
Question1 (2 points)

Suppose we collect data for a group of students in a class with variables

`X1`	hours studied per week
`X2`	GPA
`Y`	receive an A

We fit a logistic regression and produce estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) (1 point) Estimate the probability that a student who studies for 40 h and has a GPA of 3.5 gets an A in the class.

(b) (1 point) How many hours would the student in part (a) need to study to have 50% change of getting an A in the class?

Question2 (5 points)

PricewaterhouseCoopers (PwC) surveyed 1000 online shoppers in the U.S.A and China. One question asked is whether the online shopper followed brands they purchased through social media. The result is given below

	Social Media	
Country	**No**	**Yes**
U.S.A	487	513
China	72	928

(a) (1 point) What is the odds of online shoppers who follow brands through social media in each country?

(b) (1 point) What is the odds ratio for comparing U.S.A. online shoppers with Chinese online shoppers?

(c) (1 point) Write the logistic regression model for this problem using the log odds of following brands through social media as the response variable and country as an indicator predictor (U.S.A = 1).

(d) (1 point) Numerical optimisation gives the estimated slope of -2.5043 and standard error of 0.1377. Transform this result to the odds scale and compare it with your answer in part (b).

(e) (1 point) Construct a 95% confidence interval for the odds ratio and state the conclusion based on this interval.

Question3 (3 points)

When the number of predictors/features p is large, there tends to be deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it is tires into the fact that non-parametric approaches often perform poorly when p is large.

(a) (1 point) Suppose that we have a set of observations, each with measurements on only one feature X, that is, $p = 1$. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to the that test observation. For instance, in order to predict the response

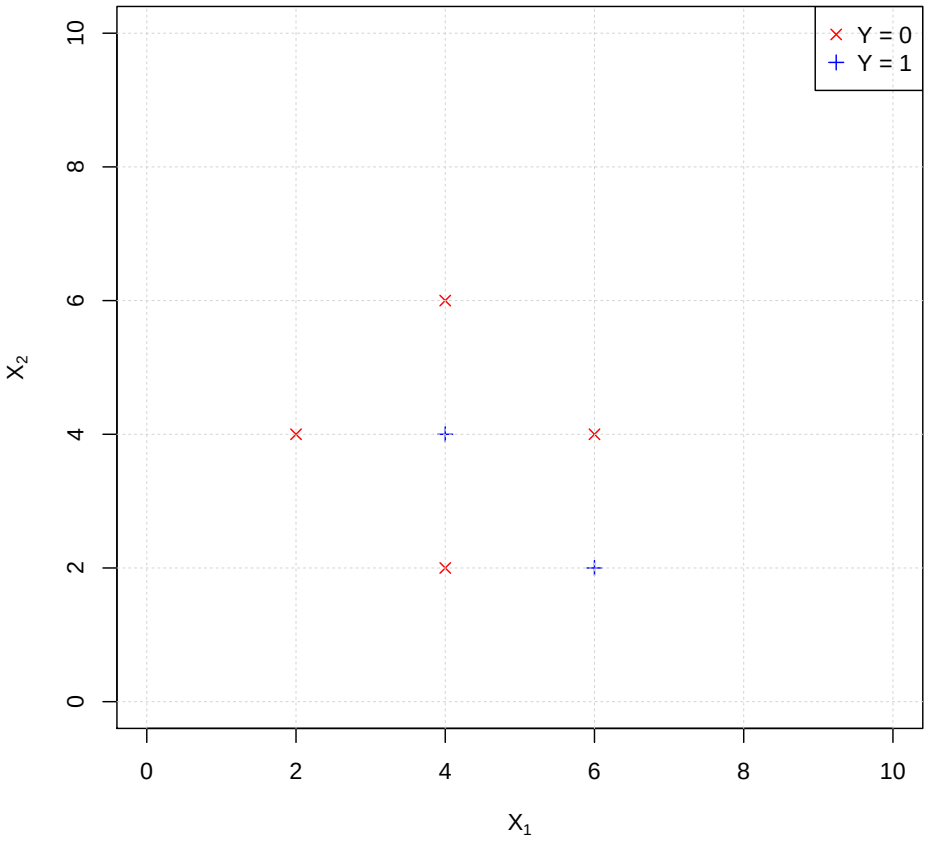
for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of available observations will we use to make the prediction.

(b) (1 point) Now suppose that we have a set of observations, each with measurements on two features X_1 and X_2, that is, $p = 2$. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range X_1 and within 10% of the range of X_2 closest to the test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

(c) (1 point) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centred around the test observation that contains, on average, 10% of the training observations. For $p = 100$, what is the length of each side of the hypercube?

Question4 (4 points)

Consider the K-nearest neighbour (KNN) classifier using Euclidean distance.



