Imperial College London
BUSINESS SCHOOL

MSc Business Analytics 2018/19
Machine Learning
Wolfram Wiesemann

Solution 1

1. Individual Assignment: Predicting Plant Types with *k*-Nearest Neighbours

(a) We calculate (μ, σ) = (5.79, 0.91) for the sepal length and (μ, σ) = (3.1, 0.35) for the sepal width. After normalisation, the data set becomes:

#	sepal length	sepal width	class
1	0.55	-0.58	virginica
2	-0.75	0.82	setosa
3	-0.1	-1.71	virginica
4	-0.86	1.11	setosa
5	-1.08	0.82	setosa
6	0.87	-0.58	versicolor
7	0.55	0.54	versicolor
8	0.55	0.82	versicolor
9	0.22	0.82	versicolor
A	-1.19	0.26	???
B	-0.86	0.54	???
C	0.33	-0.58	???
D	1.09	0.26	???
E	-1.04	-2.28	???
F	2.07	-0.3	???

(b) We calculate the pairwise distances as follows:

	1	2	3	4	5	6	7	8	9
A	1.93	**0.71**	2.25	**0.90**	**0.57**	2.23	1.76	1.83	1.52
B	1.81	**0.30**	2.38	**0.56**	**0.35**	2.07	1.41	1.44	1.12
C	**0.21**	1.78	1.21	2.07	2.00	**0.54**	**1.15**	1.43	1.41
D	1.00	1.93	2.31	2.13	2.24	**0.87**	**0.61**	**0.78**	1.03
E	**2.59**	3.17	**1.42**	3.43	3.12	**2.84**	3.34	3.67	3.51

	1	2	3	4	5	6	7	8	9
F	**1.55**	3.04	2.59	3.26	3.35	**1.23**	**1.74**	1.89	2.16

Our *k*-Nearest Neighbours classifier then makes the following predictions:

	3 nearest neighbours			prediction
A	2 (setosa)	4 (setosa)	5 (setosa)	**setosa**
B	2 (setosa)	4 (setosa)	5 (setosa)	**setosa**
C	1 (virginica)	6 (versicolor)	7 (versicolor)	**versicolor**
D	6 (versicolor)	7 (versicolor)	8 (versicolor)	**versicolor**
E	1 (virginica)	3 (virginica)	6 (versicolor)	**virginica**
F	1 (virginica)	6 (versicolor)	7 (versicolor)	**versicolor**

It turns out that the actual values are setosa (A), setosa (B), versicolor (C), virginica (D), setosa (E) and virginica (F) — so with 9 data points and 2 features, we were only able to predict 50% of the validation samples correctly.