Datamining Assignment 1

Elkhan Ismayilzada

March 26, 2021

Exercise 1

The goal of learning in data mining is to extract potentially useful information by finding out patterns from large datasets and then to transform it to an understandable structure for future use. There are several algorithms to use for this purpose. What they aim for is to discover the best parameters for a function which will then generalize the data with an error as small as possible. A model stores information that we found from the dataset such as the patterns derived from analysis of data. By using the trained model, we can predict the outcome of new data.

Exercise 2

1. In supervised learning, the data we have received already has categories. That is to say, we have labeled training data as well as test data which then can be used to determine if our function is right or wrong. In addition, there is a relationship between the input and output. The data has target outcome that tells us what to predict and how our correct output should look like.

2. In unsupervised learning, on the other hand, the data has no labels. In this case, we have no idea how our results should look like since there is no target outcome. Therefore, to get useful information from this data, we let the model create the categories for us.

3. Regression analysis is useful when we want to answer the question of "how much?". For example, if we want to predict housing prices then we should use regression. More formally, in a regression problem, we predict results within a continuous output by mapping input variables to a continuous function.

4. In a classification problem, we want to answer the question of "is it?" or "does it?". For example, if we want to know whether the patient has tumor or not, we should use classification. In a more formal description, we try to predict results within a discrete output by mapping the input variables to discrete categories.

5. In a clustering problem, we try to make up categories for unlabeled data by grouping data points which are similar to one another. Famously known application of clustering is social network analysis. Given your Facebook friends,for instance, we can identify which are the cohesive groups of friends or which are the groups of people in which everyone knows each other

Exercise 3

We were asked to prove

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \tag{3.1}$$

Error sum of squares (SSE) is given by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

And since

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

We have

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

If we want to minimize this, then we should take derivative with respect to the parameters and equal it to zero, hence

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

Which is same as (after dividing both sides by -2)

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Now we take derivative with respect to β_1

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Which is same as (after dividing both sides by -2)

$$\sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0$$

And since we know that

$$x_i = \frac{1}{\hat{\beta}_1}\hat{y}_i - \frac{\hat{\beta}_0}{\hat{\beta}_1}$$

If we substitute x_i with this then

$$\sum_{i=1}^{n} (\frac{1}{\hat{\beta}_1}\hat{y}_i - \frac{\hat{\beta}_0}{\hat{\beta}_1})(y_i - \hat{y}_i) = 0$$

After opening up the parentheses we get

$$\frac{1}{\hat{\beta}_1} \sum_{i=1}^{n} \hat{y}_i y_i - \frac{1}{\hat{\beta}_1} \sum_{i=1}^{n} \hat{y}_i \hat{y}_i - \frac{\hat{\beta}_0}{\hat{\beta}_1} \sum_{i=1}^{n} y_i + \frac{\hat{\beta}_0}{\hat{\beta}_1} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{\hat{\beta}_1} \sum_{i=1}^{n} \hat{y}_i(y_i - \hat{y}_i) - \frac{\hat{\beta}_0}{\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

And since we know that the second term is zero (after clearing constants)

$$\sum_{i=1}^{n} \hat{y}_i(y_i - \hat{y}_i) = 0$$

If we open up the parentheses of (3.1)

$$\sum_{i=1}^{n} y_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i\hat{y}_i + \bar{y}\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i(y_i - \hat{y}_i) - \bar{y}\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Exercise 4

4.1

Firstly, we are going to prove expected values of the estimators $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Notice that numerator is the sample covariance between x and y and denominator is sample variance of x. If we open up the parentheses of numerator we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Notice that the summation of the second term of the numerator is zero because

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

So, in the end we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Now we can find the expected value of β_1

$$E[\hat{\beta}_1] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} E[y_i]$$

And since

$$E[y_i] = \beta_0 + \beta_1 x_i$$

We have

$$E[\hat{\beta}_1] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i)$$

If we open the parentheses, we get

$$E[\hat{\beta}_1] = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\beta_0 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\beta_1$$

Notice that because of the numerator, the first term goes to zero, the numerator and the denominator of the second term are equal, therefore

$$E[\hat{\beta}_1] = \beta_1$$

Now, we can move on to expected value of $\hat{\beta}_0$

We know that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Therefore,

$$E[\hat{\beta}_0] = E[\bar{y} - \hat{\beta}_1 \bar{x}] = E[\bar{y}] - E[\hat{\beta}_1]\bar{x}$$

