Matlab - 1

Mikhail Andreev

October 9, 2015

1 Gaussian Discriminant Analysis

1.1 Part a

Helper functions located in online submission.

1.2 Part b

Mean vectors:

Average mean vector for each split				
u_1	5.0117	3.4261	1.4584	0.2443
u_2	5.9366	2.7642	4.2538	1.3238
u_3	6.5768	2.9674	5.5399	2.0263

Variances for QDA:

Average variances for QDA for each split				
class = 1	0.1233	0.1379	0.0285	0.0101
class = 2	0.2556	0.0940	0.2136	0.0373
class = 3	0.3781	0.0971	0.2828	0.0718

Variances for LDA:

Average variances for LDA for each split				
All classes	0.2519	0.1106	0.1744	0.0394

CCR Results:

	CCR Statistics			
Split	LDA Mean	QDA Mean	LDA Variance	QDA Variance
1	0.9683	0.9383	0.0001975	0.0016
2	0.9682	0.95	0.0001331	0.0003
3	0.969	0.959	0.0002544	0.0004
4	0.9644	0.9522	0.000214	0.0011
5	0.97	0.9675	0.0002847	0.0004
6	0.97	0.9557	0.0005646	0.0004
7	0.975	0.965	0.0002006	0.0005
8	0.982	0.97	0.0003067	0.0009
9	0.9675	0.9775	0.0008403	0.0005
10	0.9733	0.97	0.0004444	0.0006

Confusion matrix of best split in LDA:

Confusion Matrix Best CCR			
Labels	1	2	3
1	15.2	0	0
2	0	16.5	0.5
3	0	0.4	17.4

Confusion matrix of worst split in LDA:

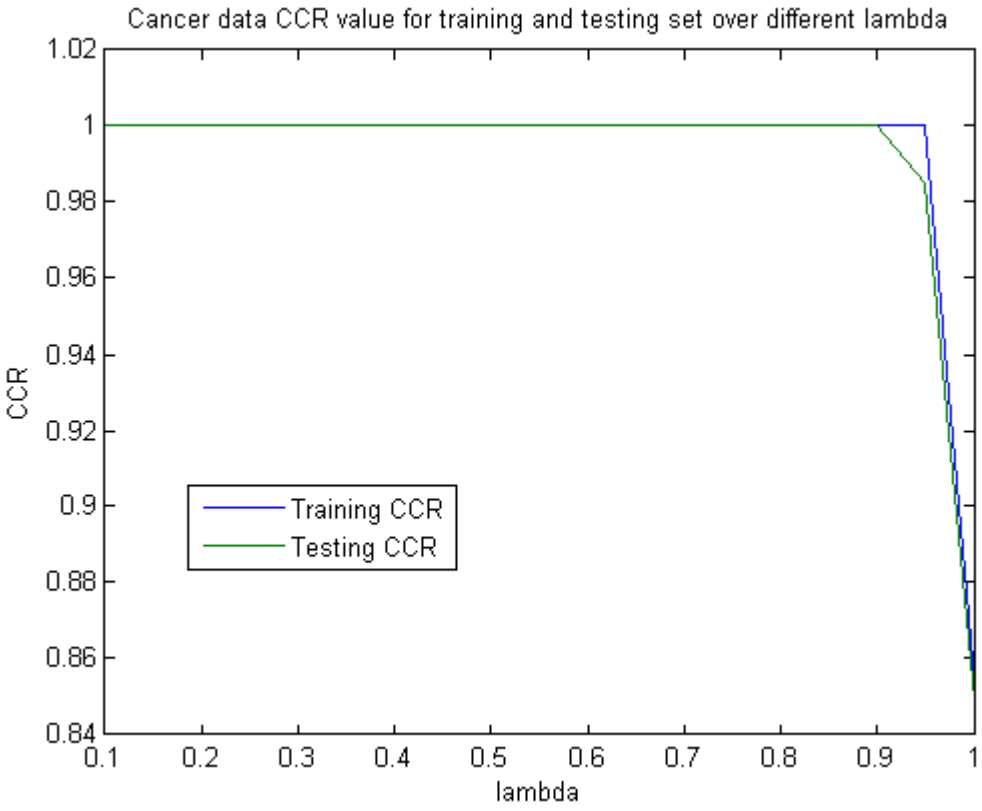
Confusion Matrix Worst CCR			
Labels	1	2	3
1	29	0	0
2	0	29.1	1.6
3	0	1.6	28.7

1.3 Part c

Helper functions located in online submission.

