COMS20011 - Data-Driven Computer Science

Problem Sheet 1 - Data Acquisition and Distances

January 2024

1. **Refreshing your memory:**

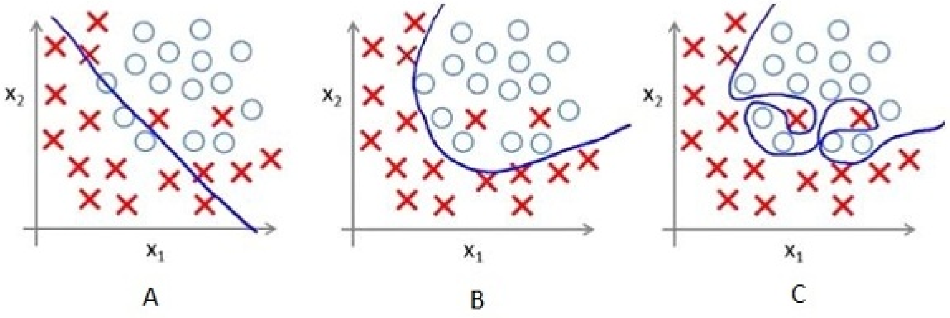
   For the set of measurements: -3, 2, 4, 6, -2, 0, 5

   calculate by hand:

   mean, median, variance, standard deviation

   Answer:
   mean = 1.7, median = 2, variance = 12.2, standard deviation = 3.5

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).

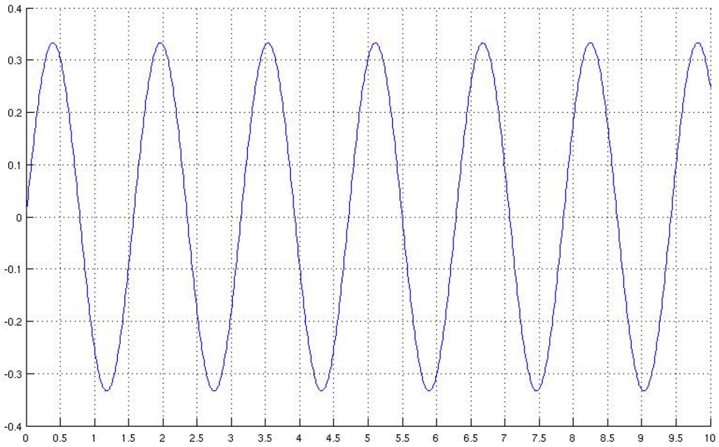
   (c) Model B is more robust than A and C because it will perform best on unseen data.

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

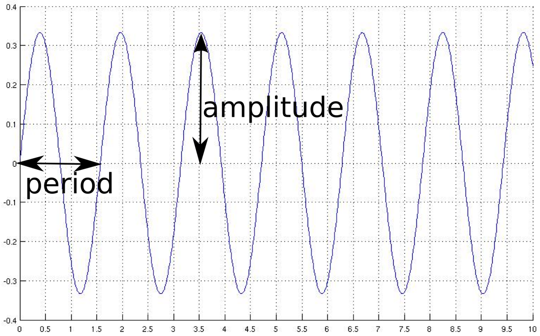
   (e) Model C is overfitting the training data compared to A and B.

   Answer:
   The following are TRUE: (a), (c), and (e)

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



Answer:



