Data Mining

Regression

Academic year 2019/2020

Queen Mary University of London

EXERCISES

A data scientist without computing power is akin to a surgeon without a scalpel. The following exercises involve numerical calculations and you might be able to do some of them by hand. However our focus is on data science concepts rather than calculations and anyway, why would you want to do it by hand when computers can do it for you? Get familiar with a computing environment (for instance Matlab, Python, R) and start using it! After all, any future data science project you will be working on will involve many data points and calculations, and most likely you will not be able to cope with them by hand.

EXERCISE ♯1. Consider the following simple dataset, where ID denotes the sample identifier and x and y are two attributes:

ID	x	y
1	2	1
2	3	2
3	1	2
4	1	1
5	0	1
6	5	3
7	4	3
8	6	7
9	5	6
10	3	5

In this exercise we will use this dataset to test the following five linear models:

- $f_1(x) = 2x + 1$
- $f_2(x) = x$
- $f_3(x) = 2x - 1$
- $f_4(x) = x - 0.5$
- $f_5(x) = x + 0.5$

Which candidate model would you choose if you use the MSE to quantify their quality? Which one would you choose if you use the MAE? Note that MSE and MAE could, but in general will not, produce the same solution.

EXERCISE ♯2. In this exercise, we will evaluate the MSE surface of a generic linear model $y = w_0 + w_1 x$. The MSE surface represents the dependence of the MSE on the values of the model's parameters, in this case w_0 and w_1. We will simplify this study by fixing one parameter and calculating the MSE for different values of the other parameter. Consider the dataset $\{x_i, y_i), 1 \leq i \leq 10\}$ presented in the previous exercise.

- Assuming $w_0 = 0$, calculate the MSE of 9 linear models where w_1 =0, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5. Plot the MSE values vs w_1 and identify the value of w_1 that minimises the MSE.

- Assuming $w_1 = 1$, calculate the MSE of 9 linear models where $w_0 =$ -4, -3, -2, -1, 0, 1, 2, 3, 4, 5. Plot the MSE values vs w_0 and identify the value of w_0 that minimises the MSE.
- Can you guess the shape of the MSE surface as a function of both w_0 and w_1 by looking at the MSE curves previously obtained? Sketch the MSE surface.
- Obtain the exact solution for w_0 and w_1 that minimises the MSE on the dataset. Compare the exact solution with the values obtained by exploring the MSE surface.
- Plot the MSE surface corresponding to all the pairs w_1 and w_0, where w_1 =0, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5 and $w_0 =$ -4, -3, -2, -1, 0, 1, 2, 3, 4, 5. Identify the MSE of the exact solution on the MSE surface.

EXERCISE ♯3. Show that the linear model $y = w_0 + w_1 x$ that minimises the sum of the squared errors in a dataset $\{x_i, y_i), 1 \leq i \leq N\}$, is defined by the coefficients

$$\begin{aligned} w_1 &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \\ w_0 &= \bar{y} - w_1 \bar{x} \end{aligned}$$

where $\bar{x} = \sum_{i=1}^{N} x_i$ and $\bar{y} = \sum_{i=1}^{N} y_i$.

EXERCISE ♯4. Show that the linear model $y = \boldsymbol{w}^T \boldsymbol{x}$ that minimises the sum of the squared errors in a dataset $\{\boldsymbol{x}_i, y_i), 1 \leq i \leq N\}$, is defined by the coefficients $\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{y}$, where X and y are obtained from the dataset as described in the lecture notes.

EXERCISE ♯5. Show that the exact solutions provided in the two previous exercises are equivalent for a simple linear regression model.

EXERCISE ♯6. Consider the following simple datasets:

Table 1

ID	x	y
1	1	1
2	2	2

Table 2

ID	x	y
1	1	1
2	2	2
3	3	3

Table 3

ID	x	y
1	1	1
2	2	2
3	2	2.2
4	2	1.9
5	2	2.3
6	2	2

Table 4

ID	x	y
1	1	1
2	1	0.8
3	1	1.2
4	2	1.9
5	2	2.3
6	2	2

Table 5

ID	x	y
1	1	1
2	1	0.8
3	1	1.2
4	1	1.1
5	1	0.7
6	2	2

Let us formulate a regression problem for each dataset, taking x as the predictor and y as the label.

- Plot each dataset and suggest a linear solution based on visual inspection alone (note that implicitly you will be using a notion of *quality*!).
- Using the notation developed for multiple linear regression problems, create the matrix X and vector $\boldsymbol{y}$ for each dataset.

- Obtain the coefficients of the regression model $y = w_0 + w_1 x$ by using the least squares solution $\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{y}$ and plot the resulting linear models.
- Write down the matrices $X^T X$, $(X^T X)^{-1}$ and $(X^T X)^{-1} X^T$. How can you quantify the dependence of each coefficient on the individual samples in the dataset? How would you interpret the strength of such dependence for each dataset?

EXERCISE ♯7. Consider the following simple dataset:

ID	x	y
1	1	2.2
2	2	3.5
3	3	3.9
4	5	2.9
5	7	5
6	8	6.2
7	2.5	3
8	8	4.8

Let us consider a regression problem for this dataset, taking x as the predictor and y as the response. We will use this dataset to train different polynomial models and then will calculate the train MSE.

