Questions 3

Dr. Roman Belavkin

BIS4435

Question 1

What is a model and its error, and how can they be defined in terms of information?

Answer: *A model is a simplified representation of another object. Because the model contains fewer details and fewer number of states, it is less complex, and, therefore, it is less uncertain than the object it represents. It means also that the model can be described using less information than the original object. The greater certainty makes the model more predictable, which can be used to study and predict the real object. However, the loss of information leads to errors, which are the differences between the model's predictions and the behaviour of the real object.*

Question 2

What are the linear models, and why are they called 'linear'?

Answer: *Linear models are represented by linear functions. In the simplest case, these functions represent a line on a plane, hence the name 'linear'. In more complex cases, linear functions can represent a plane in a 3–dimensional space and a hyperplane in more dimensions.*

Question 3

What are the data–driven models, and what does the data represent?

Answer: *Data–driven models are based on data that represents the object to be modelled. Usually, these models use large datasets from datawarehouses. The data in this case is the reality that the model has to simplify. More precisely, the data represents measurements of the object of interest or the part of reality that is to be modelled.*

Question 4

Why plotting the data on a chart can be useful?

Answer: *Charts visualise the data and can be used to check for some visible relationships in the data before a model is created. Their use, however, is limited to two or three dimensions.*

Question 5

How can the errors of the model be measured? Why is it important for the model?

Answer: *The error is the difference between the data (y) and the model's prediction of this data ($f(x)$). The difference is measured by the cost function $c(y, f(x))$. There can be different cost functions:*

- *Absolute error*
$$c(y, f(x)) = |y - f(x)|$$
- *Squared (or quadratic) error*
$$c(y, f(x)) = (y - f(x))^2$$
- *Binary error*
$$c(y, f(x)) = \begin{cases} 0, & \text{if success (e.g. } y = f(x)) \\ 1, & \text{otherwise} \end{cases}$$

Which cost function is used is important, because it influences the choice of the optimal model. For the same data, the mean absolute error can be minimised by one model, and the mean square error by another.

Question 6

Compute linear mean-square model for the following set of data:

Monthly Income (£K)	Home Owner
2	0
1	0
6	1

Answer: *Let us denote Monthly Income by x and Home Ownership by y. The computations can be done as follows:*

a) *Compute expected values $E\{x\}$ and $E\{y\}$*
$$E\{x\} = \frac{2+1+6}{3} = 3\,, \quad E\{y\} = \frac{0+0+1}{3} = 1/3$$

b) *Find covariance Cov(x,y)*

$$Cov(x,y) = \frac{(2-3)(0-1/3)+(1-3)(0-1/3)+(6-3)(1-1/3)}{3} = 1$$

c) *Find variance Var(x)*

$$Var(x) = \frac{(2-3)^2+(1-3)^2+(6-3)^2}{3} = 14/3$$

d) *Find the slope*

$$b = \frac{Cov(x,y)}{Var(x)} = \frac{3}{14} = 0,21$$

e) *Find the intercept*

$$a = E\{y\} - bE\{x\} = -0,31$$

The resulting linear model is

$$f(x) = -0,31 + 0,21x$$

Question 7

Suppose that a multiple regression model using $m = 6$ input variables should be created based on some data. What is the minimum number of cases the dataset should have?

Answer: *The dataset should have at least $m + 2 = 8$ cases. This is because the mean-square regression requires several points not laying in one hyperplane. A linear function of m variables describes a hyperplane in a $m+1$ dimensional space. Thus, $m + 1$ points always lay in one hyperplane, and we need one extra point not laying on this hyperplane. Hence, a minimum $m + 2$ cases are required.*

Question 8

Consider the following data:

Monthly Income (£K)	Monthly Expenses (£K)	Home Owner	Credit Score
2	1	0	3
1	2	0	1
6	2	1	5
3	1	1	4
3	2	0	2

Suppose you are building a linear mean-square model based on this data to predict the customers' credit scores. Let us denote the three input variables (the first three columns of the table) as x_1, x_2 and x_3, while the output variable as y. Answer the following questions:

a) Does this dataset contain sufficient number of cases to build the model?

b) The regression coefficients for this model are $b_1 = 0,69$, $b_2 = -1,31$ and $b_3 = 0,56$. Write the linear equation for $f(x_1, x_2, x_3)$.

c) Using the model, compute the credit score for the following customer:

Monthly Income (£K)	Monthly Expenses (£K)	Home Owner	Credit Score
5	3	1	?

