COMS20011 - Data-Driven Computer Science

Problem Sheet 1 - Data Acquisition and Distances

January 2024

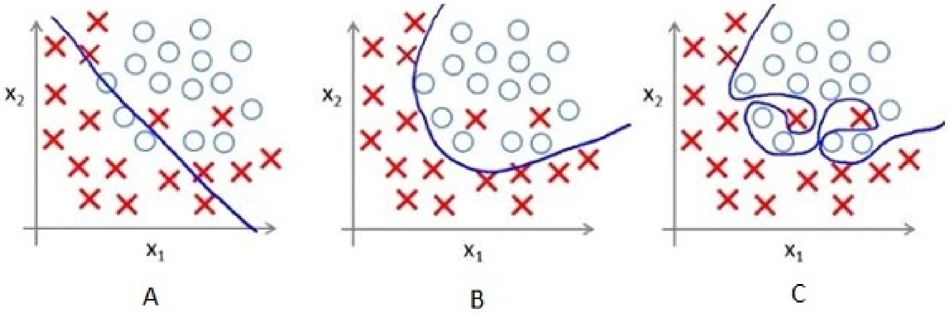
1. **Refreshing your memory:**

   For the set of measurements: -3, 2, 4, 6, -2, 0, 5

   calculate by hand:

   mean, median, variance, standard deviation

2. Below are three scatter plots (A,B,C left to right) of some training data for measurements of two features (x_1, x_2) of different kinds of fish. Also shown, are hand-drawn decision boundaries for modelling regression on the data:



   Which of the statements below are TRUE conclusions:

   (a) The training error in model A is maximum compared to models B and C.

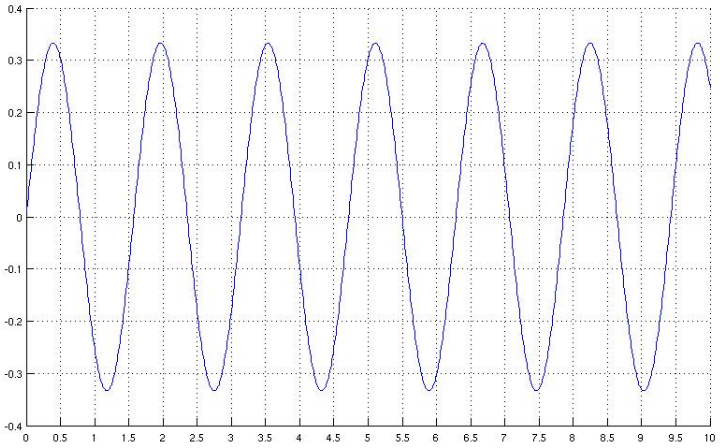
   (b) The best model for this regression problem is C because it has minimum training error (zero).

   (c) Model B is more robust than A and C because it will perform best on unseen data.

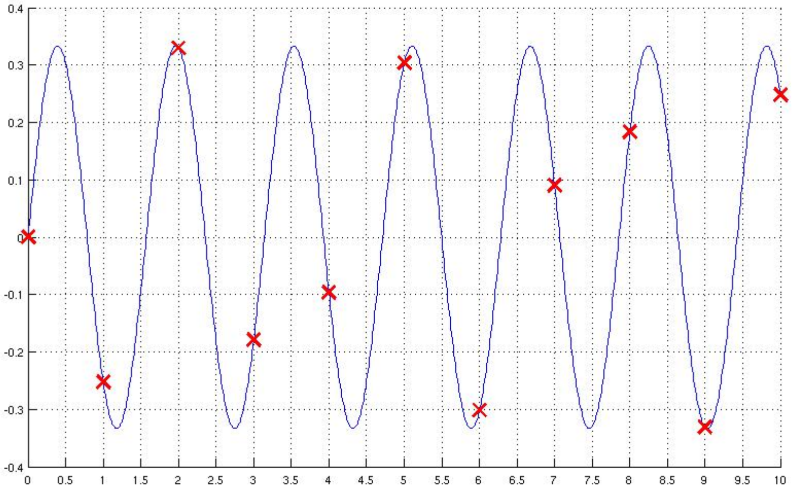
   (d) All models will perform the same because we have not seen the testing data.

   (e) Model C is overfitting the training data compared to A and B.

