EC500 - Introduction to Learning From Data

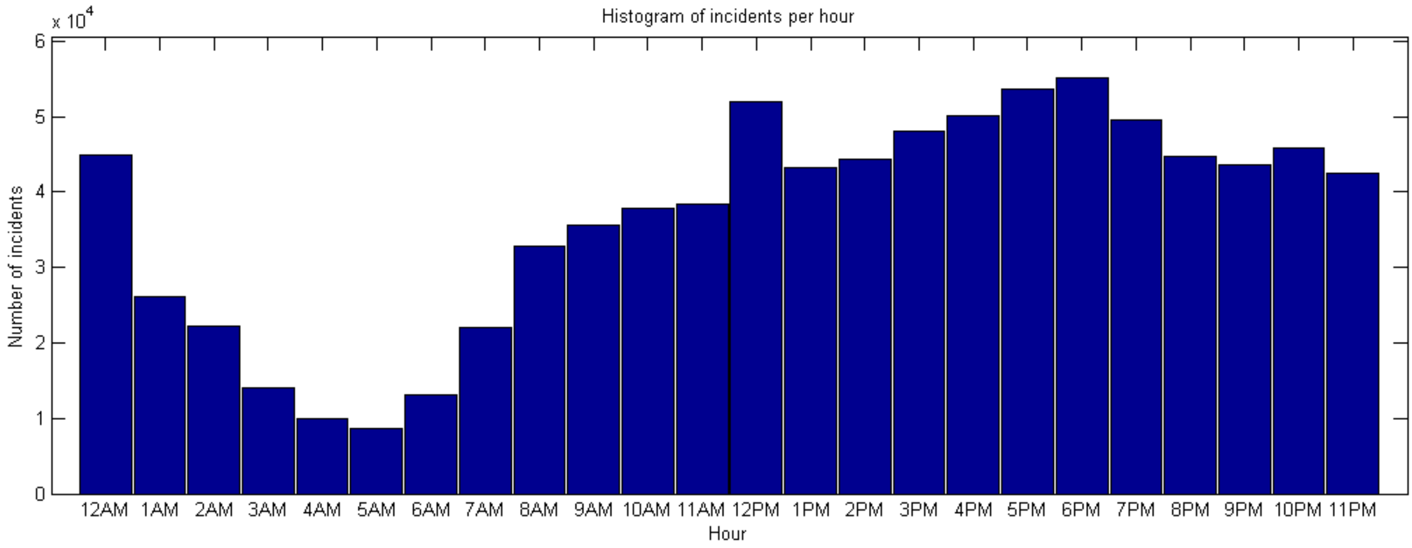
Matlab - 2

Mikhail Andreev

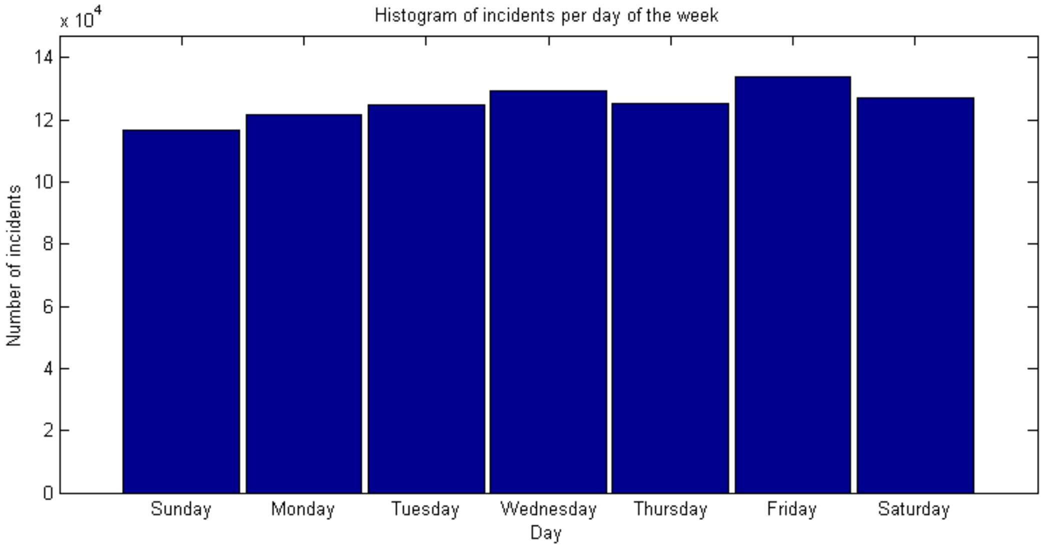
November 3, 2015

1 San Francisco Crime Prediction

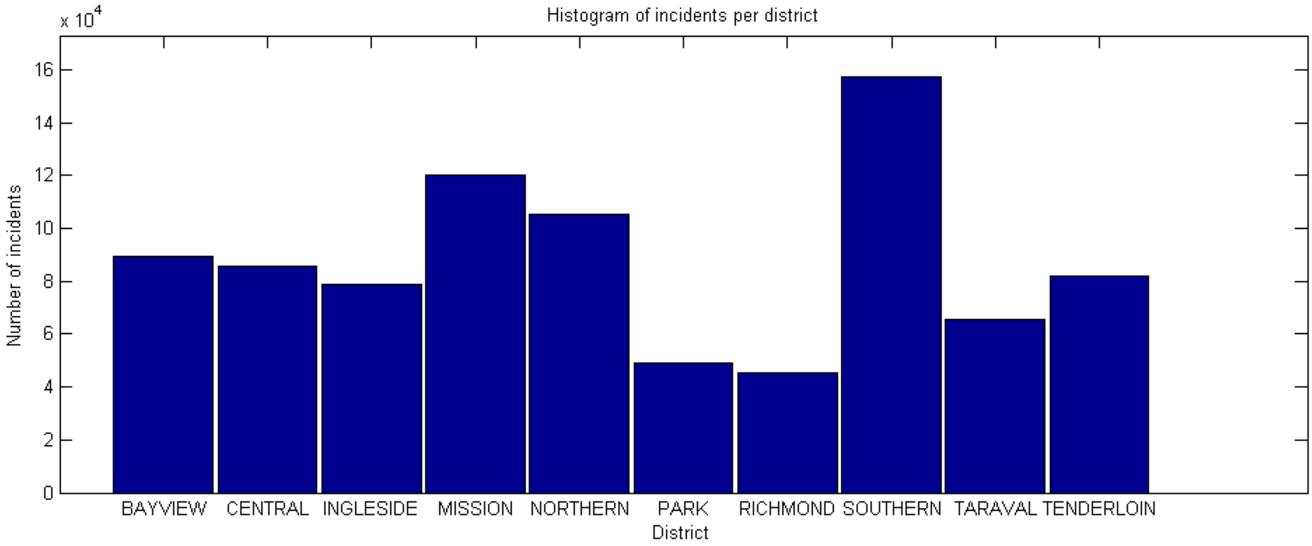
1.1 Part a



Here we can see the number of incidents that occur at each hour of the day. There is a clear drop in the number of incidents as the night progresses, steadily rising through the morning. At around lunch time the number of incidents spikes