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 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)]

Answer:

(a) x_2 has a very small variance 0.01 and mean close to x_1, so its probably not very informative (note that high variance often means that there is more information). However, by normalising the data you might get a different for x_2 result, but you would need to re-evaluate the covariance matrix. x_3 has mean different from x_1, but also significantly high negative correlation (-4; i.e. inversely proportional) thus it is more dependent on x_1. x_4 has low covariance with x_1 and large variance, thus would be a good choice as it seems to encode variability not explained by x_1. Therefore x_4 is the variable that should be selected because it provides "information" not provided by x_1.

(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 $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

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

$$| \begin{bmatrix} 4-\lambda & -0.1 \\ -0.1 & 9-\lambda \end{bmatrix} | = 0 \tag{2}$$

$$(4-\lambda)(9-\lambda) - 0.01 = 0 \tag{3}$$

$$36 - 13\lambda + \lambda^2 = 0 \tag{4}$$

$$\lambda = \frac{13 \pm \sqrt{169 - 144}}{2} \tag{5}$$

$$\lambda_1 = 4, \ \lambda_2 = 9 \tag{6}$$

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

$$\begin{bmatrix} 4 & -0.1 \\ -0.1 & 9 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 4 \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} 4v_{11} - 0.1v_{12} \\ -0.1v_{11} + 9v_{12} \end{bmatrix} = \begin{bmatrix} 4v_{11} \\ 4v_{12} \end{bmatrix} \tag{8}$$

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

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

$$v_{11} = 50v_{12} \tag{10}$$

using the vector norm [2] *we get*

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

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

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

$$\text{for } \lambda = 4 : v_1 \sim \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{13}$$

$$\text{for } \lambda = 9 : v_2 \sim \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{14}$$

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)

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{15}$$

$$= \frac{1}{2\pi\sqrt{35.99}} e^{-\frac{1}{2}(\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix})^T \frac{1}{35.99} \begin{bmatrix} 9 & 0.1 \\ 0.1 & 4 \end{bmatrix} (\begin{bmatrix} 3 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.6 \end{bmatrix})} \tag{16}$$

$$= 0.0263 \tag{17}$$

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 $y = a_0 + a_1x + a_2x^2$ - Use an online calculator to invert the matrix

Answer:

[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.

(a) *Using the matrix formula*

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

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

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

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

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

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

(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)

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?

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

iteration	μ	σ	*number of outliers*
{-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

As you can see from the example above, whether we can identify the outliers depends on how many iterations we do, but also crucially on the ϵ chosen.

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