.5632 % | 0.5006 % | 0.4693 % |

**Data Set:** [Iris](http://www.hlt.utdallas.edu/~vgogate/ml/2015s/homeworks/promoters_data.tar.gz)

**Iterations = 30**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.3333 % | 20.6667 % | 5.3333 % |
| Naïve Bayes | 4 % | 3.3333 % | 6 % |
| J48 | 4 % | 6 % | 6 % |

**Iterations = 100**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.3333 % | 10.000 % | 6.6667 % |
| Naïve Bayes | 4 % | 3.3333 % | 6 % |
| J48 | 4 % | 4.6667 % | 6 % |

**Iterations = 150**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 33.3333 % | 11.3333 % | 7.3333 % |
| Naïve Bayes | 4 % | 4 % | 6 % |
| J48 | 4 % | 4.6667 % | 6 % |

**Data Set:** Post-Operative Patient Data Set

**Iterations = 30**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 30 % | 30 % | 30 % |
| Naïve Bayes | 32.2222 % | 30 % | 32.2222 % |
| J48 | 30 % | 32.2222 % | 44.4444 % |

**Iterations = 100**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 30 % | 30 % | 30 % |
| Naïve Bayes | 32.2222 % | 31.1111 % | 32.2222 % |
| J48 | 30 % | 31.1111 % | 44.4444 % |

**Iterations = 150**

|  |  |  |  |
| --- | --- | --- | --- |
| Base Learner | Vanilla | Bagging | Boosting |
| Decision Stump | 30 % | 30 % | 30 % |
| Naïve Bayes | 32.2222 % | 31.1111 % | 32.2222 % |
| J48 | 30 % | 31.1111 % | 44.4444 % |

1. Which algorithms + data set combination is improved by Bagging?

* Data Set: **Chess** – Improvement in accuracy (Reduction in Error)
  + Naïve Bayes
  + Decision Stump
  + J48
* Data Set: **Iris** – Improvement in accuracy (Reduction in Error)
  + Decision Stump.
  + Slightly improvement in Naïve Bayes.
* Data Set: **Post-Operative Patient Data Set** (Reduction in Error)
  + Naïve Bayes.

2. Which algorithms + data set combination is improved by Boosting?

* Data Set: **Chess** – Improvement in accuracy (Reduction in Error)
  + Naïve Bayes
  + Decision Stump
  + J48
* Data Set: **Iris** – Improvement in accuracy (Reduction in Error)
  + Decision Stump.
* Data Set: **Post-Operative Patient Data Set** (Reduction in Error)
  + No algorithm shows improvement.

3. Can you explain these results in terms of the bias and variance of the learning algorithms applied to these domains? Are some of the learning algorithms unbiased for some of the domains? Which ones?

* Data Set: **Chess –** Error Rate Observations
  + J48 has low bias and low variance as there is slight reduction in error after Bagging and Boosting.
  + Decision Stump & Naïve Bayes have high bias.
* Data Set: **Iris** – Error Rate Observations
  + Decision Stump has high bias and high variance. Bagging/Boosting improves the performance.
  + Naïve Bayes has slight variance.
* Data Set: **Post-Operative Patient Data Set** - Error Rate Observations
  + Decision Stump and Naïve Bayes are unbiased.
    - Unbiased Algorithms and their Data sets:
  + Decision Stump, J48 and Naïve Bayes are unbiased on Post-Operative Patient Data Set.
  + J48 and Naïve Bayes for Iris data set are unbiased.

**Conclusions:**

Bagging reduces the variance and Boosting helps to reduce bias. Sometimes, Bagging/Boosting won’t help in accuracy improvement. Unbiased algorithms for particular data set will not ensure accuracy improvement after Boosting. There is more error introduced in case of J48 algorithm in *Post-Operative Patient Data Set* after Bagging and Boosting.

-----K-means clustering on images-----

Table has been constructed considering the following formulae.

1) Compression Ratio =

2) Variance =

- Mean.

Mean and variance has been calculated over five different observations.

Here = 5.

Input Image: **Koala.jpg** (Original Size = 780,831 bytes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | Compression Ratio | | | | | Average | Variance |
| 2 | 5.9417 | 6.0024 | 5.9701 | 5.9988 | 5.9952 | 5.98157 | 0.00052750 |
| 5 | 4.4464 | 4.4464 | 4.4247 | 4.4483 | 4.4424 | 4.44168 | 0.00007542 |
| 10 | 4.7326 | 4.7348 | 4.7460 | 4.7846 | 4.7961 | 4.75873 | 0.00007544 |
| 15 | 4.9019 | 4.9701 | 4.9677 | 4.9407 | 4.9726 | 4.95049 | 0.00072527 |
| 20 | 4.9800 | 4.9800 | 5.0530 | 5.0403 | 5.0125 | 5.01303 | 0.00090468 |

Input Image: **Penguins.jpg** (Original Size = 777,835 bytes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | Compression Ratio | | | | | Average | Variance |
| 2 | 9.157509 | 9.157509 | 9.149130 | 9.157509 | 9.149130 | 9.1541559 | 0.000016 |
| 5 | 7.183908 | 7.320644 | 7.235890 | 7.230657 | 7.251631 | 7.2442770 | 0.001957 |
| 10 | 6.680026 | 6.613756 | 6.640106 | 6.680026 | 6.613756 | 6.6454013 | 0.000885 |
| 15 | 6.648936 | 6.653359 | 6.671114 | 6.570302 | 6.734006 | 6.6551311 | 0.002742 |
| 20 | 6.675567 | 6.697923 | 6.752194 | 6.711409 | 6.770480 | 6.7213335 | 0.001221 |

**Tradeoff between Image Quality and Degree of Compression:**

Degree of compression is firstly decreasing till K=10 and it is increasing with slight value after it. Also, output images show improvement in image quality as there is increase in value of K. So, K=10 is good tradeoff value between degree of compression and image quality, considering both the images.

**Conclusions:**

Degree of compression is changing by slight value after K=10 taking both the images into account. Actually, it is increasing slightly after K=10.

We can say, after certain value of K, Degree of compression will have only slight change in its value.

If image quality is concerned, increase in value of K ensures the image quality improvement as pixels are divided into more number of relevant clusters.