drastically, then falls again. After that the number of incidents steadily rises until the peak at 6PM, then begins to fall.

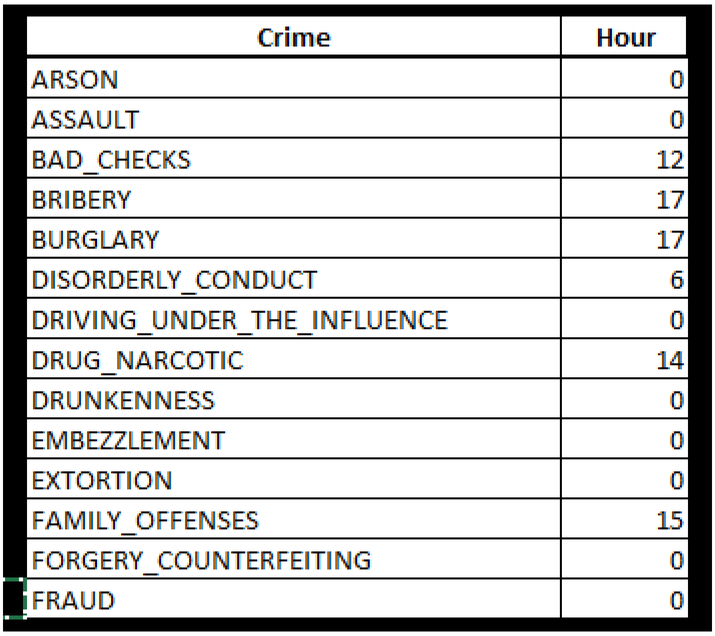


Here we have the distribution of the number of incidents each day. There is a slight bump in the number of incidents on Friday, but there are no significant changes.

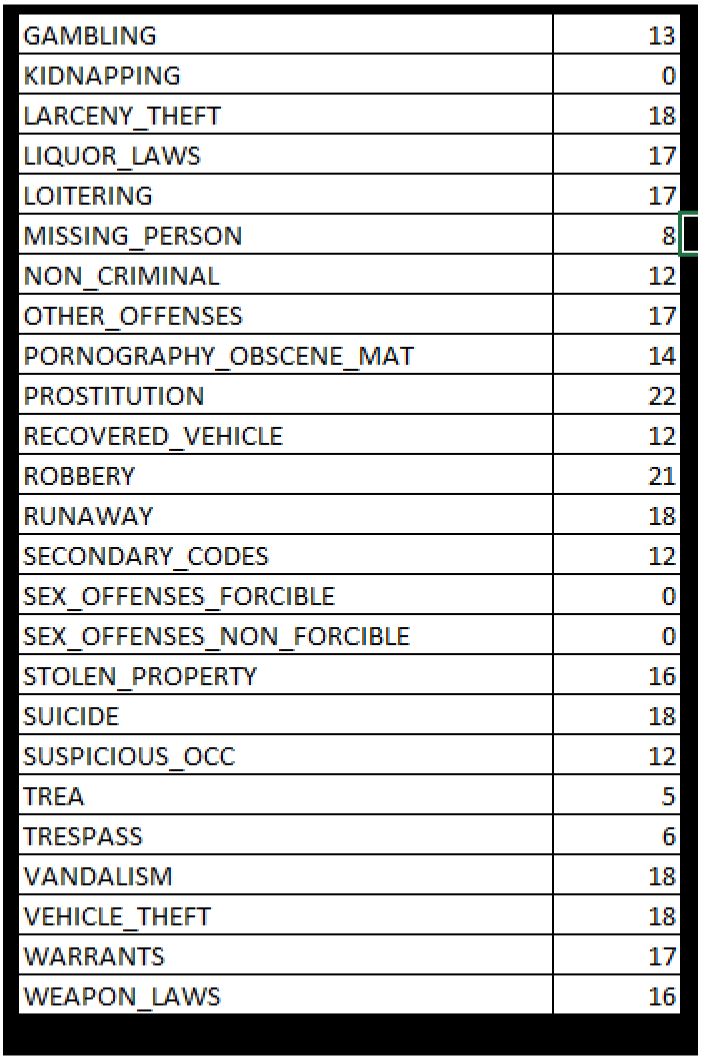


This shows the distribution of the incidents in each district. The distribution shows which parts of the city may be more dangerous.