- Write down the mathematical expressions for a linear solution, a quadratic solution and a cubic solution, and identify their coefficients (i.e. the models' parameters).
- Using the notation developed for multiple linear regression problems, create the matrices X and vectors $\boldsymbol{y}$ corresponding to each one of the models.
- Obtain the coefficients for each model by using the solution $\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{y}$ and plot the resulting models.
- Calculate the train MSE of each solution and compare them. Which one would you say is the best model?
- What would you expect the train MSE to be for a polynomial model of order 8 or higher?
- Obtain the coefficients of a polynomial model of order 8, plot it and calculate its train MSE.

EXERCISE ♯8. The file *AnimalsHRvsBM.csv* contains the body mass and resting heart rate of several animals, as recorded by Noujaim et al. in their 2004 article *From Mouse to Whale: A Universal Scaling Relation for the PR Interval of the Electrocardiogram of Mammals* published in the scientific journal Circulation. Let us consider a regression problem aiming at predicting the heart rate of an animal based on its weight.

- Plot all the samples in the dataset, analyse its distribution and suggest a model that fits it. Can you see outliers? What type of model would you use (linear, quadratic...)?
- Note that the values of the body mass attribute span several orders of magnitude, from tens to millions of grams. Create a new dataset that contains the logarithm of the body mass and the resting heart rate of the animals in the original dataset. Plot the new dataset as a point cloud and suggest a regression model by visually inspecting the point cloud.

- Fit a linear model, a quadratic model and a cubic model to the new dataset. Plot the solutions and analyse them.
- Create a new dataset that contains the logarithm of the body mass and the logarithm of the resting heart rate as recorded in the original dataset. Plot this new dataset as a point cloud and suggest a regression solution by visually inspecting it.
- Fit a linear model to this last dataset and use this model to obtain a new model that fits the original dataset. Note that this is called *exponential regression*!

EXERCISE ♯9. Consider four polynomial models of degrees 1, 2, 3 and 4 and three datasets consisting of 8 samples $\{(x_i, y_i), 1 \leq i \leq N\}$ and created as follows:

- **Dataset 1**: x_i is randomly drawn from a uniform distribution $U(0, 1)$ and $y_i = 3x_i$.
- **Dataset 2**: x_i is randomly drawn from a uniform distribution $U(0, 1)$ and $y_i = 3x_i + n_i$, where n_i is randomly drawn from a gaussian distribution $N(2, \sigma^2)$.
- **Dataset 3**: x_i and y_i are both randomly drawn from a uniform distribution $U(0, 1)$.

Assume that the first four samples of each dataset are used to train the four polynomial models using the MMSE criterion and the remaining four samples are used for testing. Answer the following questions for each dataset:

- How will the training MSE and test MSE change with the order of the MMSE polynomial?
- What do you expect the coefficients of each MMSE polynomial to be?
- What would be the impact of increasing the number of training and testing samples from 4 to 50, on the MSE values and the coefficients of each MMSE polynomial?

EXERCISE ♯10. Consider a dataset consisting of 100 samples $\{(x_i, y_i), 1 \leq i \leq 100\}$, 50 of which will be used for training and 50 for testing. The following options are proposed to create the training dataset and the test dataset:

- The first 50 samples in the dataset will be used for training and the remaining 50 for testing.
- 50 numbers between 1 and 100 will be randomly generated without repetition. Samples whose index i is one of those numbers will be used for training, the remaining 50 samples will be used for testing.
- Samples whose index i is an even number will be used for training, whereas samples indexed by an odd number will be used for testing.

Which option would you implement and why?

EXERCISE ♯11. A fraction F of a dataset is used for training a model, leaving the fraction $1 - F$ of the dataset for testing. Discuss the impact on the training MSE and test MSE of:

- Values of F close to 1.
- Values of F close to 0.

EXERCISE ♯12. Consider a dataset consisting of 100 samples $\{(x_i, y_i), 1 \leq i \leq 100\}$, where y_i is related to x_i by the model $y_i = x_i + n_i$ and n_i is a random number drawn from a gaussian distribution $N(0, \sigma^2)$, where $\sigma^2 = 1$. In addition, the values x_i are distributed uniformly from 0 to 2.

Assume that we want to fit a linear model $f(x) = w_0 + w_1 x$ by using the MMSE criterion and let us write $y = f(x) + e$.

- What is the nature of the error e?
- What do you expect the coefficients of the MMSE solution to be?
- What value would you expect for the test MSE?

Now assume that the true underlying model is $y_i = x_i + x_i^2 + n_i$ and we fit the linear model above. Answer the previous questions for this new scenario.

EXERCISE ♯13. Consider a dataset $\{(x_i, y_i)\}$ where y_i is a continuous label and x_i is a discrete attribute that can take up to 4 different values A, B, C and D. As a discrete attribute, there is no notion of distance nor ordering between any 2 of its possible values. How could you create a regression model that predicts the label y based on the attribute vector x?

EXERCISE ♯14. Consider a regression scenario where we are interested in predicting several labels. Using vector notation, we represent the response as $\boldsymbol{y}$ to signify that the output consists of several quantities. Given a dataset $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, how would you create a regression model that predicts $\boldsymbol{y}$ from $\boldsymbol{x}$?