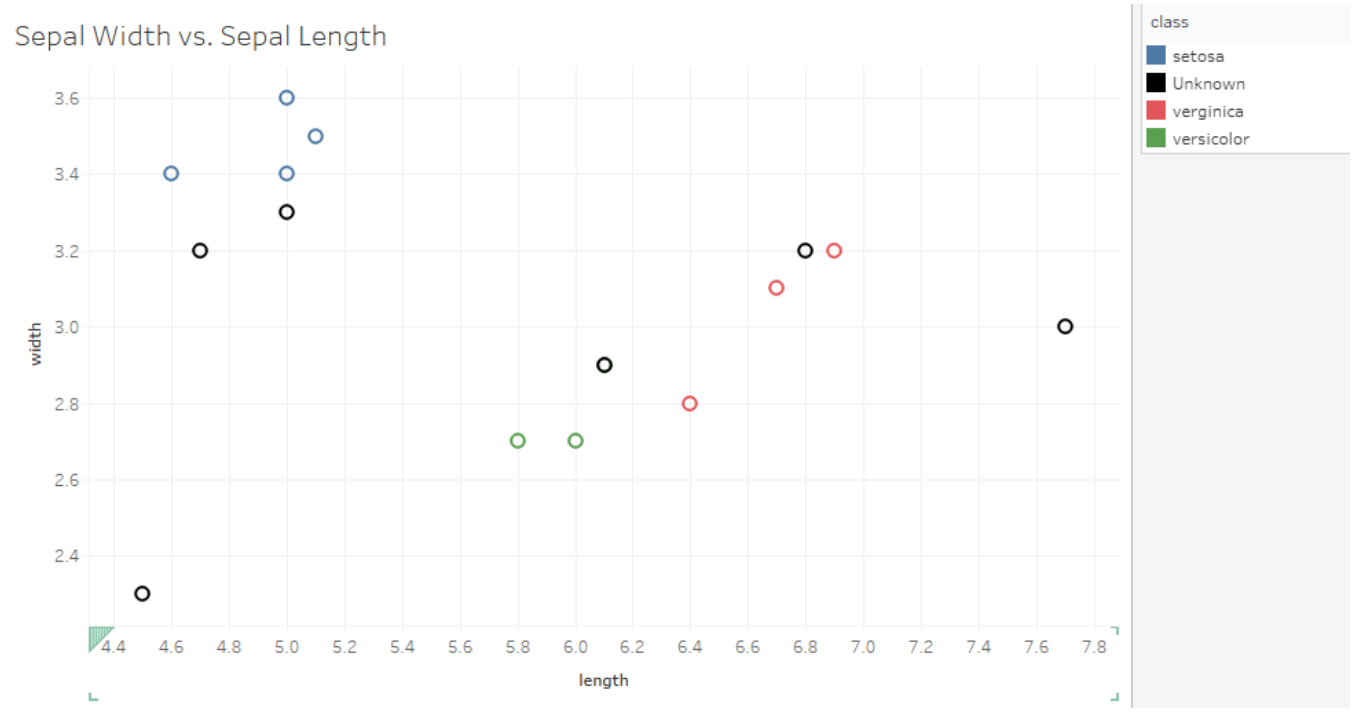
Assignment 1 – Nearest Neighbors

MSBA 315: Predictive Analytics and Machine Learning- Spring 2020

Name: Abbas Nassereddine

The following graph visualizes the position of the given and unknown points in the set:



The following table shows all the calculations performed along with the resulting labels according to the K-Neighbors algorithm which finds the closest neighbors of an unknown point and labels it according to the label of the majority of labels.