The list of crimes with the hour they are most likely to occur.



Crime	Hour
ARSON	0
ASSAULT	0
BAD_CHECKS	12
BRIBERY	17
BURGLARY	17
DISORDERLY_CONDUCT	6
DRIVING_UNDER_THE_INFLUENCE	0
DRUG_NARCOTIC	14
DRUNKENNESS	0
EMBEZZLEMENT	0
EXTORTION	0
FAMILY_OFFENSES	15
FORGERY_COUNTERFEITING	0
FRAUD	0



GAMBLING	13
KIDNAPPING	0
LARCENY_THEFT	18
LIQUOR_LAWS	17
LOITERING	17
MISSING_PERSON	8
NON_CRIMINAL	12
OTHER_OFFENSES	17
PORNOGRAPHY_OBSCENE_MAT	14
PROSTITUTION	22
RECOVERED_VEHICLE	12
ROBBERY	21
RUNAWAY	18
SECONDARY_CODES	12
SEX_OFFENSES_FORCIBLE	0
SEX_OFFENSES_NON_FORCIBLE	0
STOLEN_PROPERTY	16
SUICIDE	18
SUSPICIOUS_OCC	12
TREA	5
TRESPASS	6
VANDALISM	18
VEHICLE_THEFT	18
WARRANTS	17
WEAPON_LAWS	16

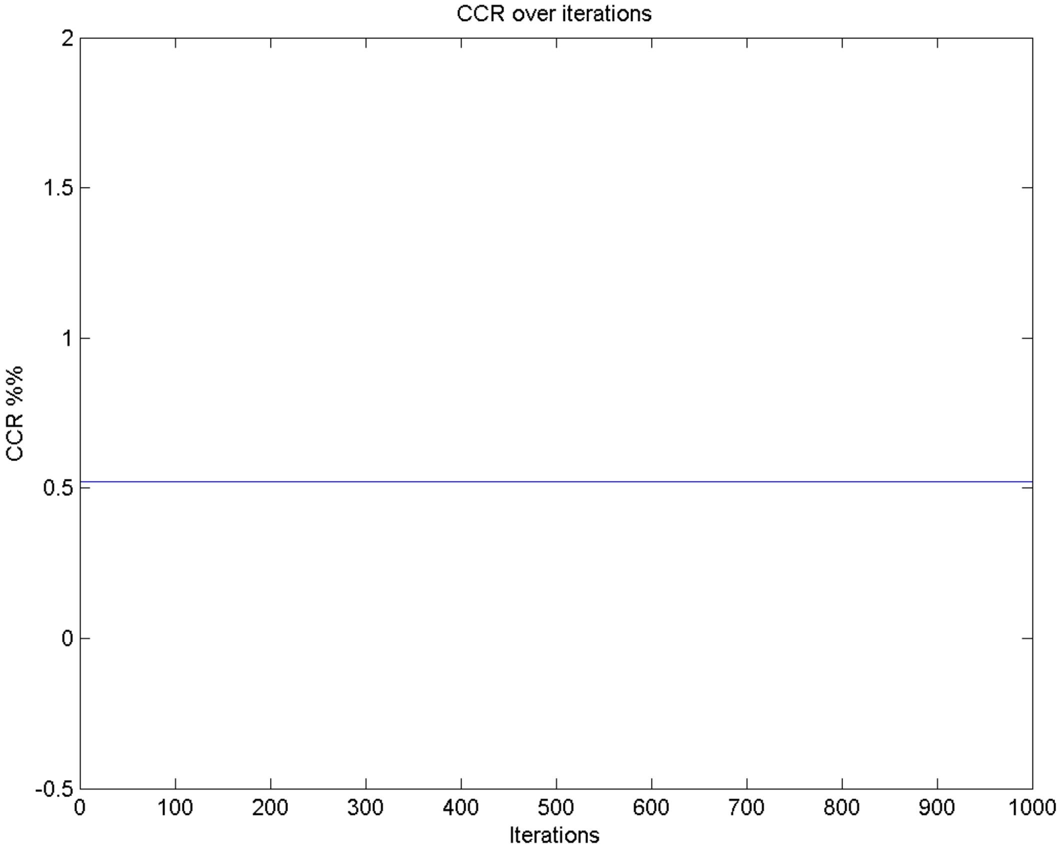
Here is the most likely hour that each crime is going to occur. The most likely hours that crimes occur are around 10PM, 0AM, 12PM, and 6PM.

The crime most likely to occur in Bayiew is other offenses
The crime most likely to occur in Central is larceny and theft
The crime most likely to occur in Ingleside is other offenses
The crime most likely to occur in Mission is other offenses
The crime most likely to occur in Northern is larceny and theft
The crime most likely to occur in Park is larceny and theft
The crime most likely to occur in Richmond is larceny and theft
The crime most likely to occur in Southern is larceny and theft
The crime most likely to occur in Taraval is larceny and theft

