COMS21202: Symbols, Patterns and Signals

Problem Sheet 2: Outliers and Deterministic Models

1. You collected a four dimensional dataset of values $\mathbf{x} = (x_1, x_2, x_3, x_4)$ and calculated the mean to be $(3, 2.6, -0.4, 2.6)$. When calculating the covariance matrices for x_1 against itself and the other variables, the following set of covariance matrices was found

	x_1	x_2	x_3	x_4
x_1	$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 0.02 \\ 0.02 & 0.05 \end{bmatrix}$	$\begin{bmatrix} 2 & -1.4 \\ -1.4 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 & 0.5 \\ 0.5 & 3 \end{bmatrix}$

   (a) You were 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 $\mathbf{x}$. Which other variable would you pick: x_2, x_3 or x_4 and why?

   (b) Calculate the eigen values and eigen vectors 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 $\mathbf{x}$ [Note: only use the two variables you picked in (a)]

2. Derive the formulas for least square line fitting presented in slide 17 from Lecture 3.

   You need to prove that solving for the two unknowns a and b from the two equations:

   $\frac{\partial R}{\partial a} = -2\sum_i (y_i - (a + bx_i)) = 0$

   and

   $\frac{\partial R}{\partial b} = -2\sum_i (x_i(y_i - (a + bx_i))) = 0$

   results in the following optimal solution

   $a_{LS} = \bar{y} - b\bar{x}$ and $b_{LS} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2}$

3. For the following 2-D data points:

$$(1, 1)\ \ (3, 2)\ \ (5, 2)\ \ (6, 4)\ \ (7, 4)\ \ (8, 3)\ \ (9, 4)\ \ (10, 5)$$

   (a) Using the **matrix form** for least squares, determine the best fitting line

   (b) Using the **algebric form** for least squares, determine the best fitting line

   (c) Confirm your answers using Matlab

   (d) Using the **matrix form** for least squares, determine the best fitting polynomial $y = a_0 + a_1 x + a_2 x^2$ - Use Matlab to invert the matrix

4. One method to avoid the effect of outliers on means and variances is to use "random sampling". Random sampling selects a sample of points, and estimates the error along with the number of 'outliers'.

   For the set A = {-3, 2, 0, 4, -9, 3, 2, 3, 3, 1, -12, 2}

   Follow this algorithm to estimate the correct mean of this sample (without the effect of outliers)

   Step1: Take 75% of the points at random

   Step2: Calculate the mean of the sampled points

Step3: Estimate the inliers from the set A (i.e. the number of points with Euclidean distance less than ϵ from the mean) [use ϵ = 5 for your tests]

Step4: Recalculate the mean and standard deviation from all inliers

Step5: Repeat for N times [use N = 5 for your tests]

Can you decide on the best optimal mean given your algorithm?

Assume that the outliers in the data were {-9, -12}. Were you able to find the correct mean?

What are the advantages and disadvantages of random sampling?

5. {Extra}: Study the algorithm of RANSAC (Random Sampling Consensus) and see how line fitting can be correctly estimated in the presence of outliers