We can write the first term as

$$E[\bar{y}] = \frac{1}{n}\sum_{i=1}^{n} E[y_i] = \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)$$

And since the expected value of $\hat{\beta}_1$ is β_1 after substitution

$$E[\hat{\beta}_0] = \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

If we open up the parentheses of the first term

$$\frac{1}{n}\sum_{i=1}^{n} \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n\beta_0}{n}$$

As a result

$$E[\hat{\beta}_0] = \beta_0$$

Now, we are going to prove the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$

For β_1

$$Var(\hat{\beta}_1) = Var(\frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2})$$

We can substitute denominator with S_{xx} to make it look more simple

$$Var(\hat{\beta}_1) = Var(\frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}})$$

Let's take S_{xx} and other x related parts out of Var

$$Var(\hat{\beta}_1) = \frac{1}{(S_{xx})^2}\sum_{i=1}^{n} (x_i - \bar{x})^2 Var(y_i) = \frac{1}{(S_{xx})^2} S_{xx}\sigma^2$$

After cancelling out S_{xx} we get

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Finally, for $\hat{\beta}_0$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x}) = Var(\bar{y}) + Var(-\hat{\beta}_1 \bar{x}) + 2Cov(\bar{y}, -\bar{x}\hat{\beta}_1)$$

The variance of $\hat{y}$ can be found by

$$Var(\hat{y}) = Var(\frac{y_1 + y_2 + \cdots + y_n}{n}) = Var(\frac{1}{n}y_1 + \frac{1}{n}y_2 + \cdots + \frac{1}{n}y_n) = \frac{1}{n^2}Var(y_1) + \frac{1}{n^2}Var(y_2) + \cdots + \frac{1}{n^2}Var(y_n)$$

if we substitute with σ

$$Var(\bar{y}) = \frac{1}{n^2}[\sigma^2 + \sigma^2 + \cdots + \sigma^2] = \frac{1}{n^2}[n\sigma^2] = \frac{\sigma^2}{n}$$

Since we know the variance of $\hat{\beta}_1$

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \hat{x}^2(\frac{\sigma^2}{S_{xx}}) - 2\hat{x}Cov(\bar{y}, \hat{\beta}_1)$$

The covariance of the two will be zero because

$$Cov(\bar{y}, \hat{\beta}_1) = Cov(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{j=1}^{n} \frac{x_j - \bar{x}}{S_{xx}} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{x_j - \bar{x}}{nS_{xx}} Cov(y_i, y_j) = 0$$

Now if we merge them in one fraction

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2(\frac{\sigma^2}{S_{xx}}) = \sigma^2 \frac{S_{xx} + n\bar{x}^2}{nS_{xx}}$$

Now if replace S_{xx} in the numerator with its formula and open up the parentheses we get

$$Var(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) + n\bar{x}^2}{nS_{xx}}$$

Notice that we can cancel out something in the numerator

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + 2n\bar{x}^2 = \sum_{i=1}^{n} x_i^2$$

So in the end we have

$$Var(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}$$

4.2

The reason why we have to use t-test is because we do not know the population variance (σ^2). We never observe it. Therefore, we have to use the estimate of σ^2. By using t-test, we can get information about how significance each feature variables are. That is to say, how much they affect to the output.

Exercise 5

We have

$$Y = \hat{\beta}X \tag{5.1}$$

The matrix X is not a square matrix so we multiply both sides with its transpose

$$X^T Y = \hat{\beta}X^T X \tag{5.2}$$

If we multiply both sides with $(X^T X)^{-1}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5.3}$$

Let's name $(X^TX)^{-1}X^T$ as c (c stands for constant) so that

$$\hat{\beta} = cY \tag{5.4}$$

The variance of $\hat{\beta}$ is

$$Var(\hat{\beta}) = Var(cY) = c^T cVar(Y) = c^T\sigma^2 I_n c \tag{5.5}$$

This means

$$Var(\hat{\beta}) = (X^TX)^{-1}X^T\sigma^2 I_n((X^TX)^{-1}X^T)^T \tag{5.6}$$

And using the fact that both X^TX and its inverse are symmetric

$$((X^TX)^{-1})^T = (X^TX)^{-1} \tag{5.7}$$

Which means

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \tag{5.8}$$

As a result

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1} \tag{5.9}$$

So let V_{ii} be i^{th} diagonal matrix of $(X^TX)^{-1}$ then

$$Var(\hat{\beta}) = \sigma^2 V_{ii} \tag{5.10}$$

Exercise 6

We have

$$L = \frac{1}{(2\pi\sigma_{ML}^2)^{\frac{N}{2}}}\exp[\frac{-\sum_{t=1}^{N}(x_t-\mu_{ML})^2}{2\sigma_{ML}^2}]$$

