Machine Learning 6375.002

Assignment 3

Problem 1:

Implementation of Email classifier using perceptron algorithm and comparing the accuracy of the classifier with that of logistic regression and naïve bayes.

Experimental Results:

Naive bayes Classifier:

Including stop words:

Spam Accuracy	Ham Accuracy
85.38 %	97.70 %

After removing stop words:

Spam Accuracy	Ham Accuracy
86.15 %	97.12 %

Perceptron Classifier

Learning Rate	Iterations	Spam Accuracy (With stop words)	Ham Accuracy (With stop words)	Spam Accuracy (Without stop words)	Ham Accuracy (Without stop words)
0.001	25	87.69 %	94.54%	83.07%	95.97%
0.025	50	90.76 %	93.39%	79.23%	97.12%
0.050	75	90.76 %	93.39%	79.23%	
0.075	100	90.76 %	95.40 %	83.07 %	96.55%
0.090	110	90.76 %	94.25 %	86.15 %	95.40 %
0.095	125	93.84 %	94.25 %	78.46 %	96.83 %
0.1	135	90.76 %	93.39 %	79.23 %	97.12%
0.2	150	90.76 %	93.39 %	79.23 %	97.12 %
0.3	160	90.76 %	95.40 %	83.07 %	96.55 %
0.4	175	90.76 %	93.39 %	79.23 %	97.12 %
0.5	200	92.30 %	95.11 %	85.38%	95.11%
0.7	220	91.53%	94.54%	77.69%	97.41%
0.8	200	90.76%	93.39%	79.23%	97.12%
0.46	165	93.07%	93.39%	84.61%	95.97%
1	300	91.53%	95.11%	82.3%	96.2%
0.0001	300	91.53%	93.39%	92.30%	95.40%
0.0002	350	91.53%	93.39%	92.30%	95.40%

0.00002	400	93.07%	93.37%	88.46%	95.56%
0.00015	350	91.53%	93.39%	92.30%	95.40%
0.003	500	86.15%	94.54%	89.23%	96.26%

Logistic Regression:

Learning Rate	Regularization Factor	Iterations	Spam Accuracy (with stopwords)	Ham Accuracy (with stopwords)	Spam Accuracy (Without stop words)	Ham Accuracy (Without stop words)
0.25	0.1	150	86.15%	95.68%	86.92%	96.55%
0.30	0.2	160	88.46%	94.25%	92.30%	94.82%
0.27	0.15	170	88.46%	93.30%	91.24%	94.40%
0.35	0.35	175	86.15%	95.68%	84.61%	97.12%
0.25	0.5	100	6.15%	95.43%	92.34%	92.24%
0.5	0.5	150	76.9 %	95.57%	80%	97.98%
0.5	0.1	175	84.35%	95.11%	91.53%	94.82%
0.56	0.18	180	86.62%	96.4%	90.30%	94.82%
0.75	0.28	160	82.45%	90.87%	84.61%	97.41
0.80	0.35	180	85.67%	93.48%	87.69%	95.42%
0.45	0.43	200	73.4%	88.62%	80.76%	97.7%
0.64	0.2	190	77.8%	86.34%	82.3%	90.5%
0.025	0.01	190	76.92%	96.26%	13.07%	99.76%
0.8	0.8	250	53%	87.26%	63.24%	94.35%
0.34	0.24	165	88.24%	94.32%	92.46%	94.82%
0.25	0.1	50	83.84%	95.4%	85.62%	97.12%
0.30	0.2	50	88.46%	93.67%	79.23%	97.41%
0.0025	0.4	25	3.84%	98.67%	65.38%	89.36%
0.025	0.45	40	43.27%	98.27%	32.30%	99.42%

Conclusion.

The experimental results of Naïve bayes, logistic regression and perceptron classifier for Spam filtering are shown. Different values of iterations and learning rate are tested out and it was found out the perceptron and logistic regression classifier performed well.

K-means clustering:

Image compression is done using k-means.

Experimental Results:

K value	Penguin Image Compression ratio	Koala Image Compression ratio
2	9.03	5.9
5	7.16	4.43
10	6.54	4.7
15	6.65	4.88
20	6.71	4.91
50	6.77	5.11
75	6.83	5.18
100	6.83	5.25
125	6.83	5.25

Findings:

The compressed images can be found in the compressed images directory after unzipping the ML_Assignment3.zip.

Average Compression ratio for Penguin: 7.03

Variance of Compression ratio for Penguin: 0.52

Average Compression ratio for Koala: 5.06

Variance of Compression ratio for Koala: 0.15

The best K value for penguin: 125

The best K value for koala: 100

Conclusion:

As the compression ratio increases, the image quality decreases. This is the trade off found between image quality and compression ratio.