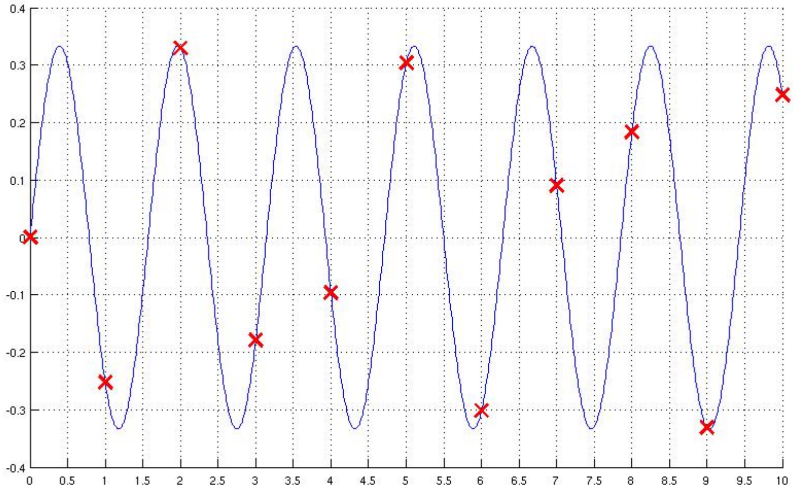
period = 1.57 seconds

frequency = 0.6 [note that we did not label frequency as it is equal to 1/period]

amplitude = 0.3

There will be more on frequency analysis later in the unit.

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 ($2^3 = 8$).



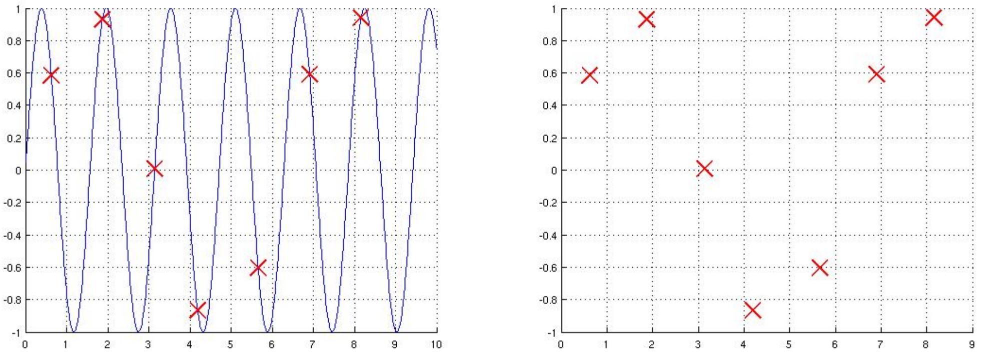
Answer:
Assuming we use 8 levels of sound that correspond to the horizontal grid lines, then each sample is approximated by the binary representation of the closest (rounding) vertical line. The digitisation would thus be:

100 001 111 010 011 111 000 100 101 000 110 (11 in total each corresponding to a sampling point as above)

What is the sampling rate in this case??

Answer:
Sampling rate is 1 Hz (1 sample per second)

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



Answer:
Follow the example above again...

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

    Answer:

    - *11*
    - *6.3*
    - *6*
    - *2*
    - *3*
    - *2 (2 insertions)*
    - *4 (2 substitutions and 2 insertions)*
    - *3 (2 deletions and 1 substitution)*
    - *WUP ('water', ice') = 0.67, WUP('water', tap') = 0.8, WUP('water', 'river') = 0.83, WUP('water', 'liquid') = 0.94*
      D ('water', ice') = 0.33, D('water', tap') = 0.2, D('water', 'river') = 0.17, D('water', 'liquid') = 0.06
      Order = {'ice', 'tap', 'river', 'liquid'}

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.

Answer:

(a) *Calculate the Hamming distance.*

(b) *Calculate the difference in the number of minutes relative to a suitable starting time.*

(c) *You might wish to propose to use the following measure between message M_1 and M_2*

$$D_{NS}(M_1, M_2) = \sum_i \min_j hamming(w_{1i}, w_{2j}) \tag{1}$$

but this is not symmetric (note that distance measures need to be symmetric). For example, if M_1 = {'a', 'c', 'e'} and M_2 = {'b', 'a'}, $D_{NS}(M_1, M_2) = 2$, but $D_{NS}(M_2, M_1) = 1$. One way to make it symmetric is to

$$D_S(M_1, M_2) = (D_{NS}(M_1, M_2) + D_{NS}(M_2, M_1))/2 \tag{2}$$

(d) *You can use dynamic time warping.*

(e) *You can use a similar approach to the one in (c), with a semantic distance measure between words like WUP, where*

$$D(M_1, M_2) = (\sum_i \min_j D_{WUP}(w_{1i}, w_{2j}) + \sum_i \min_j D_{WUP}(w_{2i}, w_{1j}))/2 \tag{3}$$