Answer:

a) *The dataset must have at least $m + 2$ different cases, where m is the number of input variables. So, the minimum number of cases is 5. Hence, the dataset contains the required number (the table has five rows).*

b) *The we can use the following form of the linear mean-square model:*

$$y = E\{y\} + b_1(x_1 - E\{x_1\}) + b_2(x_2 - E\{x_2\}) + b_3(x_3 - E\{x_3\})$$

The values of the regression coefficients are already given, so we only need to compute the mean values for each variable (the means approximate the expected values)

$$E\{x_1\} = 3\,, \quad E\{x_2\} = 1,6\,, \quad E\{x_3\} = 0,4\,, \quad E\{y\} = 3$$

The expected values represent the centre of gravity point. The linear model is

$$y = 3 + 0,69(x_1 - 3) - 1,31(x_2 - 1,6) + 0,56(x_3 - 0,4)$$

c) *In this case, $x_1 = 5$, $x_2 = 3$ and $x_3 = 1$. The credit score is*

$$y = 3 + 0,69(5 - 3) - 1,31(3 - 1,6) + 0,56(1 - 0,4) = 2,88$$

Question 9

How to test whether a model is good in forecasting?

Answer: *Usually, the dataset is split into two parts. One is used for creating the model, and another for testing the model. Because the model has not 'seen' the cases from the testing part of the data, they can be used to estimate how good the model is for forecasting. The errors between the testing data and the model's predictions can estimate how reliable is the forecast.*

Question 10

Compute correlation for the following set of data:

Monthly Expenses (£K)	Home Owner
1	0
2	1
1	1
2	0

What does the value of the correlation mean in this case?

Answer: *The correlation is defined as*

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

We can compute it as follows:

a) *Find the means:*

$$\begin{aligned} E\{x\} &= \frac{1+2+1+2}{4} = \frac{3}{2} \\ E\{y\} &= \frac{0+1+1+0}{4} = \frac{1}{2} \end{aligned}$$

b) *Find the covariance*

$$\begin{aligned} Cov(x,y) &= \frac{1}{4}[(1-3/2)(0-1/2) + (2-3/2)(1-1/2) \\ &+ (1-3/2)(1-1/2) + (2-3/2)(0-1/2)] = 0 \end{aligned}$$

c) *Find the variances*

$$\begin{aligned} Var(x) &= \frac{(1-3/2)^2 + (2-3/2)^2 + (1-3/2)^2 + (2-3/2)^2}{4} = \frac{1}{4} \\ Var(y) &= \frac{(0-1/2)^2 + (1-1/2)^2 + (1-1/2)^2 + (0-1/2)^2}{4} = \frac{1}{4} \end{aligned}$$

d) *The correlation is*

$$Corr(x, y) = \frac{0}{\sqrt{1/41/4}} = 0$$

Zero correlation means that the variables are uncorrelated, and, therefore, they are not related linearly. This, however, does **not** *mean that the variables are independent. There still can be some non-linear dependency between the variables.*

Question 11

What are the outliers? Why is the mean-square model sensitive to outliers? What does it mean for the mean-square model?

Answer: *Outliers are the cases in the dataset that are far from the expected value (or from the centre of gravity). They also lay far from the hyperplane represented by the linear model. Because the mean-square model uses the square of the distance as the measure of the error, it is very sensitive to the outliers. Few outliers can increase significantly the average error, and change the model and its performance. The outliers are the most unusual cases in the database, and, therefore, they are the most interesting from information point of view.*

Question 12

What are the main differences between the mean-square and the mean absolute models?

Answer:

- *The models are based on minimising errors measured by different cost functions: Absolute or mean-squared errors.*
- *The mean absolute model may not have a unique optimal solution, while the mean-square model always has a unique optimal solution.*
- *There is no analytical solution for the mean absolute model. The solution for the mean-square model can be found using the method of least squares (or the least squares regression).*
- *The mean absolute model is robust to outliers, while the mean-square model is not.*