Problem Set 5

1. Consider running a k-NN classifier using Euclidean distance on the data set from Figure 1, whereby each points belongs to one of two classes: $+$ and $\circ$. [50]

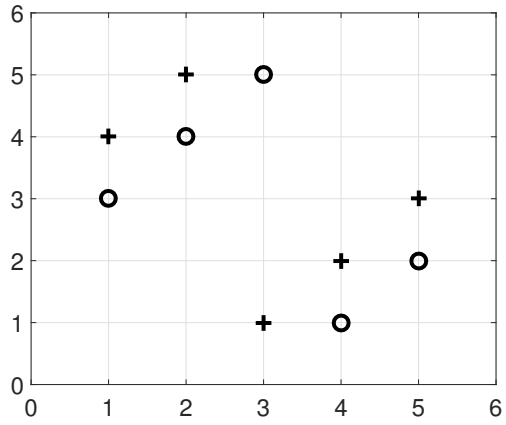

Figure 1: Points belonging to two classes

   (a) What is the 10-fold cross validation error when $k = 1$? [15]

   (b) Which of the values $k \in \{3, 4, 5, 9\}$ yields the minimum number of 10-fold cross validation errors? [20]

   (c) Give a distance metric, instead of the Euclidean distance, such that the 10-fold cross validation error of 1-NN is $\frac{4}{10}$. [15]

2. (a) Consider the points $\{1, 2, 3, 4\}$ in the 1-dimensional Euclidean space. Does Lloyd's K-means algorithm, for $K = 2$, always yield the optimal clustering? You can assume that, in case of ties, the algorithm makes the most favourable choice. [15] [50]

   (b) Consider the points $\{1, 2, 3, a\}$ in the 1-dimensional Euclidean space, where a is a real number satisfying $a > 3$. Find a *small* value for a such that Lloyd's K-means algorithm, for $K = 2$, is sub-optimal. (Full points are awarded if your value a is sufficiently small, e.g., $a - 0.1$ would not be a valid solution.) [35]