Applied HW: k-Nearest Neighbours:

Training data set containing six observations, three predictors, and one quantitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

We wish to use this data set to predict Y = X1 = X2 = X3 = 0 using KNN

1. Compute the Euclidean distance between each observation and the test point:

   $X_1 = X_2 = X_3 = 0$.

   Euclidean distance: $d(p,q) = \text{sqrt}(\text{sum}(q_i - p_i)^2)$
   (where the sum is taken over all points)

Observation	Euclidean Formula	Euclidean Distance
1	$\sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2}$	3.0
2	$\sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2}$	2.0
3	$\sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2}$	3.162
4	$\sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2}$	2.236
5	$\sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2}$	1.414
6	$\sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2}$	1.732

2. What is our prediction with K=1? Why?

   KNN: We predict the nearest response variable based on the nearest neighbour to the test point.

   The smallest distance is 1.414 coming from observation 5, which correlates to the response “Green”.

   Our prediction for Y when K=1 is “Green”

3. What is our prediction with K=3? Why?

   With K=3, we predict the three nearest response variables. The three smallest distances are:

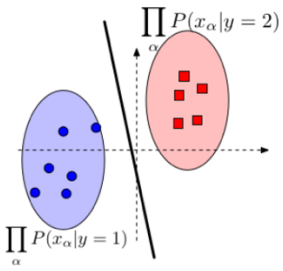
   1.414 from observation 5 with response “Green”
   1.732 from observation 6 with response “Red”
   2 from observation 2 with response “Red”

   So our prediction for Y is “Red” since it’s the majority.

4. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

   If the decision boundary is highly non-linear, then we would expect the best value of K to be small. A smaller K-value would allow the model to be more adaptive, unlike a higher K-value that may not capture the patterns as well since there are more values to average.

   Example of linear Bayes decision boundary:



   Reference of linear Bayes decision boundary:
   https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html

Pass: you must submit correct distances, predictions with explanations, and whether K should be large or small with an explanation.