Assignment 1 - Individual

Machine Learning

MSc Business Analytics 2020/2021

For each of the attributes, namely petal length and petal width, the mean μ and standard deviations σ are computed. For the petal length, the mean μ_{length} is

$$\mu_{length} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

where X_i represents the length of a petal numbered i.

$$\mu_{length} = \frac{5.6 + 1.5 + 5.0 + ... + 5.9 + 1.3 + 6.1}{15} = 3.76 \tag{2}$$

(correct to 2 decimal places)
For the petal length, the standard deviation (σ_{length}) is

$$\sigma_{length} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{length})^2} \tag{3}$$

$$\sigma_{length} = \sqrt{\frac{(5.6 - 3.76)^2 + (1.5 - 3.76)^2 + ... + (6.1 - 3.76)^2}{15}} = 1.98 \tag{4}$$

(correct to 2 decimal places)

For the petal width, the mean μ_{width} is

$$\mu_{width} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{5}$$

where Y_i represents the width of a petal numbered i.

$$\mu_{width} = \frac{1.8 + 0.2 + 2.0 + ... + 2.3 + 0.3 + 2.3}{15} = 1.27 \tag{6}$$

(correct to 2 decimal places)
For the petal width, the standard deviation (σ_{width}) is

$$\sigma_{width} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_{width})^2} \tag{7}$$

$$\sigma_{width} = \sqrt{\frac{(1.8-1.27)^2 + (0.2-1.27)^2 + ... + (2.3-1.27)^2}{15}} = 0.91 \quad (8)$$

(correct to 2 decimal places)
The data is then normalised using the following equations for Z-normalisation.

$$X_{N,i} = \frac{X_i - \mu_{length}}{\sigma_{length}} \quad (9)$$

$$Y_{N,i} = \frac{Y_i - \mu_{width}}{\sigma_{width}} \quad (10)$$

where $X_{N,i}$ is the normalised value of X_i and $Y_{N,i}$ is the normalised value of Y_i. The normalised values are shown in the table below with example calculations for the first row. For consistency, all values are given to two decimal places.

X_i	Y_i	$X_{N,i}$	$Y_{N,i}$
5.6	1.8	$\frac{5.6-3.76}{1.98} = 0.93$	$\frac{1.8-1.27}{0.91} = 0.58$
1.5	0.2	-1.14	-1.18
5.0	2.0	0.63	0.80
1.3	0.3	-1.24	-1.07
1.6	0.2	-1.09	-1.18
4.6	1.3	0.42	0.03
6.0	2.5	1.13	1.35
5.6	2.4	0.93	1.24
4.5	1.6	0.37	0.36
1.3	0.2	-1.24	-1.18
1.4	0.2	-1.19	-1.18
4.7	1.4	0.47	0.14
5.9	2.3	1.08	1.13
1.3	0.3	-1.24	-1.07
6.1	2.3	1.18	1.13

Table 1: Values for normalised petal length and petal width (last two columns)

The Euclidean distance from each training point, T_i to each prediction point P_j is calculated using the following equation. Thus for training points, $i = 1, 2, ..., 9$ and for prediction points, $j = 10, 11, ..., 15$.

$$D_{i,j} = \sqrt{(T_{X,i} - P_{X,j})^2 + (T_{Y,i} - P_{Y,j})^2} \quad (11)$$

where $T_{X,i}$ is the length of T_i and $T_{Y,i}$ is the width of T_i. Similarly, $P_{X,j}$ is the length of P_j and $P_{Y,j}$ is the width of P_j.

For example, the Euclidean distance between training point 1 (T_1) and prediction point 10 P_{10} is

$$D_{1,10} = \sqrt{(0.93 - -01.24)^2 + (0.59 - -1.18)^2} = 2.80 \tag{12}$$

(correct to 2 decimal places)
All Euclidean distances (correct to 2 decimal places) are shown in the table below.

Classes of T_i		P_1	P_2	P_3	P_4	P_5	P_6
virginica	T_1	2.80	2.79	**0.63**	0.57	2.73	**0.60**
setosa	T_2	**1.10**	**0.05**	2.09	3.21	**0.15**	3.27
virginica	T_3	2.72	2.72	0.68	**0.56**	2.64	0.65
setosa	T_4	**0.11**	**0.12**	2.10	3.20	**0.00**	3.27
setosa	T_5	**0.15**	**0.10**	2.05	3.17	**0.19**	3.24
versicolor	T_6	2.06	2.06	**0.12**	1.28	2.00	1.33
versicolor	T_7	3.47	3.47	1.38	**0.23**	3.39	**0.23**
versicolor	T_8	3.25	3.25	1.19	**0.19**	3.17	**0.28**
versicolor	T_9	2.23	2.23	**0.24**	1.05	2.16	1.12

Table 2: Euclidean distances between training and prediction points.

The three nearest T_i neighbours (with smallest Euclidean distances to P_i) for every P_i are denoted in bold in every column. These neighbours are further highlighted in the table below for the purpose of predictions.

P_i	P_i's neighbours	P_i's neighbour classes	P_i class prediction
P_1	T_2,T_4,T_5	**setosa, setosa, setosa**	setosa
P_2	T_2,T_4,T_5	**setosa, setosa, setosa**	setosa
P_3	T_1,T_6,T_9	virginica, **versicolor, versicolor**	versicolor
P_4	T_3,T_7,T_8	virginica, **versicolor, versicolor**	versicolor
P_5	T_2,T_4,T_5	**setosa, setosa, setosa**	setosa
P_6	T_1,T_7,T_8	virginica, **versicolor, versicolor**	versicolor

Table 3: Class prediction

The predicted classes are given in the last column of Table 3. The prediction is made by analysis of the classes (Column 3) of the three nearest neighbours (Column 2) of P_i. The class of P_i is taken to be the majority class of its three nearest neighbours. For example, P_6 has neighbours T_1,T_7,T_8 which have classes virginica, **versicolor, versicolor** respectively. There are 2 versicolor candidates and only one virginica candidate. Thus P_6 is allocated to the versicolor class.