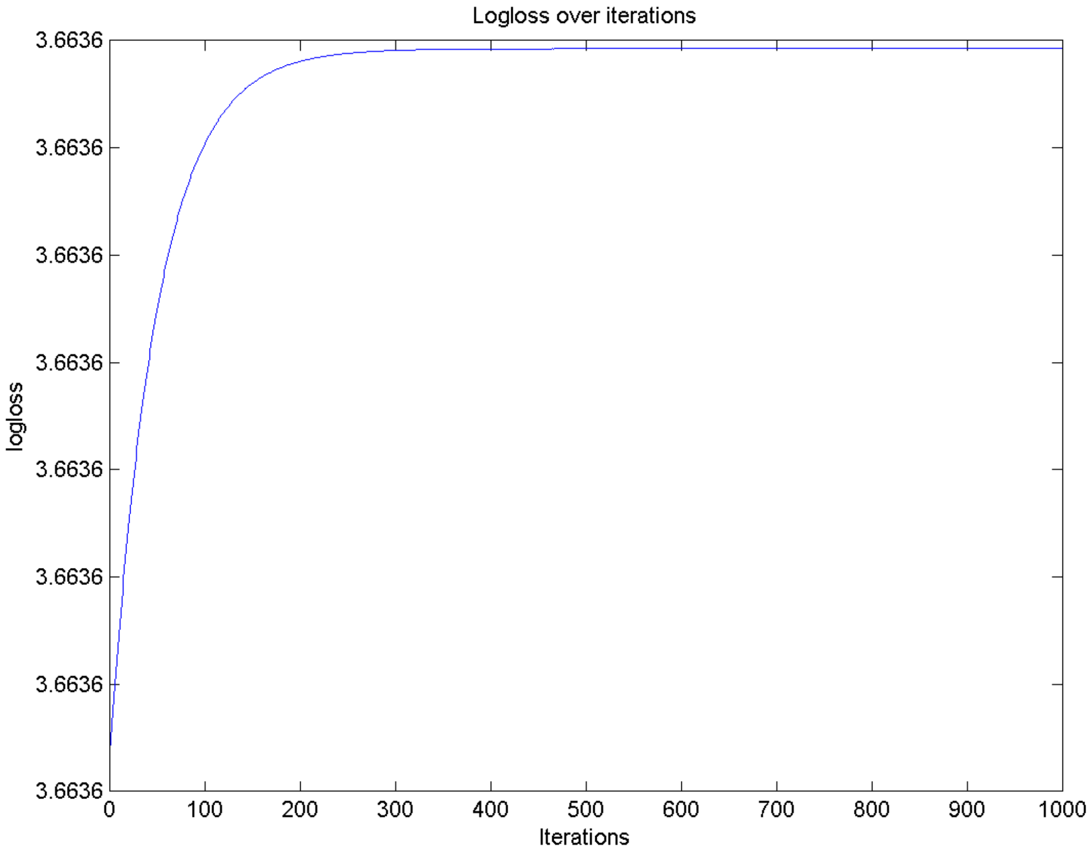
The crime most likely to occur in Tenderloin is drug and narcotic

From this spread, we can see that larceny and theft is the most likely incident that occurs in many regions.

1.2 Part b



This graph indicates that the CCR rate did not noticeably change over the different iterations of the parameters. Unfortunately, this also indicates the CCR rate was very low for the experiment. The end result was only .5% correct detection.

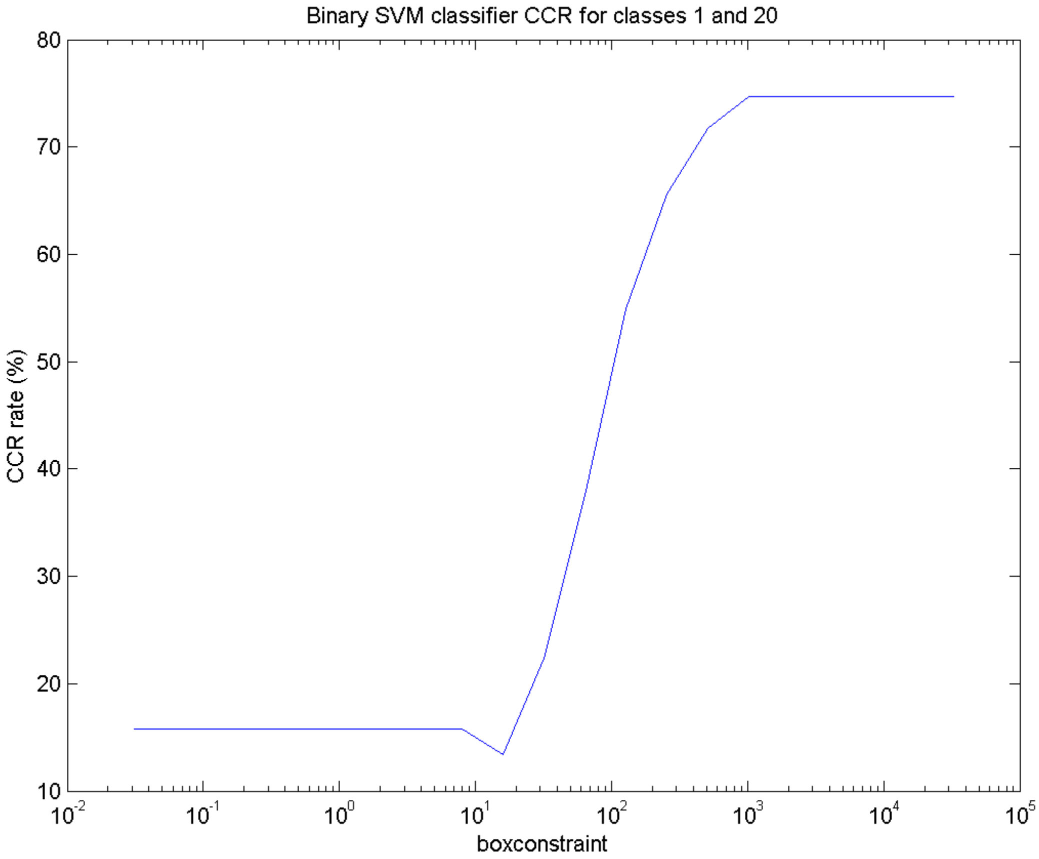


As iterations increase, the logloss value approaches an asymptote of around 3.6636, and levels off.

