HOMEWORK 4

Name: Vaishakhi Kulkarni

Net Id: vpk140230

1) Support Vector Machine:

To run support vector machines experiment use LIBSVM, follow the following steps:

- Download libsvm
- Navigate to the libsvm directory through terminal (on MAC OS)
- Copy the training.new and validation.new to the folder
- Run the following command for Learning:
    - ./svm-train -t 0 training.new
    - This will generate a file named “training.new.model”
    - Use this model file to predict the rating
- Run the following command:
    - ./svm-predict validation.new training.new.model test.out1
    - It will give you the output and also store it in test.out1 file

Accuracy with different kernels

Kernel Used	**Accuracy**
0	85.7143
1	74.2857
2	77.1429
3	45.7143

The accuracy decreases as we move from linear classifier to higher dimensional classifier. Thus, we conclude that the accuracy decreases drastically when kernel is 3.We can observe that data is getting classified very correctly up to kernel 2.For kernel 0 data is getting linearly classified.

Perceptron using Weka for given dataset is : 85.71%

The accuracy of a single perceptron is same as SVM kernel 0.

2. Boosting:

I have selected 3 classifiers Naïve Bayes, Logistic Regression and Decision stump also performed Bagging and boosting for iterations 30,100 and 150.
I have recorded accuracies in % for 3 data sets. (HAVE ATTACHED ALL DATASETS IN ARFF FORMAT)

To understand the effect of our experiment on variance and bias, we need to understand the effect of bagging and boosting as the techniques on Variance and bias first.
As we do bagging on classification problem, we tend to average the variance and in boosting, we average the bias.

CAR DATA SET:

ACCURACIES FOR CAR EVALUATION DATA SET

Iterations	Base Learner	Vanilla	Bagging	Boosting
30	Naïve Bayes	85.5324	85.0694	90.1042
30	Logistic Regression	92.4769	93.1134	93.1134
30	Decision Stump	70.0231	70.0231	70.0231

Iterations	Base Learner	Vanilla	Bagging	Boosting
100	Naïve Bayes	85.5324	85.5903	90.1042
100	Logistic Regression	93.1713	93.1134	93.1134
100	Decision Stump	70.0231	70.0231	70.0231

Iterations	Base Learner	Vanilla	Bagging	Boosting
150	Naïve Bayes	85.5324	85.4167	90.1042
150	Logistic Regression	93.1713	93.1134	93.1134
150	Decision Stump	70.0231	70.0231	70.0231

1. No change in accuracy after bagging and boosting on the car data set. So this means the Car data set is very much pure.
2. Logistic for Car data set is unbiased.
3. Logistic have highest accuracy for Vanilla, Bagging and Boosting.
4. Naïve Bayes is somewhat biased and has low variance.
5. Logistic regression is unbiased.
6. Iterations are not affecting much for variance and accuracy calculations in Vanilla while very slit changes present in Bagging and boosting.

IRIS DATASET:

ACCURACIES FOR IRIS EVALUATION DATA SET

Iterations	Base Learner	Vanilla	Bagging	Boosting
30	Naïve Bayes	96	96.6667	94
30	Logistic Regression	95.3333	96.6667	95.3333
30	Decision Stump	66.6667	79.3333	94.6667

Iterations	Base Learner	Vanilla	Bagging	Boosting
100	Naïve Bayes	96	93.3333	94
100	Logistic Regression	95.3333	95.3333	95.3333
100	Decision Stump	66.6667	90	93.3333

Iterations	Base Learner	Vanilla	Bagging	Boosting
150	Naïve Bayes	96	96	96
150	Logistic Regression	95.3333	95.3333	94.6667
150	Decision Stump	66.6667	95.3333	94.6667

1. IRIS is unbiased for Logistic Regression and Naïve Bayes.
2. For Naïve Bayes and Logistic Regression the variance is very less for Bagging and Boosting. Accuracy is improved by using Bagging and Boosting.
3. Gain at most accuracy while performing Bagging.
4. For Decision stump in Bagging and Boosting we have high accuracy but if used for Vanilla the accuracy is very less.
5. Decision Stump + Bagging gives higher accuracy in all specified iterations.
6. Decision Stump + Boosting gives higher accuracy in all specified iterations.
7. Decision Stump has high variance therefore bagging and boosting both helped to improve accuracy.
8. When iterations are increased the accuracy of Logistic regression remains consistent after applying bagging as well as boosting.
9. Have fluctuation in behavior of Naïve Bayes for Bagging as well as boosting as iterations increases.
10. Iterations are not affecting much for variance and accuracy calculations in Vanilla while very slit changes present in Bagging and boosting.
11. Naïve Bayes and logistic regression has acceptable bias and variance.

BREASTCANCER DATASET:

ACCURACIES FOR BREASTCANCER EVALUATION DATA SET

Iterations	Base Learner	Vanilla	Bagging	Boosting
30	Naïve Bayes	71.6783	72.3776	72.7273
30	Logistic Regression	68.8811	68.8811	68.8811
30	Decision Stump	68.5315	74.8252	72.7273

Iterations	Base Learner	Vanilla	Bagging	Boosting
100	Naïve Bayes	71.6783	72.7273	71.6783
100	Logistic Regression	69.2308	68.1818	69.2308
100	Decision Stump	68.5315	75.1748	71.6783

Iterations	Base Learner	Vanilla	Bagging	Boosting
150	Naïve Bayes	71.6783	72.7273	64.6853
150	Logistic Regression	69.2308	68.8811	68.8811
150	Decision Stump	68.5315	75.1748	70.2797

1. Naïve Bayes + Bagging increases the accuracy
2. Logistic Regression + Bagging and Boosting either remains consistent or reduces slightly.
3. Decision Stamp + Bagging increases the accuracy.
4. Decision Stamp + Boosting increases the accuracy.
5. When iterations 150 for Naïve Bayes the accuracy decreases drastically after applying boosting.
6. Dataset is unbiased for Naïve Bayes and Logistic Regression.
7. While for Decision Stump has high variance. The accuracy increases as bagging and boosting is applied on the dataset.
8. When iterations are increasing and reaches at 150 the accuracy reduces drastically for Naïve Bayes after applying Boosting on it.

3) **K Means Clustering on images**

I. The output images for different values of K refer to folder Koala and Penguin

ii. For Each value of K ran for 3 iterations

iii. Compression ratio = original size / compressed size

iv. Variance: 1/3 * summation of (xi-mean) ^ 2

K means Compression for Koala.jpg (780,831 bytes)

K	Compression ratio with original images(3 runs)			Mean	Variance
2	5.968	5.968	5.967	5.9677	0.0000013
5	4.433	4.4215	4.4226	4.4257	0.00002607
10	4.717	4.731	4.778	4.742	0.00068133
15	4.9699	4.974	4.973	4.9725	0.00000241
20	4.9374	5.030	4.9766	4.9824	0.00144155

K means Compression for Penguins.jpg (777,835 bytes)

K	Compression ratio with original images(3 runs)			Mean	Variance
2	9.1513	9.1520	9.1520	9.1517	0.00000009
5	7.0294	7.3780	7.1525	7.1866	0.0243
10	6.5633	7.051	6.515	6.7100	0.0554
15	6.8292	6.7289	6.6506	6.7362	0.00534
20	6.5692	6.5915	6.6529	6.6045	0.001251

From above observations we can see that the compression ratio reduces and then again increases a bit. As k increases the quality of the image increases. When value of k is 10 the image looks good having good quality of image in Koala.jpg while the images looks good when k= 10 in Penguins.jpg.