

ML Problem Solving

K-Nearest Neighbor Classifier

1. Consider the following dataset, where each row is an observation. First and second columns are input features and the true output is the last column.

$$D = [[3, 3, 1], [1, 1, 1], [-1, 0, 1], [2, 2, 0], [-2, 2, 0]]$$

Draw the data points in a coordinate system to represent your input space. Mark class 1 as an 'X' and class 0 as a 'O'. Hint: labeling datapoints help (e.g., A, B, C...).

2. Create a 5×5 distance matrix using euclidian distance. Hint: this is a symmetric matrix.

3. Predict the classes using the provided datapoints for $k \in \{1, 3, 5\}$. Write down the accuracy for each k . Which one is the best? Why? Are these training or testing accuracies?
4. For each k , draw a confusion matrix (consider that rows represent predictions and columns represent actual classes).
5. For new data points in $T = [[3, 1, 1], [0, 2, 0]]$, predict their classes for each k . Did the 'best' k change? If so, why?