1.4 Part d

Resulting graph:



2 Naive Bayes Text Document Classifiers

2.1 Part a

Total number of unique words in training set: 53975
Total number of unique words in testing set: 47376
Total number of unique words in entire dataset: 61188
Average document length of training set: 245.39 words.
Average document length of testing set: 239.43 words.
Total number of unique words in testing set that are not in the training set: 7213.

2.2 Part b

200778 (16.41%) $\beta_{w,c}$'s are non-zero.
6958 (92.71%) documents have $P(Y = c|x) = 0$ for all $c = 1, ..., 20$. This occurs because the documents contain words that are not seen in the training samples. Since the total probability is the multiplication of all the underlying probabilities for each word, any word in a test sample that has not been seen in a training example for a class, will result in a probability of 0 for that class.
The test CCR = 5.44%.

2.3 Part c

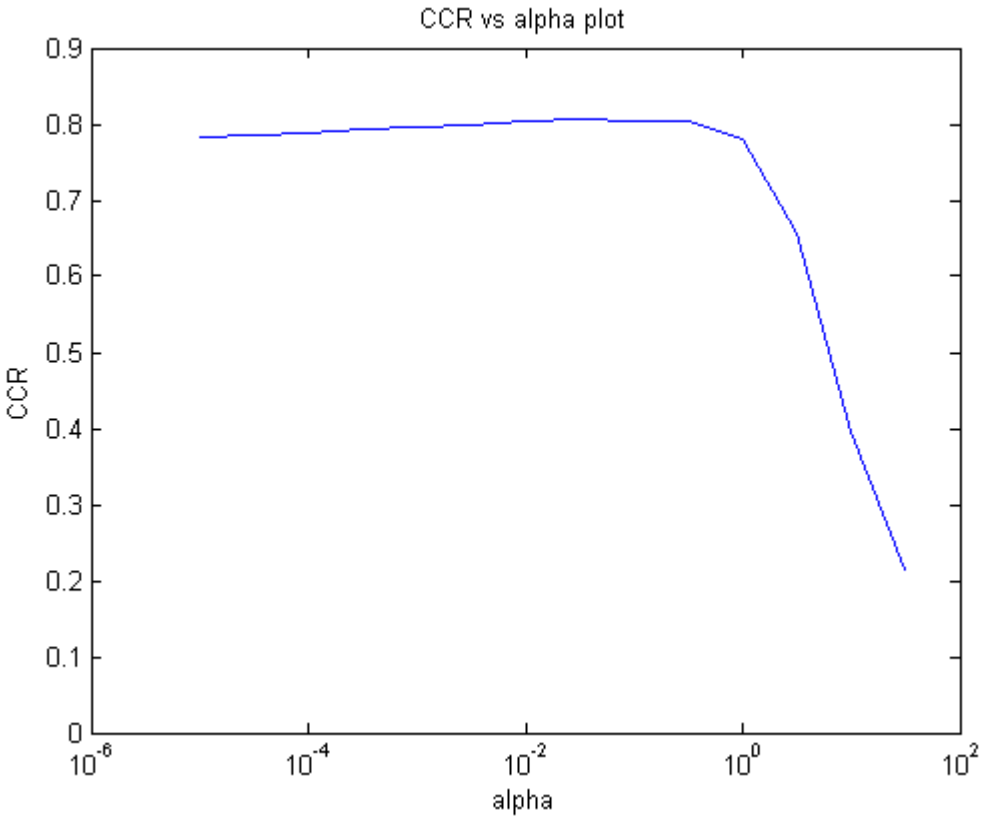
200778 (18.60%) $\beta_{w,c}$'s are non-zero.
The test CCR = 7.16%.

