ANKUSH PHULIA
2014CS50279

Assignment 2 – Machine Learning

TEXT CLASSIFICATION

NAÏVE BAYES

- Training Set Accuracy : 97.192%
- Test Set Accuracy : 95.477%
- Random Guess Accuracy : 12.5%   Improvement : 82.977% difference (or 7.63 times)
- Majority Guess Accuracy : 49.474%   Improvement : 46.003% difference (or 1.92 times)

CONFUSION MATRIX

guess/act	trade	earn	money-fx	crude	acq	ship	interest	grain
trade	72	1	1	1	0	0	0	0
earn	2	1056	0	1	24	0	0	0
money-fx	5	1	80	0	0	0	1	0
crude	3	0	0	118	0	0	0	0
acq	3	2	1	1	689	0	0	0
ship	7	0	0	9	3	17	0	0
interest	8	0	19	0	0	0	54	0
grain	3	1	0	2	0	0	0	4

- ‘Earn’ has the highest diagonal entry in the table, as earn had the highest number of training examples
- Fewer training examples lead to lower test accuracies
- ‘Interest’ was most confused with ‘money-fx’
- ‘Grain’ had very few test examples, and was correctly classified the least
- ‘Trade’ was the cause of the most misclassification, followed by ‘acq’. ‘Grain’ caused the least misclassification

STOP-WORD REMOVAL AND STEMMING

guess/act	trade	earn	money-fx	crude	acq	ship	interest	grain
trade	72	1	1	1	0	0	0	0
earn	2	1055	0	1	24	0	1	0
money-fx	3	1	82	0	0	0	1	0
crude	3	0	0	116	2	0	0	0
acq	3	5	2	1	685	0	0	0
ship	6	0	0	5	3	22	0	0
interest	5	0	20	0	0	0	56	0
grain	2	0	0	1	0	0	0	7

- Test Set Accuracy : 95.705%   Training Set Accuracy : 97.192%
- Accuracy increases slightly, decreases for a few classes like ‘crude’, ‘acq’, and ‘earn’, same for ‘trade’ and increases for ‘grain’, ‘interest’ and ‘ship’
- ‘trade’ causes fewer misclassifications, ‘grain’ and ‘ship’ are biggest gainers

CVXOPT PACKAGE

Linear Kernel

- Support Vectors : The support vectors were examples with indices 15, 18, 33, 110 ,120, 150, 152, 172, 248, 278
- Weight vector w : A vector of length 7396, given by the formula

$$w(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x)$$

  < -5.30797151 -3.83746177 -4.4634469 ..., -0.26758164, 1.14566318, 1.18139699>
- Intercept term b : 1.832
- Average Test Accuracy : 61.66% (74 out of 120)

Gaussian Kernel

- Support Vectors : 42 training vectors were support vectors with indices 3, 9, 11, 12, 23, 34, 51, 55, 59, 60, 63, 67, 69, 70, 82, 91, 95, 107, 110, 116, 122, 123, 128, 130, 152, 158, 161, 162, 167, 198, 199, 216, 236, 241, 248, 257, 261, 263, 264, 269, 271, 278
- Intercept term b : 6.11
- Average Test Accuracy : 67.5% (81 out of 120)

LIBSVM

Linear Kernel

- Intercept term b : 1.832
- Average Test Accuracy : 61.66% (74 out of 120)
- There is small difference in the accuracy using the linear kernel
- Apparently, many support vectors ~270+, but many were close to 500

Gaussian Kernel

- Intercept term b : 6.11
- Average Test Accuracy : 67.5% (81 out of 120)
- The accuracy using linear kernel is 61.66% and using Gaussian kernel is 67.5%.
- The accuracies are the same in case of Gaussian kernel is the same. This is mainly because in cvx we using convex optimization methods and in libvsm, we use the SMO algorithm.
- Apparently, many support vectors ~260+, but many were close to 500

Cross Validation

- The value of C for which we get best test data accuracy is C > 105, however for Cross Validation Accuracy, C = 104 did the best.

Table 3: Accuracies

C value	Cross Validation Accuracy	Test Data Accuracy
1	52.1429%	56.6667% (68/120)
10	52.1429%	56.6667% (68/120)
10^2	51.7857%	61.6667% (74/120)
10^3	61.7857%	72.5% (87/120)
10^4	67.1429%	75.8333% (91/120)
10^5	65%	76.6667% (92/120)
10^6	65%	76.6667% (92/120)

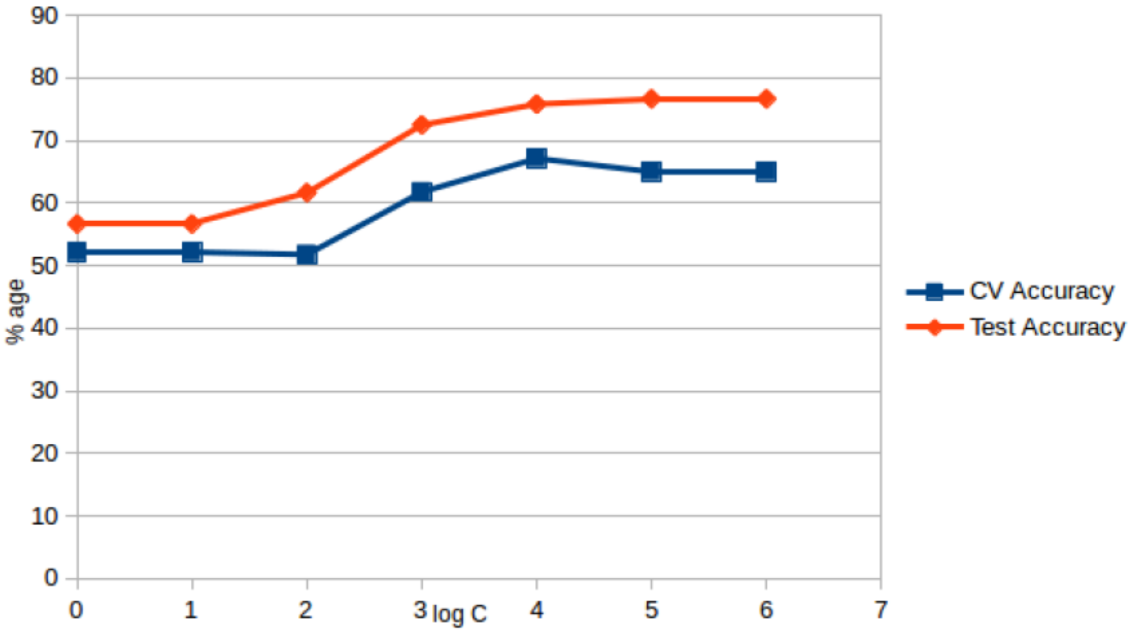

Figure 1: Accuracies