COL774 : Assignment 2

Aditi Singla
2014CS50277

11th March 2017

1. Text Classification

   (a) Naïve Bayes Classifier: The accuracies observed were as follows:

   - Over Training Data : 97.192342753 % (5331/5485)
   - Over Test Data : 95.4773869347 % (2090/2189)

   (b) Accuracies for Random Majority Baseline on test data:

   - Random Baseline : 12.8369118319% (281/2189)
   - Majority Baseline : 49.4746459571% (1083/2189)

   The accuracy obtained by Naïve Bayes Classifier is far better than the above.

   (c) Confusion Matrix: We can clearly observe that the original class and predicted class match most of the time. So, the results become more inaccurate for the classes with less training data.
   The class 'Earn' has the highest value of the diagonal-entry in the confusion matrix. This is again because a lot of training data has it as a label.
   The class 'Interest' is confused with 'money-fx' a lot of times (23.45%). (19 times wrongly as money-fx and 54 times rightly as interests) The confusion matrix for the given data is as follows:

		Predicted Class							
		Earn	Money-Fx	Trade	Acq	Grain	Interest	Crude	Ship
Actual Class	Earn	1056	0	2	24	0	0	1	0
	Money-Fx	1	80	5	0	0	1	0	0
	Trade	1	1	72	0	0	0	1	0
	Acq	2	1	3	689	0	0	1	0
	Grain	1	0	3	0	4	0	2	0
	Interest	0	19	8	0	0	54	0	0
	Crude	0	0	3	0	0	0	118	0
	Ship	0	0	7	3	0	0	9	17

Figure 1: Confusion Matrix without Stemming & Stopword Removal

   (d) Stopword Removal and Stemming:

   - Accuracy on the training data: 97.192342753 % (5331/5485)
   - Accuracy on the test data: 95.705801736 % (2095/2189)
   - Confusion matrix for the training data:

		Predicted Class							
		Earn	Money-Fx	Trade	Acq	Grain	Interest	Crude	Ship
Actual Class	Earn	1055	0	2	24	0	1	1	0
	Money-Fx	1	82	3	0	0	1	0	0
	Trade	1	1	72	0	0	0	1	0
	Acq	5	2	3	685	0	0	1	0
	Grain	0	0	2	0	7	0	1	0
	Interest	0	20	5	0	0	56	0	0
	Crude	0	0	3	2	0	0	116	0
	Ship	0	0	6	3	0	0	5	22

Figure 2: Confusion Matrix with Stemming & Stopword Removal

- We observe no change in the accuracy over training data but a slight increase over testing data. A very few entries change in the confusion matrix.

2. Facial Attractiveness Classification

   (a) List of 9 Support Vectors for Linear Kernel:

   - 15 [230.477821106]
   - 18 [415.857392485]
   - 33 [236.953397652]
   - 110 [2.93085389314]
   - 120 [35.4879941058]
   - 150 [224.576172743]
   - 152 [474.005525269]
   - 248 [299.007203779]
   - 278 [368.551562993]

   (b) The value of b for Linear Kernel is -1.832 & the accuracy obtained is 61.67 % (74/120).

   (c) The value of b for Gaussian Kernel is -6.116 & the accuracy obtained is 67.5 % (81/120). List of 42 Support Vectors for Linear Kernel:

   - 3 [147.236931796]
   - 9 [247.384255528]
   - 11 [161.416721856]
   - 12 [380.220816175]
   - 23 [145.532096014]
   - 34 [443.749454426]
   - 51 [453.185078564]
   - 55 [282.160421689]
   - 59 [491.589726827]
   - 60 [407.114783819]
   - 63 [230.317484061]
   - 67 [440.182042574]
   - 69 [141.227570491]