2.4 Part d

The test CCR = 78.52%.
The confusion matrix:

Confusion Matrix for Naive Bayes with alpha = 1/W																				
Labels	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	249	0	0	0	0	1	0	0	1	0	0	2	0	3	3	24	2	3	4	26
2	0	286	13	14	9	22	4	1	1	0	1	11	8	6	10	1	2	0	0	0
3	1	33	204	57	19	21	4	2	3	0	0	12	5	10	8	3	1	0	5	3
4	0	11	30	277	20	1	10	2	1	0	1	4	32	1	2	0	0	0	0	0
5	0	17	13	30	269	0	12	2	2	0	0	3	21	8	4	0	1	0	1	0
6	0	54	16	6	3	285	1	1	3	0	0	5	3	6	4	0	1	1	1	0
7	0	7	5	32	16	1	270	17	8	1	2	0	7	4	6	0	2	1	2	1
8	0	3	1	2	0	0	14	331	17	0	0	1	13	0	4	2	0	0	6	1
9	0	1	0	1	0	0	2	27	360	0	0	0	3	1	0	0	1	1	0	0
10	0	0	0	1	1	0	2	1	2	352	17	0	1	3	3	5	2	1	5	1
11	2	0	1	0	0	0	2	1	2	4	383	0	0	0	0	1	2	0	1	0
12	0	3	0	3	4	1	0	0	0	1	1	362	2	2	2	0	9	0	5	0
13	3	20	4	25	7	4	8	11	6	0	0	21	264	9	7	1	3	0	0	0
14	5	7	0	3	0	0	3	5	4	1	0	1	8	320	8	7	6	5	8	2
15	0	8	0	1	0	3	1	0	1	0	1	4	6	5	343	3	2	1	12	1
16	11	2	0	0	0	2	1	0	0	0	0	0	0	2	0	362	0	1	2	15
17	1	1	0	0	0	1	1	2	1	1	0	4	0	5	2	1	303	5	23	13
18	12	1	0	1	0	0	1	2	0	2	0	2	1	0	0	6	3	326	18	1
19	6	1	0	0	1	1	0	0	0	0	0	5	0	10	6	2	63	6	196	13
20	39	3	0	0	0	0	0	0	1	1	0	1	0	2	6	27	10	3	7	151

