Informatics 2B - Coursework 2

Task 2 - k-NN and Gaussian Classification

s1765026 - University of Edinburgh

April 2019

Task 2.1

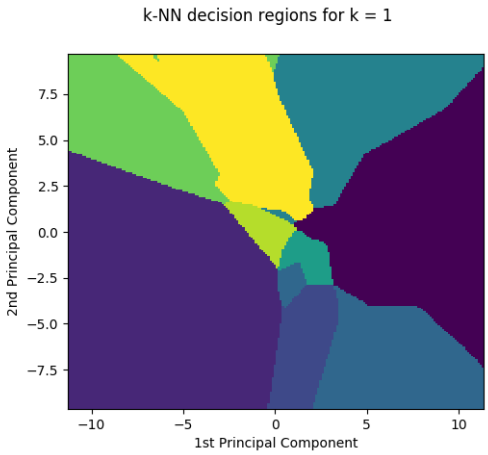
K	1	3	5	10	20
Runtime (secs)	32.46727	32.48070	32.47686	32.51035	32.51912
Number of Samples (N)	3998				
Number of errors (Nerrs)	108	112	124	132	156
Accuracy (acc)	97.29865%	97.19860%	96.89845%	96.69835%	96.09805%

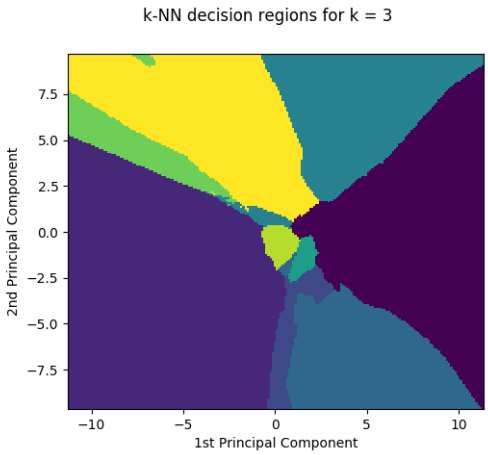
The runtimes presented in the table include the overhead (32.23439 secs) of computing the distance and sorting of k-NN. This is only computed once, but if we were running the function for a single value of k the time it would take is approximately what is presented on the table rather than a negligible value (excluding the overhead). That is the reason the decision was made to use such value.

Task 2.2

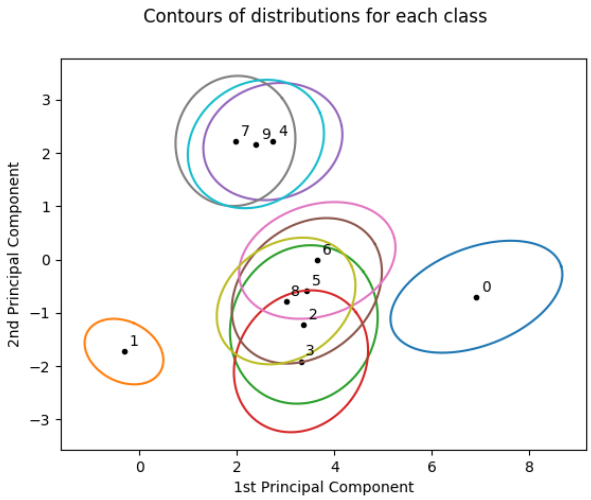
To tackle this question we had to reduce the dataset to a conservative size. We choose the first 2000 samples as this is guaranteed to fit in memory of any modern machine.

This limitation was imposed by the fact that vectorising the square distance function requires a large amount of memory. In all other parts of the assignment, the memory available is more than enough, but in this case the best we can hope for is to either not vectorise our code, making it very slow, or to use only a subset of the data points, which is what has been done.





Task 2.3



Task 2.4

K	1	2	3	4	5
Correlation (r_{12})	-0.20890	0.15840	0.06222	-0.43908	0.59212

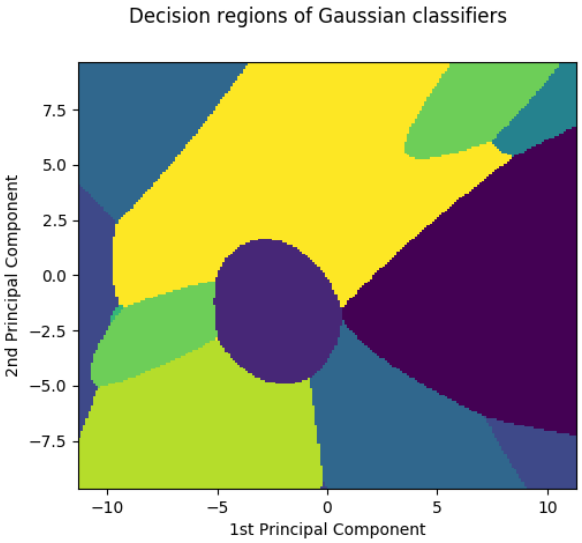
K	6	7	8	9	10
Correlation (r_{12})	-0.31049	0.31548	0.54231	0.11689	0.44342

Overall Correlation: 7.78629×10^{-17}

Task 2.5

Runtime (secs)	5.01578
Number of Samples (N)	3998
Number of errors (Nerrs)	199
Accuracy (acc)	95.02251%

Task 2.6



Task 2.7

Ratio	0.9	0.8	0.7	0.6	0.5	0.4	0.3
Accuracy	94.9725%	94.9725%	94.9975%	95.2226%	95.0725%	94.9475%	95.1476%

Task 2.8

L	2	5	10
Runtime (secs)	9.90424	24.32799	48.79008
Number of Samples (N)	3998		
Number of errors (Nerrs)	175	121	93
Accuracy (acc)	95.62281%	96.97349%	97.67384%