1.3 Part c

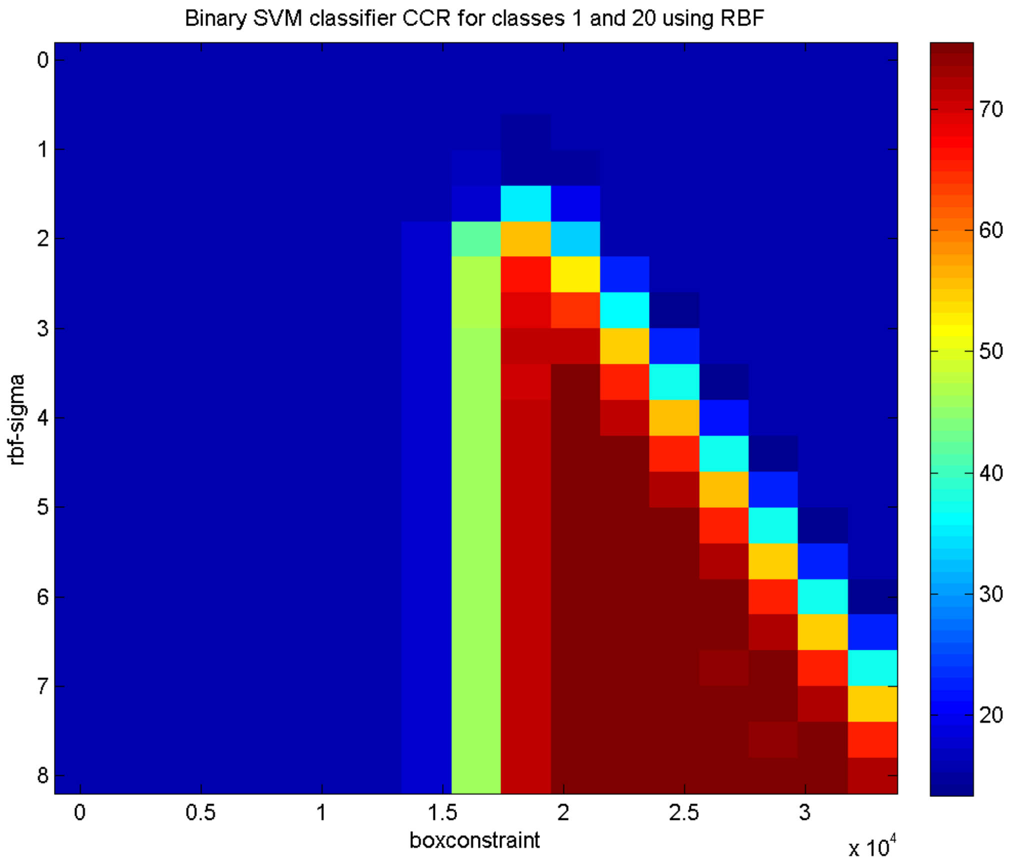
2 SVM Classifier for Text Documents

2.1 Part a



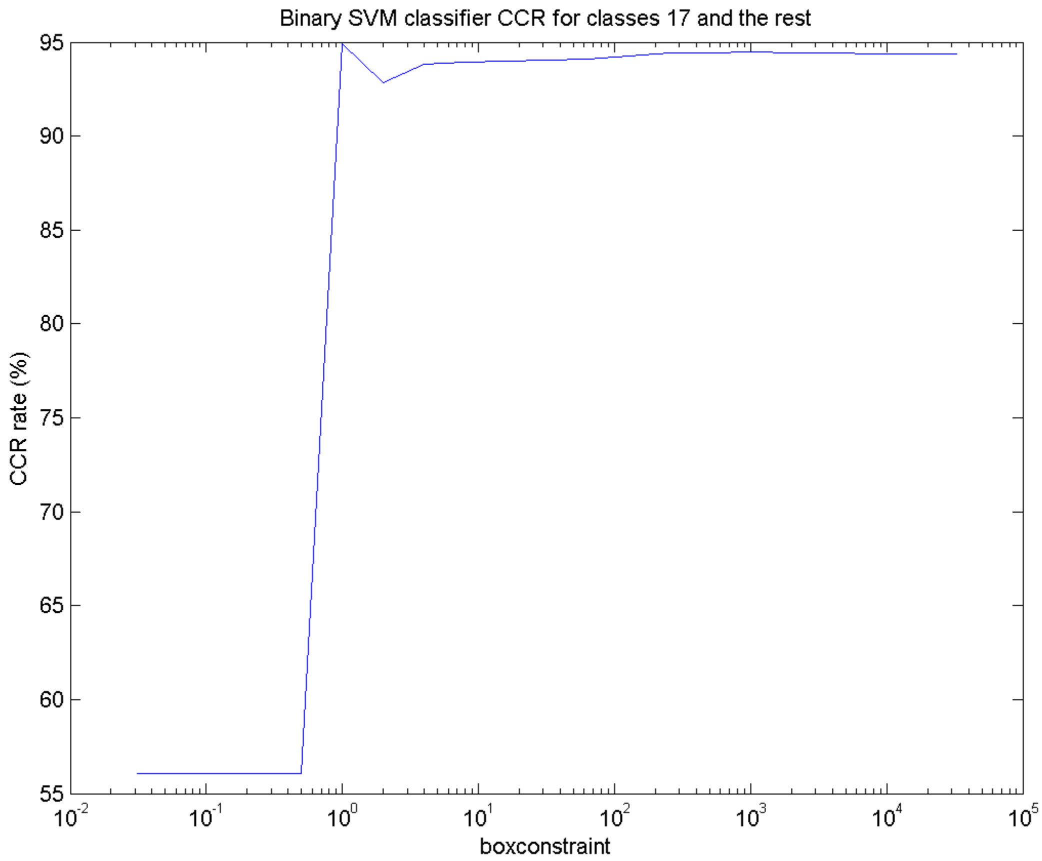
The best CCR was 80.32% which was achieved when $C^* = 2^{10}$. This can be clearly seen from the curve, where the CCR rate increases with boxconstraint. Eventually it hits a plateau value, after which boxconstraint does not improve the CCR.