3. On the $sin(x)$ signal below, label the following terms and approximate their values: period, frequency and amplitude

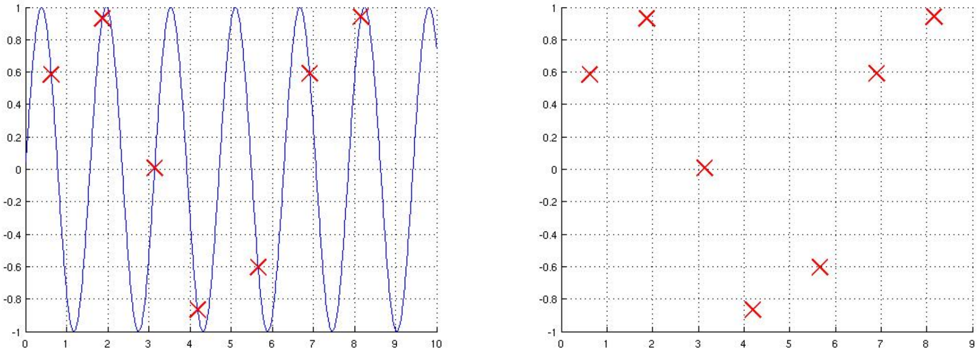


4. For the signal above, convert it into its digital representation using the sampled points. You need to think about the number of bits you would represent each sample as. This is referred to as **Quantisation**. Example, if you need 8 different levels of sound, then 3 bits are sufficient ($2^3 = 8$).



What is the sampling rate in this case??

5. Repeat the digitization and reconstruction step for this data below, can you notice any difference?



6. **Distance measures:** Calculate the following distance measures for the data provided:

   - A = (4,5,6), B = (2, -1, 3) - Distance Measure Manhattan Distance L_1
   - P = (4,5,6), Q = (2, -1, 3) - Distance Measure 3-norm L_3
   - E = (4, 5, 6), F = (2, -1, 3) - Distance Measure Chebyshev Distance L_∞
   - A1 = ‘Shot’, A2 = ’Chop’ - Distance Measure Hamming Distance
   - A1 = ‘weather’, A2 = ’further’ - Distance Measure Hamming Distance
   - A1 = ‘Tank’, A2 = ’Thanks’ - Distance Measure Edit Distance
   - A1 = ‘water’, A2 = ’further’ - Distance Measure Edit Distance
   - A1 = ‘plankton’, A2 = ’plants’ - Distance Measure Edit Distance
   - *** OPTIONAL *** Order, ascendingly, the following words {‘tap’, ‘river’, ‘liquid’, ‘ice’} based on their WUP relatedness to: ‘water’. Use 1-WUP as the distance measure and the online http://ws4jdemo.appspot.com

7. **Distance measures:** Assume you were given a set of whatsapp messages, each with a timestamp (yy-mm-dd hh:mm) and text content (word, word, ...). Propose a distance measure for:

   - calculating whether one message is an exact copy of the other message
   - calculating whether one message was sent before the other message
   - calculating whether one message contains the same set of words as the other message
   - calculating whether one message contains the other message (with potential extras at the start and the end)
   - calculating whether both messages discuss the same topic

   Check your distance measures satisfy: non-negativity, reflexive, symmetric and triangule inequality.

8. You collected a four dimensional dataset of values **x** $= (x_1, x_2, x_3, x_4)$ and calculated the mean to be $(3, 2.6, -0.4, 2.6)$, and the covariance matrix to be

$$\begin{bmatrix} 4 & 0.1 & -4 & -0.1 \\ 0.1 & 0.01 & -0.1 & 0 \\ -4 & -0.1 & 4 & 0.1 \\ -0.1 & 0 & 0.1 & 9 \end{bmatrix}$$

(a) You are asked to only select two variables, x_1 and another variable, to take forward for a machine learning algorithm that predicts future values of the variable **x**. Which other variable would you pick: x_2, x_3 or x_4 and why?

(b) Calculate the eigenvalues and eigenvectors for your chosen covariance matrix

(c) Using the probability density function of the normal distribution in two dimensions, calculate the probability that the following new data $(3, 2.61, 0, 3)$ belongs to the dataset **x** [Note: only use the two variables you picked in (a)]