- 70 [192.826454929]
- 82 [305.710538548]
- 91 [290.53485948]
- 95 [329.312064562]
- 107 [157.868702586]
- 110 [61.2275237613]
- 116 [135.761014136]
- 122 [235.257073944]
- 123 [423.341381983]
- 128 [215.987204922]
- 130 [377.777219612]
- 152 [122.596744893]
- 158 [141.957576084]
- 161 [26.0306849757]
- 162 [481.260309416]
- 167 [329.503278399]
- 198 [364.824057461]
- 199 [235.873135478]
- 216 [90.9562539599]
- 236 [456.45523821]
- 241 [105.453578849]
- 248 [315.180121094]
- 257 [353.989531053]
- 261 [95.6323680836]
- 263 [461.411839882]
- 264 [68.7961627526]
- 269 [77.6955676424]
- 271 [9.68642070999]
- 278 [346.585781307]

The results obtained for Linear SVM are better than the linear SVM. The Gaussian Kernel fits the data better than the Linear SVM.

(d) The observations made using LIBSVM Library are:

- Linear Kernel: Accuracy: 61.67% (74/120)
- Gaussian Kernel: Accuracy: 67.5% (81/120)
- We can clearly observe same accuracies for both the cases: Linear and Gaussian Kernel. This shows that both the libraries use the same optimisation. But the number of support vectors are way too high in LIBSVM (Linear: 269 & Gaussian: 256) as compared to CVXPY. This could be because LIBSVM classified vectors with alpha*y 500, as support vectors, which aren't included by us in case of CVXPY.

(e) Cross Validation: Values for different values of C on Gaussian Kernel:

- c=1: 56.0714%
- c=10: 56.0714%
- c=100: 56.0714%
- c=1000: 58.5714%
- c=10000: 67.1429%
- c=100000: 64.6429%
- c=1000000: 64.6429%

Test Accuracy for different values of C on Gaussian Kernel:

- c=1: 56.6667%
- c=10: 56.6667%
- c=100: 56.6667%
- c=1000: 65.0000%
- c=10000: 74.1667%
- c=100000: 76.6667%
- c=1000000: 76.6667%

The value of C with the best Cross Validation Accuracy is C = 10000, but the test accuracy keeps increasing with C, in the range that we have checked.

The graph for the observations:

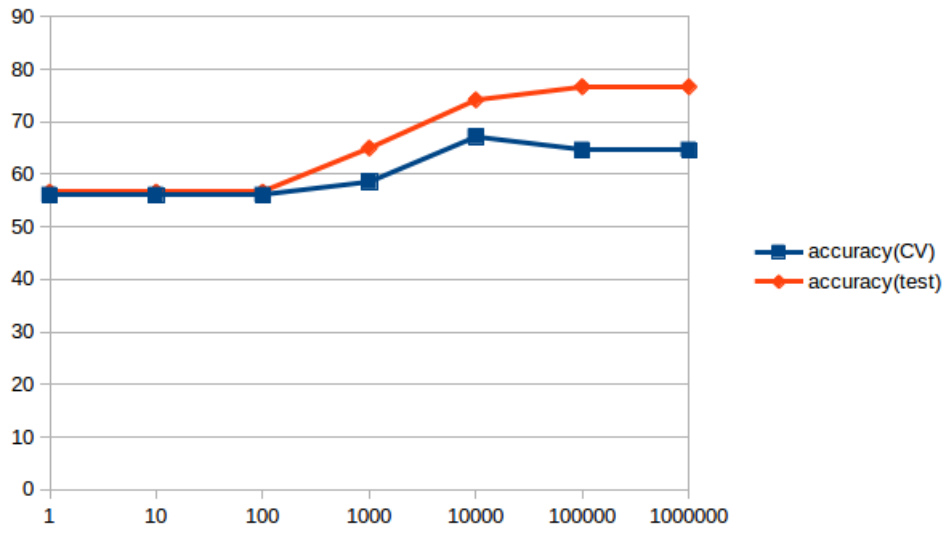

Figure 3: Graph for Cross Validation and Test Accuracies