Homework 4 Report

1. EM Algorithm

Randomization strategy: first run k-means several iterations to get the initialization parameters (**Bishop – Pattern Recognition And Machine Learning, p. 438**). I use Weka's k-means for initialization, and used 3 different initializations iterations, 1, 2, and the following is the clustering centers and variances after initialization:

Iteration Number	Initialization Results	
1	Center[0]	25.000356758
1	Center[1]	16.584711014
1	Center[2]	9.7929782265
1	Variance[0]	4.6696715082
1	Variance[1]	0.0314442732
1	Variance[2]	23.625294502
2	Center[0]	25.486654429
2	Center[1]	15.482717626
2	Center[2]	5.5992657029
2	Variance[0]	0.998096618
2	Variance[1]	0.8874658517
2	Variance[2]	1.6671038293

Notice that when we set iteration number to 2, the results is close to the final results after using EM, so I did not use iteration number greater than 2 for initialization.

The following gives the log likelihood after running different number of iteration of EM algorithms under 2 initialization settings (1 and 2), and setting variance to 1 or not:

Do Not Set Variance To 1

Initialization	EM Iterations	Log Likelihood/10000
1	1	-1.8616417607
1	3	-1.8434107284
1	7	-1.7482363939
1	10	-1.6237056923
1	12	-1.5099772255
1	15	-1.509977175
2	1	-1.5099771783
2	3	-1.509977175

Set Variance To 1

Initialization	EM Iterations	Log Likelihood/10000
1	1	-1.51746051
1	2	-1.5100775131
1	3	-1.5100775131
2	1	-1.5100775131
2	2	-1.5100775131

From the results, we see that if we set the variance to 1, the EM algorithm will converge much faster, it shows that if we know some distribution information about the data, we can get a much faster convergence.

2. Boosting & Bagging

Three data set chooses: Heart Disease Data Set (Dataset 1), Pima Indians Diabetes Data Set (Dataset 2), Ionosphere Data Set (Dataset 3), all of them are from UCI Repository, and you can find the converted .arff file in the embkk/dataset directory.

For Heart Disease Data Set, if the missing rates if >50%, the the attribute is committed, the other missing value are being modified as the average value giving the class. The following is the experiment results (Iteration number: 30, 100, 150, I marked the best performance of a row as red):

Iteration Number: 30

Dataset 1:

Base Learner	Vanilla	Bagging	Boosting
J48	22.4489795918	21.5136054422	20.7968901846
Logistic Regression	20.4081632653	20.7482993197	20.4506802721
Decision Stump	21.3151927438	20.5782312925	20.3325774754

Dataset 2:

Base Learner	Vanilla	Bagging	Boosting
J48	26.171875	25.3255208333	25.4836309524
Logistic Regression	24.4791666667	24.7916666667	25.146484375
Decision Stump	25.6944444444	25.1953125	25.1591435185

Dataset 3:

Base Learner	Vanilla	Bagging	Boosting
J48	8.547008547	11.0398860399	11.3146113146
Logistic Regression	9.8290598291	11.1111111111	11.3603988604
Decision Stump	12.3456790123	12.0607787274	11.0477999367

Iteration Number: 100

Dataset 1:

Base Learner	Vanilla	Bagging	Boosting
J48	22.4489795918	21.4285714286	20.7968901846
Logistic Regression	20.4081632653	20.6802721088	20.4506802721
Decision Stump	21.3151927438	20.6349206349	20.2947845805

Dataset 2:

Base Learner	Vanilla	Bagging	Boosting
J48	26.171875	25.1953125	25.1302083333
Logistic Regression	24.4791666667	24.6614583333	24.8372395833
Decision Stump	25.6944444444	25.0868055556	24.8119212963

Dataset 3:

Base Learner	Vanilla	Bagging	Boosting
J48	8.547008547	11.0398860399	11.2739112739
Logistic Regression	9.8290598291	11.0541310541	11.3247863248
Decision Stump	12.3456790123	12.1082621083	10.8578664134

Iteration Number: 150

Dataset 1:

Base Learner	Vanilla	Bagging	Boosting
J48	22.4489795918	21.4285714286	20.8940719145
Logistic Regression	20.4081632653	20.6802721088	20.5357142857
Decision Stump	21.3151927438	20.6349206349	20.3703703704

Dataset 2:

Base Learner	Vanilla	Bagging	Boosting
J48	26.171875	25.09765625	25.1488095238
Logistic Regression	24.4791666667	24.609375	24.853515625
Decision Stump	25.6944444444	25	24.7974537037

Dataset 3:

Base Learner	Vanilla	Bagging	Boosting
J48	8.547008547	11.0398860399	11.2332112332
Logistic Regression	9.8290598291	11.1111111111	11.2891737892
Decision Stump	12.3456790123	12.1082621083	10.7628996518

Question 1:
Which algorithm+dataset combination is improved by bagging
J48+dataset1, J48+dataset2, Decision Stump+dataset1, Decision Stump+dataset2, Decision Stump+dataset3

Question 2:
Which algorithm+dataset combination is improved by boosting
J48+dataset1, J48+dataset2, Decision Stump+dataset1, Decision Stump+datatset2, Decision Stump+dataset3

Question 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?
Bagging can reduce bias, boosting can reduce both bias and variance. From the experiment results, we can see that the performance of Logistic Regression is not improved by bagging and boosting.

3. K-means

Compression Ratios for Penguins.jpg (length of color vector: 786432):
K=2 : 3.125%
K=5 : 9.376%
K=10 : 12.50%
K=15 : 12.50%
K=20 : 15.63%

Compression Ratios for Penguins.jpg (length of color vector: 786432):
K=2 : 3.125%

K=5 : 9.376%
K=10 : 12.50%
K=15 : 12.50%
K=20 : 15.63%

Using Penguins.jpg as example, show these images, K=2, 5, 10, 15, 20:

We can see K=10 is the best. K's best value depends on the color distribution of the images, and should be chosen by experiment and artificially, for images with small number of color clusters, K should be small, while for images with large number of color clusters, K should be large to keep a good quality of the compressed images.