2.2 Part b



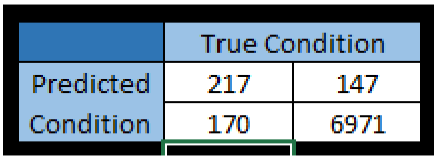
The best CCR was 79.44% which was achieved when $C^* = 2^{11}$ and rbf-sigma=2^{-3}. The graph shows the region in which the combined boxconstraint and rbf-sigma values produce an increased CCR. Outside this region, the CCR drops dramatically.

2.3 Part c



The best CCR was 95.78% which was achieved when $C^* = 2^0$. The graph shows the sharp rise and peak as the boxconstraint rises, which then falls off after C^*.

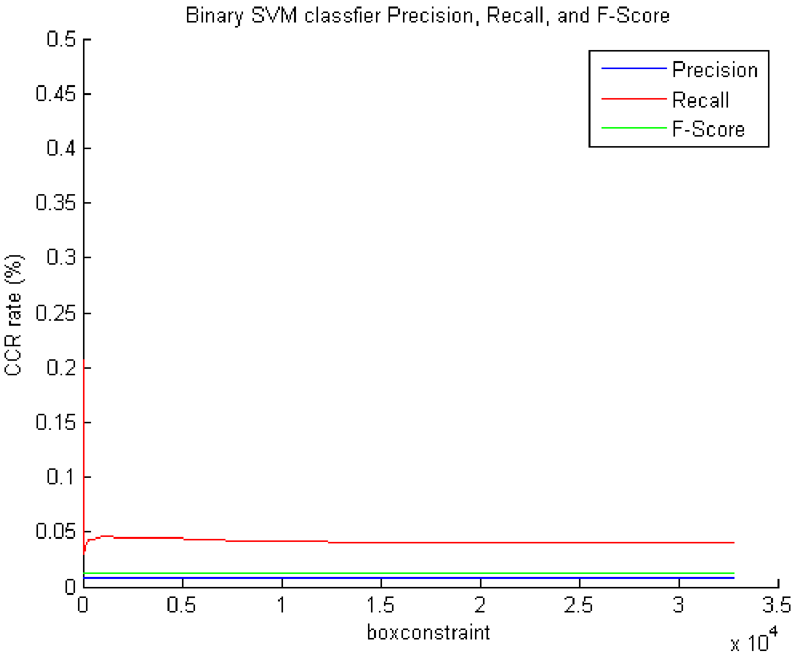
Confusion Matrix:



	True Condition	
Predicted	217	147
Condition	170	6971

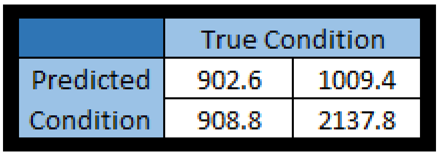
In the confusion matrix it is clear that there are many more negative samples than positive ones, however, the vast majority have been properly classified as being not class 17. The accuracy of this classification is likely due to the sheer number of negative samples which allow for accurate prediction.

2.4 Part d



Best value of C as determined by the precision and the F-score is 2^0. In the graph we can see the recall and precision, and in turn the F-score values are fairly constant throughout the different boxconstraint values.

Confusion Matrix:



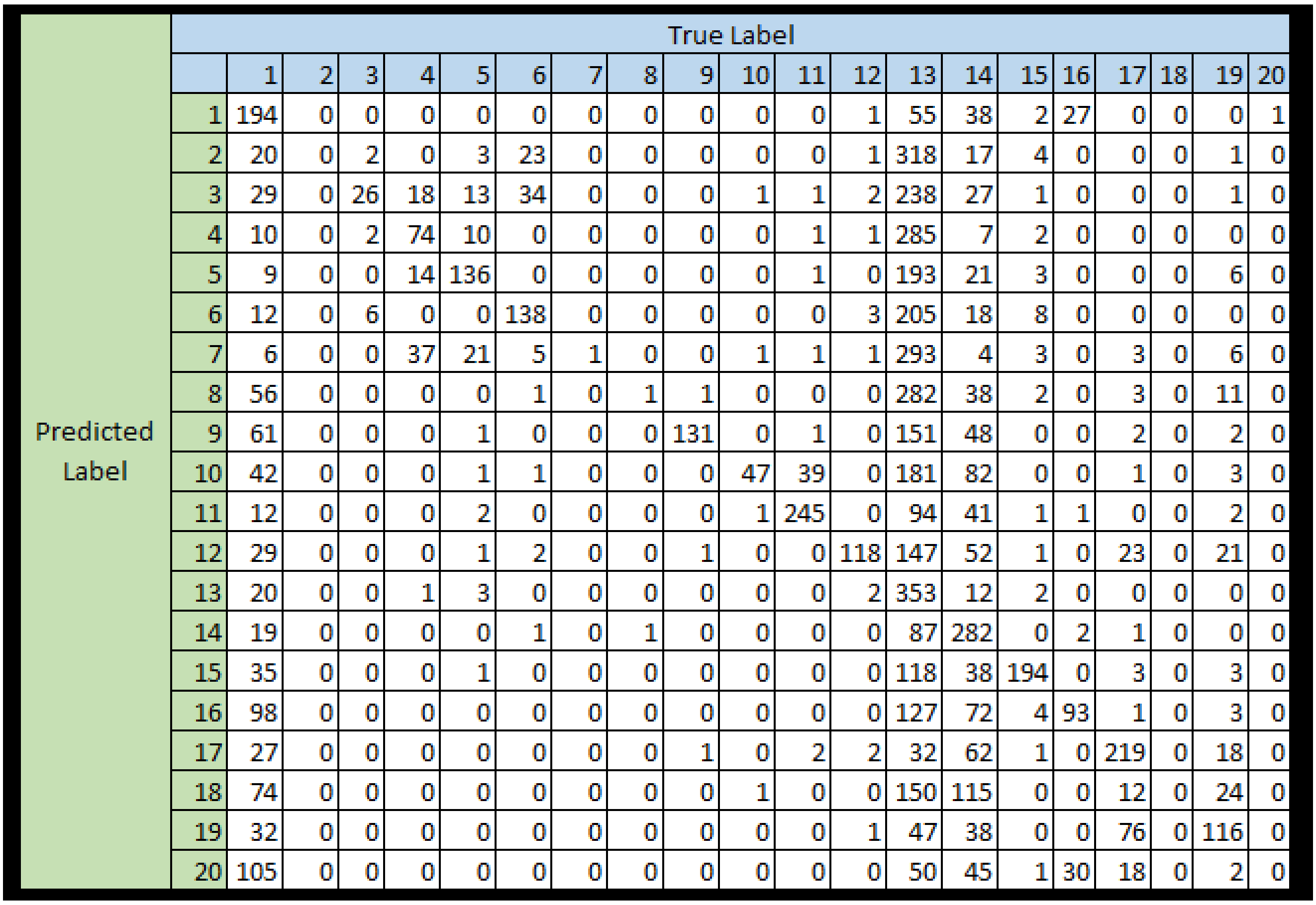
	True Condition	
Predicted Condition	902.6	1009.4
	908.8	2137.8

Since the best F-score and precision value occur at the same value of the boxconstraint, there is one confusion matrix. As can be seen there is a large amount of misclassification, and only the classification of samples as not being in class 17 can be said to be more accurate. This seems to indicate that using the CCR as the determinant of the optimal C^* value is the correct approach.

2.5 Part e

The overall CCR is 31.55%. This can likely be improved with a better boxconstraint. The training time was 100.68s, and the testing time was 98.48s.

The confusion matrix is:

