Department of Computer Science University of Bristol

COMS21202: Symbols, Patterns and Signals

Problem Sheet 2: Outliers and Deterministic Models

1. 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 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 **x** [Note: only use the two variables you picked in (a)]

2. 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 or IPython

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

3. 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)

   Step 1: Take 75% of the points at random

   Step 2: Calculate the mean of the sampled points

   Step 3: 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]. The points with $\epsilon \geq 5$ are outliers.

   Step 4: Recalculate the mean and standard deviation from all inliers

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

   Can you decide on the best mean given your algorithm?

   Assume that the outliers in the data were {-9, -12}. Were you able to find the correct mean (i.e. the mean without the outliers)?

   What are the advantages and disadvantages of random sampling?

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