Point	Length	Width	Class	Normalized_Length	Normalized_Width	Distance to Unknown Pt 1	Distance to Unknown Pt	tance to Unknown	ance to Unknown	istance to Unknown Pt	istance to Unknown Pt
1	6	2.7	versicolor	0.32	-1.33	2.32	2.28	0.63	1.88	2.34	2.44
2	5	3.4	setosa	-1.01	0.84	0.74	0.31	2.13	2.46	3.47	3.78
3	6.4	2.8	verginica	0.85	-1.02	2.57	2.42	0.50	1.35	2.95	1.83
4	4.6	3.4	setosa	-1.54	0.84	0.63	0.61	2.52	2.98	3.41	4.29
5	6.1	2.9	versicolor	0.45	-0.71	2.07	1.91	0.00	1.31	2.82	2.14
6	5	3.6	setosa	-1.01	1.46	1.30	0.93	2.61	2.69	4.08	4.03
7	5.8	2.7	versicolor	0.05	-1.33	2.13	2.14	0.74	2.04	2.12	2.68
8	5.1	3.5	setosa	-0.87	1.15	1.07	0.63	2.28	2.44	3.80	3.78
9	6.7	3.1	verginica	1.24	-0.09	2.67	2.33	1.01	0.34	3.83	1.36
10	6.9	3.2	verginica	1.51	0.22	2.91	2.53	1.41	0.13	4.23	1.23
Unknown 1	4.7	3.2	Unknown	-1.40	0.22	-	-	-	-	-	-
Unknown 2	5	3.3	Unknown	-1.01	0.53	-	-	-	-	-	-
Unknown 3	6.1	2.9	Unknown	0.45	-0.71	-	-	-	-	-	-
Unknown 4	6.8	3.2	Unknown	1.38	0.22	-	-	-	-	-	-
Unknown 5	4.5	2.3	Unknown	-1.67	-2.57	-	-	-	-	-	-
Unknown 6	7.7	3	Unknown	2.57	-0.40	-	-	-	-	-	-
Training Data Average	5.76	3.13				setosa	setosa	versicolor	verginica	versicolor	verginica
Training Data SD	0.76	0.32									

It is important to note that for the mean and SD calculations, only the training set was used in order to ensure that the model will be biased and contaminated by the test set. According to Abu Mostafa et al., this learning principle is coined “data snooping”.

$$\mu_{training\_length} = \frac{6 + 5 + 6.4 + 4.6 + 6.1 + 5 + 5.8 + 5.1 + 6.7 + 6.9}{10} = 5.76$$

$$\sigma_{training_{length}} = \frac{(6 - 5.76)^2 + (5 - 5.76)^2 + (6.4 - 5.76)^2 + (4.6 - 5.76)^2 + (6.1 - 5.76)^2 + (5 - 5.76)^2 + (5.8 - 5.76)^2 + (5.1 - 5.76)^2 + (6.7 - 5.76)^2 + (6.9 - 5.76)^2}{10}$$
$$= 0.76$$

Each datapoint for each feature is normalized using z-score normalization given by the following equation: $Z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}$ $for\ each\ i\ data\ point\ and\ j\ feature.$

- Sample calculation: $Z_{2-length} = \frac{5 - 5.76}{0.76} = -1.01$

After normalizing, we proceed to find the neighbors of each unknown point using k=3, i.e. we will be checking the 3 closest neighbors of each unknown data point to categorize it. We will be using Euclidean distance to calculate distances. The Euclidean distance is given by: $D_{1-2} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

-Sample calculation: $D_{unknown1-point\ 2} = \sqrt{(-1.01 + 1.4)^2 + (0.84 - 0.22)^2} = 0.74$

Calculating the distance between the unknown point and each point in the training set, we choose the three closest points and categorize the unknown point according to the majority of the labels on its neighbors. For example, for unknown point 4, two out of three neighbors are verginica, so the point would be labeled as verginica. Note that unknown point 3 is in itself point 5 in the training set and would therefore be directly labeled as versicolor.

As a conclusion the unknown points will be categorized as follows:

Unknown Point	Label
1	Setosa
2	Setosa
3	Versicolor
4	Verginica
5	Versicolor
6	Verginica