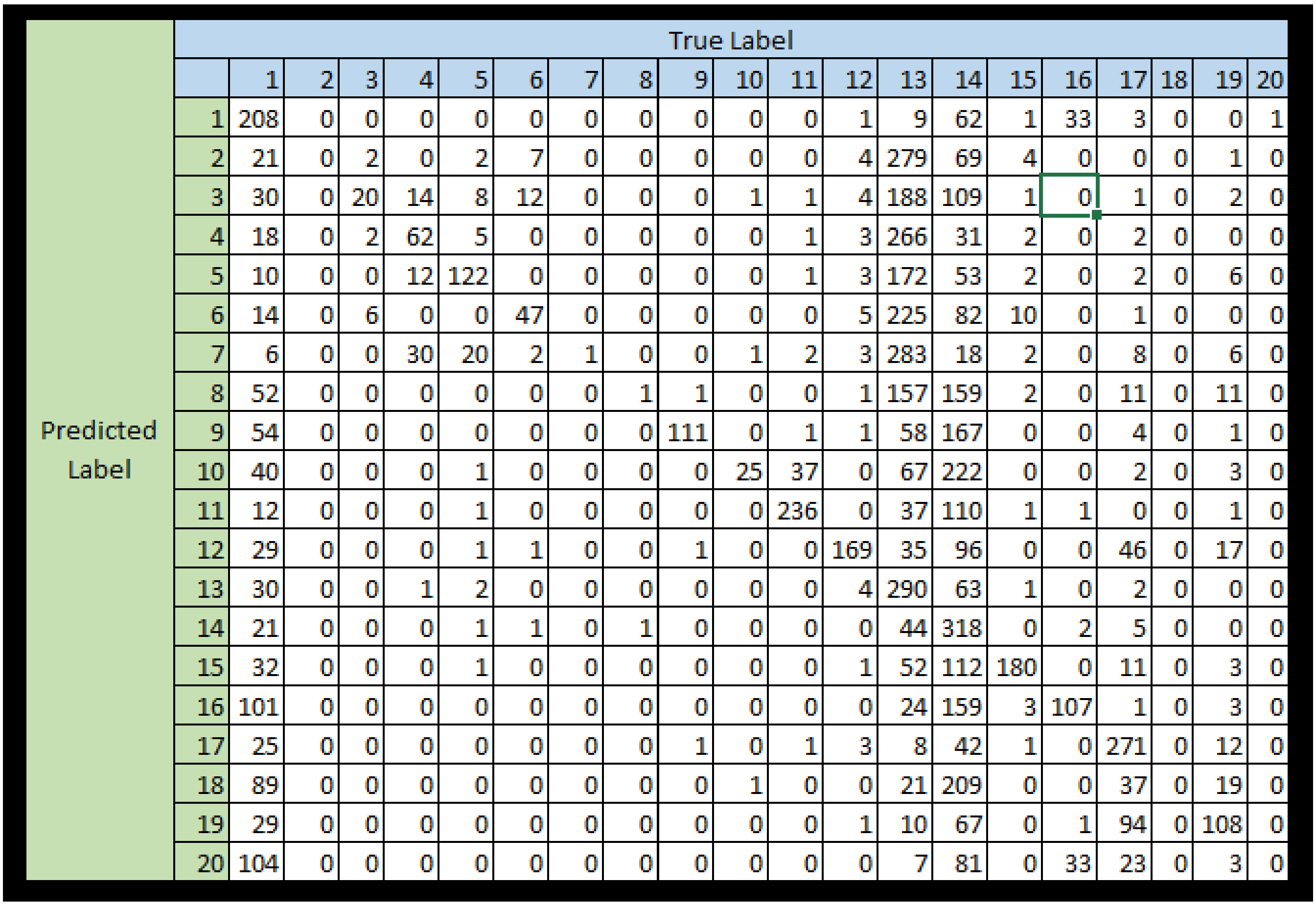


		True Label																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Predicted Label	1	194	0	0	0	0	0	0	0	0	0	0	1	55	38	2	27	0	0	0	1
	2	20	0	2	0	3	23	0	0	0	0	0	1	318	17	4	0	0	0	1	0
	3	29	0	26	18	13	34	0	0	0	1	1	2	238	27	1	0	0	0	1	0
	4	10	0	2	74	10	0	0	0	0	0	1	1	285	7	2	0	0	0	0	0
	5	9	0	0	14	136	0	0	0	0	0	1	0	193	21	3	0	0	0	6	0
	6	12	0	6	0	0	138	0	0	0	0	0	3	205	18	8	0	0	0	0	0
	7	6	0	0	37	21	5	1	0	0	1	1	1	293	4	3	0	3	0	6	0
	8	56	0	0	0	0	1	0	1	1	0	0	0	282	38	2	0	3	0	11	0
	9	61	0	0	0	1	0	0	0	131	0	1	0	151	48	0	0	2	0	2	0
	10	42	0	0	0	1	1	0	0	0	47	39	0	181	82	0	0	1	0	3	0
	11	12	0	0	0	2	0	0	0	0	1	245	0	94	41	1	1	0	0	2	0
	12	29	0	0	0	1	2	0	0	1	0	0	118	147	52	1	0	23	0	21	0
	13	20	0	0	1	3	0	0	0	0	0	0	2	353	12	2	0	0	0	0	0
	14	19	0	0	0	0	1	0	1	0	0	0	0	87	282	0	2	1	0	0	0
	15	35	0	0	0	1	0	0	0	0	0	0	0	118	38	194	0	3	0	3	0
	16	98	0	0	0	0	0	0	0	0	0	0	0	127	72	4	93	1	0	3	0
	17	27	0	0	0	0	0	0	0	1	0	2	2	32	62	1	0	219	0	18	0
	18	74	0	0	0	0	0	0	0	0	1	0	0	150	115	0	0	12	0	24	0
	19	32	0	0	0	0	0	0	0	0	0	0	1	47	38	0	0	76	0	116	0
	20	105	0	0	0	0	0	0	0	0	0	0	0	50	45	1	30	18	0	2	0

2.6 Part f

The overall CCR is 30.33%. The training time was 83.31s, and the testing time was 201.37s.

The confusion matrix is:



		True Label																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Predicted Label	1	208	0	0	0	0	0	0	0	0	0	0	1	9	62	1	33	3	0	0	1
	2	21	0	2	0	2	7	0	0	0	0	0	4	279	69	4	0	0	0	1	0
	3	30	0	20	14	8	12	0	0	0	1	1	4	188	109	1	0	1	0	2	0
	4	18	0	2	62	5	0	0	0	0	0	1	3	266	31	2	0	2	0	0	0
	5	10	0	0	12	122	0	0	0	0	0	1	3	172	53	2	0	2	0	6	0
	6	14	0	6	0	0	47	0	0	0	0	0	5	225	82	10	0	1	0	0	0
	7	6	0	0	30	20	2	1	0	0	1	2	3	283	18	2	0	8	0	6	0
	8	52	0	0	0	0	0	0	1	1	0	0	1	157	159	2	0	11	0	11	0
	9	54	0	0	0	0	0	0	0	111	0	1	1	58	167	0	0	4	0	1	0
	10	40	0	0	0	1	0	0	0	0	25	37	0	67	222	0	0	2	0	3	0
	11	12	0	0	0	1	0	0	0	0	0	236	0	37	110	1	1	0	0	1	0
	12	29	0	0	0	1	1	0	0	1	0	0	169	35	96	0	0	46	0	17	0
	13	30	0	0	1	2	0	0	0	0	0	0	4	290	63	1	0	2	0	0	0
	14	21	0	0	0	1	1	0	1	0	0	0	0	44	318	0	2	5	0	0	0
	15	32	0	0	0	1	0	0	0	0	0	0	1	52	112	180	0	11	0	3	0
	16	101	0	0	0	0	0	0	0	0	0	0	0	24	159	3	107	1	0	3	0
	17	25	0	0	0	0	0	0	0	1	0	1	3	8	42	1	0	271	0	12	0
	18	89	0	0	0	0	0	0	0	0	1	0	0	21	209	0	0	37	0	19	0
	19	29	0	0	0	0	0	0	0	0	0	0	1	10	67	0	1	94	0	108	0
	20	104	0	0	0	0	0	0	0	0	0	0	0	7	81	0	33	23	0	3	0