If we take log for both sides then

$$\log L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_{ML}^2) - \sum_{t=1}^{N}\frac{(x_t-\mu_{ML})^2}{2\sigma_{ML}^2}$$

Now if we take derivative of it with respect to μ_{ML} and make it zero then

$$\frac{\partial L}{\partial \mu_{ML}} = \frac{1}{\sigma_{ML}^2}\sum_{t=1}^{N}(x_t-\mu_{ML}) = 0$$

If we open up parentheses then

$$\frac{1}{\sigma_{ML}^2}\sum_{t=1}^{N}x_t - \frac{N\mu_{ML}}{\sigma_{ML}^2} = 0$$

After cancelling out σ_{ML}^2 and finding μ_{ML} we get

$$\mu_{ML} = \frac{1}{N}\sum_{t=1}^{N}x_t$$

In the same way, we take derivative with respect to σ_{ML}^2

$$\frac{\partial L}{\partial \sigma_{ML}^2} = \frac{1}{2\sigma_{ML}^4}\sum_{t=1}^{N}(x_t-\mu_{ML})^2 - \frac{N}{2\sigma_{ML}^2} = 0$$

Now if we take second term to right hand side and multiply both sides with $2\sigma_{ML}^4$ we get

$$\sum_{t=1}^{N} (x_t - \mu_{ML})^2 = N\sigma_{ML}^2$$

From here we can easily get σ_{ML}^2

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{t=1}^{N} (x_t - \mu_{ML})^2$$

Exercise 7

7.1

The equation to find estimators is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

So after putting the values of X and Y we get

$$\hat{\beta} = \left(\begin{bmatrix} 1 & -2 & 0 & 3 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -2 & -2 \\ 0 & 1 \\ 3 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & -2 & 0 & 3 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

And after calculation

$$\hat{\beta} = \begin{bmatrix} \frac{43}{59} \\ -\frac{21}{59} \end{bmatrix}$$

So equation will look like

$$\hat{y} = \frac{43}{59} x_1 - \frac{21}{59} x_2$$

7.2

With intercept, we should prepend one more column to the matrix X so it becomes

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & -2 & -2 \\ 1 & 0 & 1 \\ 1 & 3 & 1 \end{bmatrix}$$

So after putting the values of X and Y we get

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -2 & 0 & 3 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & -2 & -2 \\ 1 & 0 & 1 \\ 1 & 3 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -2 & 0 & 3 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

And after calculation

$$\hat{\beta} = \begin{bmatrix} \frac{155}{106} \\ \frac{36}{53} \\ -\frac{32}{53} \end{bmatrix}$$

So equation will look like

$$\hat{y} = \frac{155}{106} + \frac{36}{53} x_1 - \frac{32}{53} x_2$$

7.3

If we use Python library for fitting the data we get these results (Figure 1, Figure 2) and if we change the fractions we found from matrix calculation to its decimal version, they all match with the values that are found by using python library.

```
# Import necessary libraries . Only the following will be needed
import numpy as np
from sklearn import linear_model

X = np.array ([[1 ,2] ,[ -2 , -2] ,[0 ,1] ,[3 ,1]])
y = np.array ([0 ,1 ,2 ,3])
```

```
model = linear_model.LinearRegression(fit_intercept=False)
model.fit(X,y)
```

```
LinearRegression(fit_intercept=False)
```

```
print(model.coef_,model.intercept_)
```

```
[ 0.72881356 -0.3559322 ] 0.0
```

Figure 1: Without intercept

```
# Import necessary libraries . Only the following will be needed
import numpy as np
from sklearn import linear_model

X = np.array ([[1 ,2] ,[ -2 , -2] ,[0 ,1] ,[3 ,1]])
y = np.array ([0 ,1 ,2 ,3])
```

```
model = linear_model.LinearRegression(fit_intercept=True)
model.fit(X,y)
```

```
LinearRegression()
```

```
print(model.coef_,model.intercept_)
```

```
[ 0.67924528 -0.60377358] 1.4622641509433962
```

Figure 2: With intercept