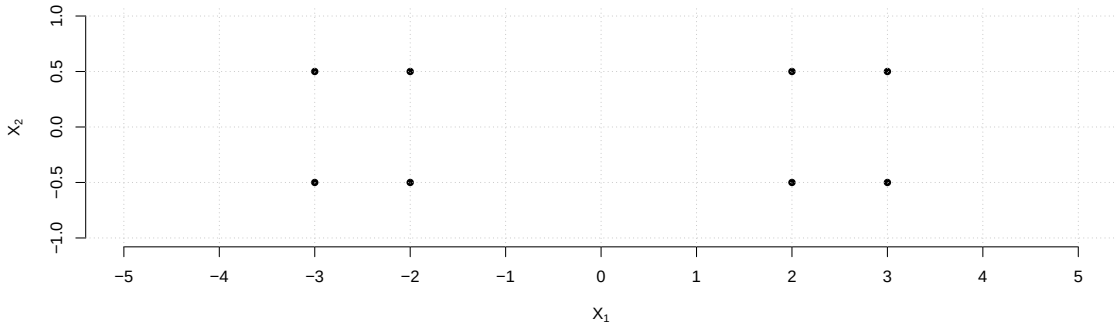
The dataset is described above. Note that a point can be its own neighbour.

(a) (3 points) Sketch the 1-nearest neighbour (1-NN) decision boundary for this dataset.

(b) (1 point) How would the point $(8, 1)$ be classified using (1-NN)?

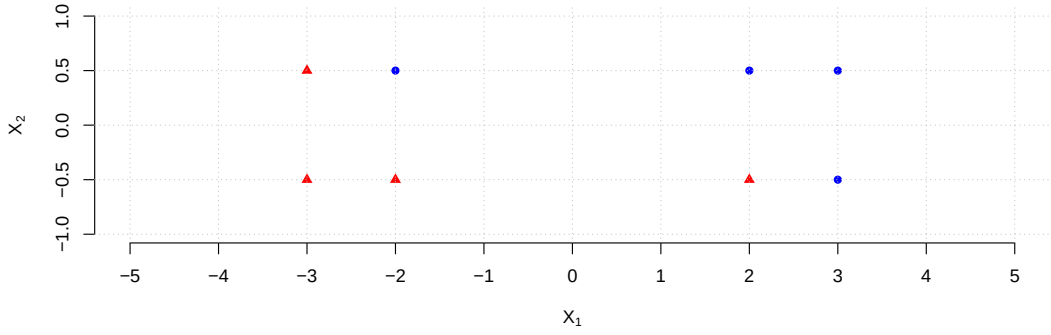
Question5 (5 points)

Consider manually assigning items in a dataset described by into 2 clusters using K-means.



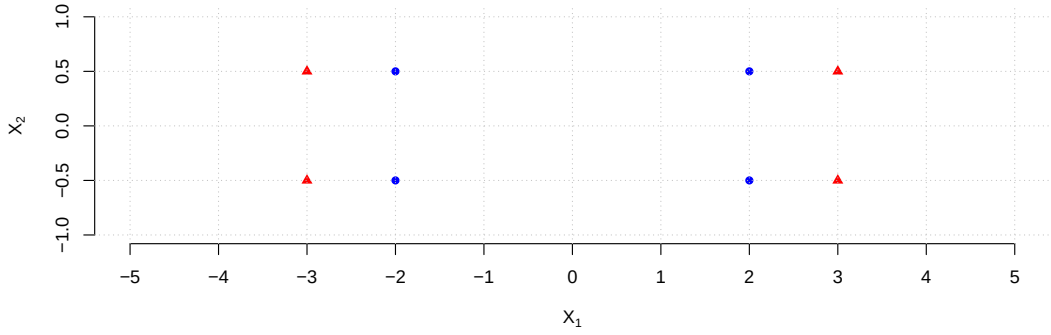
Use squared Euclidean distance as the measure of variation/dissimilarity.

(a) (2 points) Suppose we have the following initialisation.



Compute K-means clustering until convergence, show all steps for all iterations.

(b) (3 points) Suppose we have the following initialisation.



Compute again until convergence. Compare with part (a), what do you notice?

Question6 (6 points)

Suppose that we have 4 observations, for which the following matrix gives the dissimilarity.

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

For instance, the dissimilarity measure between the first and second observations is 0.3, and the dissimilarity measure between the second and the fourth observations is 0.8.

(a) (2 points) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) (2 points) Repeat part (a), this time using single linkage clustering.

(c) (1 point) Suppose that we cut the dendogram obtained in part (a) such that two clusters result. Which observations are in each cluster?

(d) (1 point) Suppose that we cut the dendogram obtained in part (b) such that two clusters result. Which observations are in each cluster?