Practical Sheet 2: k-Nearest neighbours

APL 405 - 2023W (Machine Learning for Mechanics)

[20 marks] Predicting colors with k-NN: Consider a synthetic binary classification problem ($M = 2$) with a given training dataset having $N = 10$ (statistically independent) observations of 2-dimensional input variables $\mathbf{x} = [x_1 \;\; x_2]^T$ and one categorical output y, representing either the colour Red or Blue.

i	x_1	x_2	y
1	-1	3	Red
2	2	1	Blue
3	-2	2	Red
4	-1	2	Blue
5	-1	0	Blue
6	1	1	Red

1. Write a function to calculate the Euclidean distance between two vectors. Using a test input $\mathbf{x}^* = [1 \;\; 2]^T$, compute the Euclidean distance between each training data point $\mathbf{x}^{(i)}$ and the test data point. Print the result in tabular fashion

2. Output the k-NN prediction for (i) $k = 1$ (one neighbour), and (ii) $k = 3$ (three neighbours)

3. Repeat step (2) for the test input $\mathbf{x}^* = [0 \;\; 2]^T$. What color prediction do you get for the two cases?

4. Repeat step (2) for the test input $\mathbf{x}^* = [0.5 \;\; 2]^T$. What color prediction do you get for the two cases?

5. Can you plot the decision boundaries of the two k-NN classifiers?

6. Consider a case where the scale of the input variable x_1 is 100 times greater than x_2, meaning (say) the column of x_1 is multiplied with 100. Normalize the input data (using min-max scaling) and then implement step (2) for predicting on the test input $\mathbf{x}^* = [0 \;\; 2]^T$.