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)]

Answer:

(a) *x_2 has a very small variance 0.01 and mean close to x_1, so it's probably not very informative (note that high variance often means that there is more information). x_3 has a mean different from x_1, but also significantly high negative correlation (-4; i.e. inversely proportional) thus it is less independent as a variable. x_4 has low covariance with x_1 and large variance, thus would be a good choice – it seems to encode variability not explained by x_1. Therefore x_4 is the variable that should be selected.*

(b) *Lets use x_1 and x_4 for our covariance matrix. Recall from Lecture 2 that to calculate the eigenvalues you need to solve $|A - \lambda\mathbf{I}| = 0$ where $\mathbf{I}$ is the identity matrix and $|A|$ is the determinant of matrix A, with $|A| = (ad - bc)$ for a matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.*

$$|\begin{bmatrix} 4 & -0.1 \\ -0.1 & 9 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}| = 0 \tag{4}$$

$$|\begin{bmatrix} 4-\lambda & -0.1 \\ -0.1 & 9-\lambda \end{bmatrix}| = 0 \tag{5}$$

$$(4-\lambda)(9-\lambda) - 0.01 = 0 \tag{6}$$

$$36 - 13\lambda + \lambda^2 = 0 \tag{7}$$

$$\lambda = \frac{13 \pm \sqrt{169-144}}{2} \tag{8}$$

$$\lambda_1 = 4, \ \lambda_2 = 9 \tag{9}$$

The first eigenvector v_1 is given by $Av = \lambda v$ (with $\lambda = 4$)

$$\begin{bmatrix} 4 & -0.1 \\ -0.1 & 9 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 4 \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} 4v_{11} - 0.1v_{12} \\ -0.1v_{11} + 9v_{12} \end{bmatrix} = \begin{bmatrix} 4v_{11} \\ 4v_{12} \end{bmatrix} \tag{11}$$

We now want to find a solution with vector length of 1 (i.e. $||v_1|| = 1$) [1]

$$-0.1v_{11} + 9v_{12} = 4v_{12} \tag{12}$$

$$v_{11} = 50v_{12} \tag{13}$$

using the vector norm [2] *we get*

$$v_{11} = \frac{50}{\sqrt{2501}} \sim 1 \tag{14}$$

$$v_{12} = \frac{1}{\sqrt{2501}} \sim 0 \tag{15}$$

[1] Here we use the second equation, because the first one leads to a trivial solution (0,0) in which $||v_1|| \neq 1$.

[2] Note that the vector norm is given by $\sqrt{v_{11}^2 + v_{12}^2} = 1$, $\sqrt{2500v_{12}^2 + v_{12}^2} = 1$, $\sqrt{2501}v_{12} = 1$, $v_{12} = \frac{1}{\sqrt{2501}}$. We use the norm to obtain vectors of length 1.

which leads to the following eigenvectors (similarly for $\lambda = 9$)

$$\text{for } \lambda = 4 : v_1 \sim \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{16}$$

$$\text{for } \lambda = 9 : v_2 \sim \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{17}$$

Because v_2 has a larger eigenvalue ($\lambda = 9$) it represents the axes with the most variance, which in turn indicates that x_4 contains the most variance (note that $v_{22} = 1$ and that it represents x_4), consistent with the large variance in x_4 ($\sigma^2 = 9$).

(c)

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \tag{18}$$

$$= \frac{1}{2\pi\sqrt{35.99}} e^{-\frac{1}{2}(\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix})^T \frac{1}{35.99} \begin{bmatrix} 9 & 0.1 \\ 0.1 & 4 \end{bmatrix} (\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix})} \tag{19}$$

$$= 0.0263 \tag{20}$$