2.5 Part e



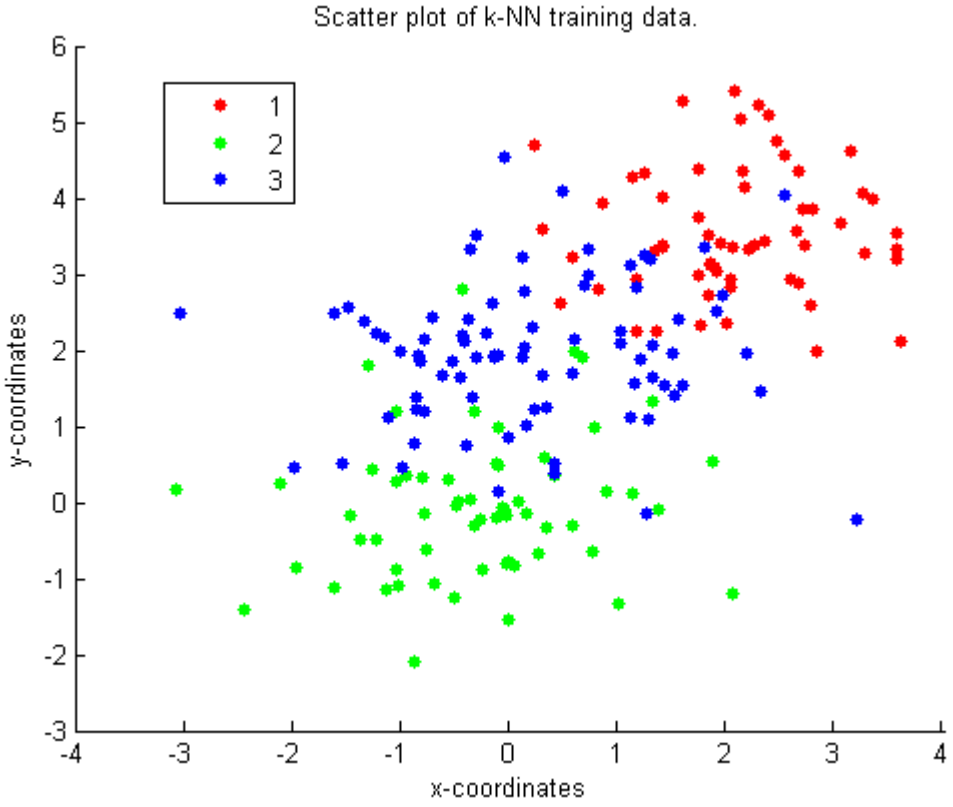
2.6 Part f

Size of new dictionary: 60698 words.
Average document length of training documents: 116.98 words.
Average document length of testing documents: 114.62 words.
The test CCR = 78.23%.

3 Nearest Neighbor Classifier

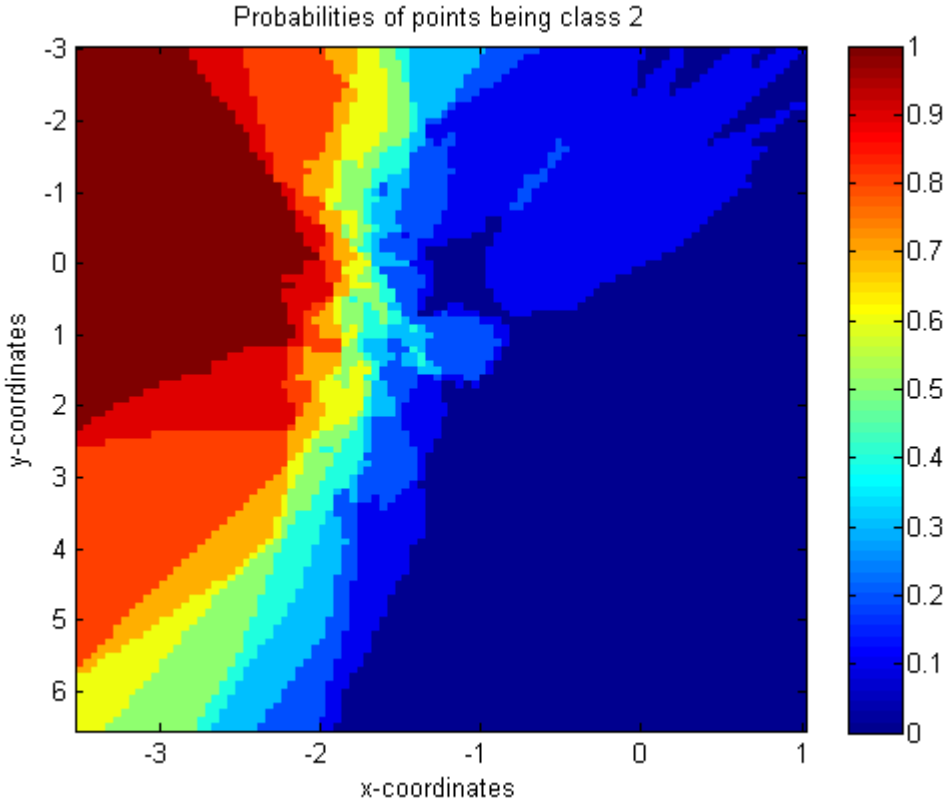
3.1 Part a

Scatter Plot:

