Exercises for exercise class 7a in MMS075, Mar 3, 2020

1. Once again, assume the same background story and data as described in Exercise 1 in Exercise class 1: a hypothetical company called Maintain-IT is responsible for a project task that needs to be repeated every year. They want to determine how the number of employees in the project affects the completion time, based on the following observations:

Year	Employees in project	Completion time (days)
1	70	20
2	30	60
3	10	100
4	90	20

For a simple linear regression model, we have computer that the least square coefficients are -1 for the slope and 100 as the intercept. Compute the training mean squared error for this model! The formula for MSE is given below:

$$MSE = \frac{(y_1 - \hat{f}(x_1))^2 + (y_2 - \hat{f}(x_2))^2 + \ldots + (y_n - \hat{f}(x_n))^2}{n}$$

This formula contains the predicted values for x_1, x_2, …, x_n by the model, and these predictions for simple linear regression can be computed as usual:

$$\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

2. The air conditioner company discussed in Exercise class 6b has constructed a logistic regression model predicting the probability of air conditioning. They define a model predicting "Yes" for a house if the estimated probability of air conditioning is at least 20% and "No" otherwise – this way, they can target those houses with commercials where they can be sufficiently certain that the house does not yet contain air conditioning. They obtain some new information about 10 houses, see below, and as a new column, add the predicted probability of air conditioning based on their model into the last column.

	price	stories	airco	gashw	Prob
1	45000	2	no	no	15%
2	48500	1	no	no	13%
3	52000	1	no	no	15%
4	53900	2	no	no	21%
5	60000	2	yes	no	25%
6	61000	2	no	no	26%
7	64500	2	no	no	29%
8	71000	2	no	no	35%
9	75500	1	yes	no	33%
10	33500	1	no	no	7%

Evaluate their classification model by the following steps:

a) Insert the predicted response of “Yes” or “No” in a new column
b) Determine the error rate for all houses in the new data
c) Determine the error rate for those houses in the new data that do not have air conditioning
d) Determine the error rate for those houses in the new data that have air conditioning

The formula for computing error rate is as follows:

$$\text{Average}\left[I(y^{\text{new}} \neq \hat{y}^{\text{new}})\right]$$

In other words, you need to determine the percentage of new points for which the model gives wrong predictions.

3. Consider a simple model A and a very flexible model B. Evaluate whether the following inequalities always hold:
   a. Training MSE for model B ≤ Training MSE for model A;
   b. Test MSE for model B ≤ Test MSE for model A.

4. In the first exercise, we computed the training MSE for the simple linear regression model predicting completion time based on the number of employees in the project. Perform Leave-One-Out Cross-Validation (LOOCV) on this dataset to estimate the test MSE! The coefficients of the simple linear regression models that are considered in this process are provided in the R outputs below.

Output 1: The coefficients for the model based on data from years 2,3,4:

```
Call:
lm(formula = CompTime ~ Employees, subset = c(2, 3, 4))

Coefficients:
(Intercept)    Employees
    100.000       -0.923
```

Output 2: The coefficients for the model based on data from years 1,3,4:

```
Call:
lm(formula = CompTime ~ Employees, subset = c(1, 3, 4))

Coefficients:
(Intercept)    Employees
     107.69        -1.08
```

Output 3: The coefficients for the model based on data from years 1,2,4:

```
Call:
lm(formula = CompTime ~ Employees, subset = c(1, 2, 4))

Coefficients:
(Intercept)    Employees
     78.571       -0.714
```

Output 4: The coefficients for the model based on data from years 1,2,3:

```
Call:
lm(formula = CompTime ~ Employees, subset = c(1, 2, 3))

Coefficients:
(Intercept)    Employees
     107.14        -1.29
```

5. Feedback quiz (optional): Go to www.menti.com and use the code 36 48 46.