

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 $(3, 2.6, -0.4, 2.6)$, 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 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)]

Answer:

- (a) x_2 has a very small variance 0.01. You could normalise the data though, but will need to evaluate the correlation again. x_3 has a significantly high negative correlation (i.e. inversely proportional) thus is less independent as a variable. x_4 seems to have a low correlation and large variance, thus would be a good choice.

(b)

$$\left| \begin{bmatrix} 4 & -0.1 \\ -0.1 & 9 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0 \quad (1)$$

$$\left| \begin{bmatrix} 4 - \lambda & -0.1 \\ -0.1 & 9 - \lambda \end{bmatrix} \right| = 0 \quad (2)$$

$$(4 - \lambda)(9 - \lambda) - 0.01 = 0 \quad (3)$$

$$36 - 13\lambda + \lambda^2 = 0 \quad (4)$$

$$(5)$$

The eigen vector will accordingly be,

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix}, \lambda = 4, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda = 9 \quad (6)$$

(c)

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \quad (7)$$

$$= \frac{1}{2\pi\sqrt{35.99}} e^{-\frac{1}{2} \left(\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix} \right)^T \frac{1}{35.99} \begin{bmatrix} 9 & 0.1 \\ 0.1 & 4 \end{bmatrix} \left(\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix} \right)} \quad (8)$$

$$= 0.0263 \quad (9)$$

2. For the following 2-D data points:

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

- Using the **matrix form** for least squares, determine the best fitting line
- Using the **algebraic form** for least squares, determine the best fitting line
- Confirm your answers using Matlab or IPython
- Using the **matrix form** for least squares, determine the best fitting polynomial
 $y = a_0 + a_1x + a_2x^2$ - Use an online calculator to invert the matrix

Answer:

(a) Using the matrix formula

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 10 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \\ 4 \\ 3 \\ 4 \\ 5 \end{bmatrix} \quad \mathbf{a} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\mathbf{a}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 8 & 49 \\ 49 & 365 \end{bmatrix} = \mathbf{H}$$

$$\mathbf{H}^{-1} = \frac{1}{519} \begin{bmatrix} 365 & -49 \\ -49 & 8 \end{bmatrix} = \begin{bmatrix} 0.703 & -0.094 \\ -0.094 & 0.015 \end{bmatrix}$$

$$\mathbf{H}^{-1} \mathbf{X}^T = \begin{bmatrix} 0.6089 & 0.4200 & 0.2312 & 0.1368 & 0.0424 & -0.0520 & -0.1464 & -0.2408 \\ -0.0790 & -0.0482 & -0.0173 & -0.0019 & 0.0135 & 0.0289 & 0.0443 & 0.0597 \end{bmatrix}$$

$$\mathbf{H}^{-1} \mathbf{X}^T \mathbf{Y} = \begin{bmatrix} 0.682 \\ 0.398 \end{bmatrix}$$

(b) $\bar{x} = 6.125$

$$\bar{y} = 3.125$$

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

$$b_{LS} = \frac{179 - 8 \times 6.125 \times 3.125}{365 - 8 \times (6.125)^2} = 0.398$$

$$a_{LS} = \bar{y} - b \bar{x}$$

$$a_{LS} = 3.125 - 0.398 \times 6.125 = 0.682$$

```
>> x = [1; 3; 5; 6; 7; 8; 9; 10];
>> x_f = [ones(8,1), x];
>> y = [1; 2; 2; 4; 4; 3; 4; 5];
>> a = inv(x_f' * x_f) * x_f' * y

a =

    0.6821
    0.3988
```

(c)

```
>> x = [1; 3; 5; 6; 7; 8; 9; 10];
>> x_f = [ones(8,1), x, x.^2];
>> y = [1; 2; 2; 4; 4; 3; 4; 5];
>> a = inv(x_f'*x_f)*x_f'*y

a =

    0.6009
    0.4395
   -0.0037
```

(d)

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)

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 $\epsilon = 5$ 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?

Answer:

Before random sampling, the mean is affected by the outliers $\mu = -0.33$

Step 1: Take 9 out of the 12 points at random. There is a random element in this algorithm so your results might be different

sample = $\{2, 0, -9, 3, 2, 3, 3, 1, 2\}$

Step 2: mean of sample = 0.78

Step 3: Calculate the distances of all points in A from the mean 0.78

distances = $\{3.8, 1.2, 0.8, 3.2, 9.8, 2.2, 1.2, 2.2, 2.2, 0.2, 12.8, 1.2\}$

Thus inliers = $\{-3, 2, 0, 4, 3, 2, 3, 3, 1, 2\}$

Step 4: Calculate the mean and std of inliers

$\mu = 1.70$

sigma = 2.00

Step 5: Repeat for N iterations

<i>iteration</i>	μ	σ	<i>number of outliers</i>
$\{-3, 2, 0, -9, 3, 3, 3, 1, -12\}$	1.44	1.94	3
$\{2, 0, 4, -9, 3, 3, 1, -12, 2\}$	1.70	2.00	2
$\{-3, 2, 0, -0, 3, 2, 3, -12, 2\}$	1.44	1.94	3
$\{-3, 0, -9, 3, 2, 3, 1, -12, 2\}$	1.44	1.94	3
$\{2, 0, -9, 3, 2, 3, 3, 1, 2\}$	1.70	2.00	2
$\{2, 0, -9, 3, 2, 3, 3, 1, 2\}$	1.70	2.00	2
$\{-3, 0, -9, 3, 2, 3, 1, -12, 2\}$	1.44	1.94	3

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