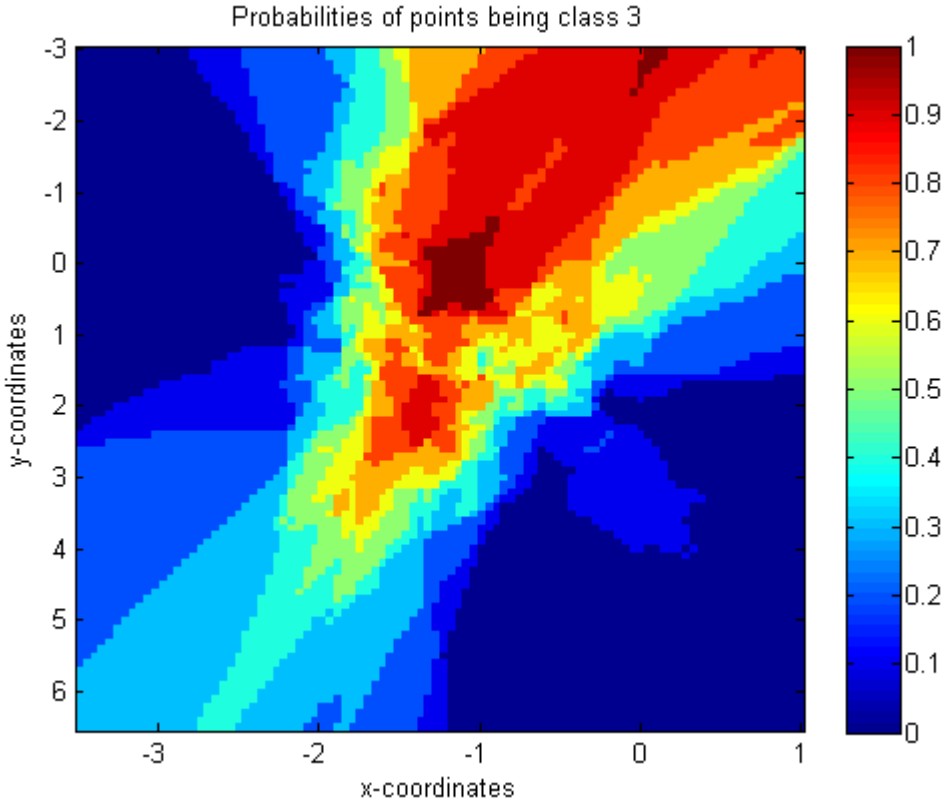


3.2 Part b

Probability color map for class = 2:

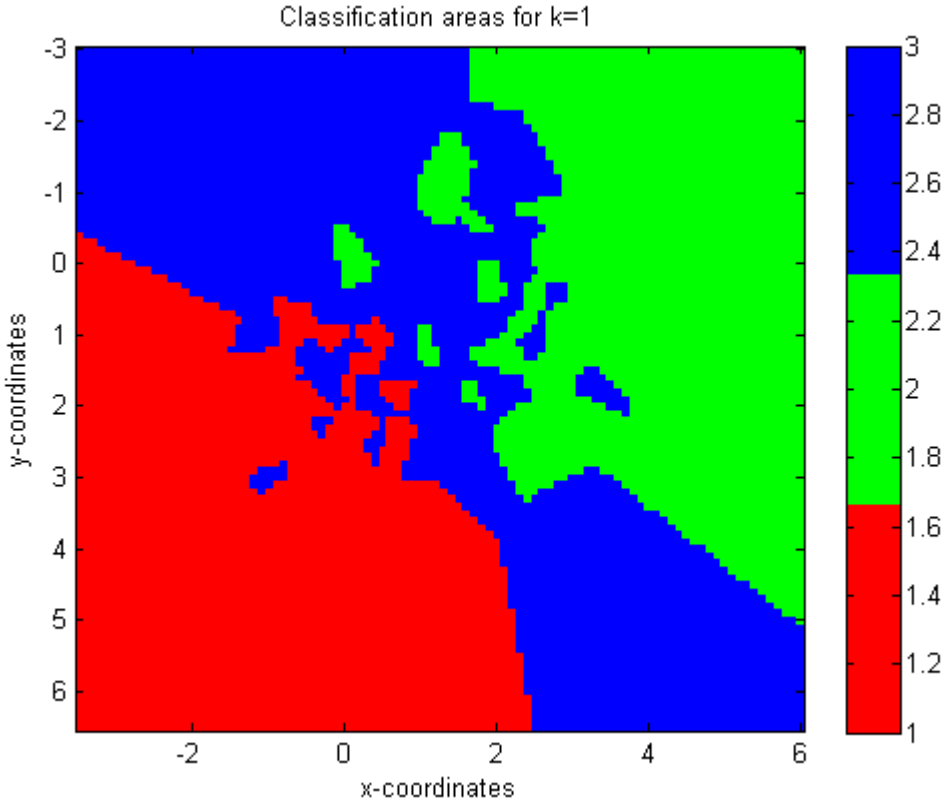


Probability color map for class = 3:

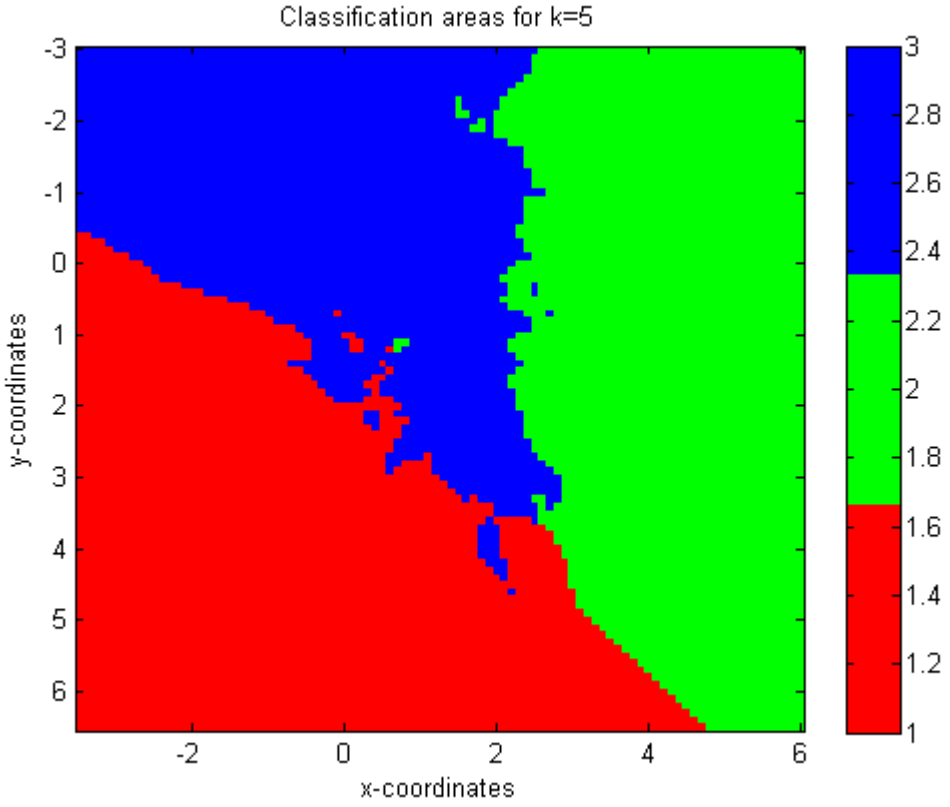


3.3 Part c

Classification coding scheme for k = 1:



Classification coding scheme for k = 5:



Here we can see that for k = 1, we only care about the closest point to each location. Because of this there is a fairly even split between the 3 classes. However, as we expand our search to more points, we see the green and red classes start to dominate as they have more points at close to proximity to much of the area.

3.4 Part d

The Test CCR = 96.31%
Confusion matrix:

Confusion Matrix for K-NN (K=1)										
Labels	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	973	2	1	0	0	1	2	1	0	0
1	0	1129	3	0	1	1	1	0	0	0
2	9	8	987	6	1	0	2	17	2	0
3	0	2	4	965	1	21	0	9	4	4
4	1	9	0	0	937	0	3	4	1	27
5	2	1	0	17	2	848	9	1	5	7
6	5	2	1	0	2	5	943	0	0	0
7	0	20	4	2	4	0	0	989	0	9
8	9	5	6	21	4	18	3	4	894	10
9	1	5	1	7	13	5	1	9	1	0