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:

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

In order to determine MSE, we need to get the predicted values for for x_1, x_2, …, x_n. The completion time that the model would predict for year 1 is as follows:

$$\widehat{\text{Time}} = 100 - 1 \times \text{Employees} = 100 - 70 = 30.$$

Therefore, the first term in the numerator of MSE is (20-30)^2, and the other terms can be computed similarly. It is convenient to include the predictions, differences and squared differences in a table extending the table above:

Year	Employees in project	Completion time (days)	Predicted completion time	Observation - prediction	Squared difference
1	70	20	30	-10	100
2	30	60	70	-10	100
3	10	100	90	10	100
4	90	20	10	10	100

The training MSE is the average of the values in the last column:

$$MSE = \frac{100+100+100+100}{4} = 100.$$

Therefore, the training MSE for this model is 100.

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	Predicted response about airco	$I(y^{\text{new}} \neq \hat{y}^{\text{new}})$
1	45000	2	no	no	15%	No	0
2	48500	1	no	no	13%	No	0
3	52000	1	no	no	15%	No	0
4	53900	2	no	no	21%	Yes	1
5	60000	2	yes	no	25%	Yes	0
6	61000	2	no	no	26%	Yes	1
7	64500	2	no	no	29%	Yes	1
8	71000	2	no	no	35%	Yes	1
9	75500	1	yes	no	33%	Yes	0
10	33500	1	no	no	7%	No	0

Evaluate their classification model by the following steps:

a) Insert the predicted response of "Yes" or "No" in a new column
See the new column above. Additionally, another column has been added, marking those houses where the predicted response was incorrect (i.e. did not agree with the observed airco status) by 1 and those with correct prediction by 0.

b) Determine the error rate for all houses in the new data
There are 10 houses in the new data, and the model gave incorrect prediction for 4 of them. Therefore, the test error rate here is 4/10 = 40%.

c) Determine the error rate for those houses in the new data that do not have air conditioning
There are 8 houses in the new data that do not have air conditioning, and the model gave incorrect prediction for 4 of them. Therefore, the test error rate for such houses is 4/8 = 50%.

d) Determine the error rate for those houses in the new data that have air conditioning
There are 2 houses in the new data that have air conditioning, and the model correctly predicted the presence of air conditioning for both. Therefore, the test error rate for houses with air conditioning is 0/2 = 0% for this test set.

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;
    Yes, this is always correct. A very flexible model built on the training points can get very close to these points, so the MSE will be very low, whereas an inflexible model has a given shape or structure and cannot follow all wiggles and curves in the training set.

    b. Test MSE for model B ≤ Test MSE for model A.
    No, this is not always the case. If our predictions follow the training points too much, then they will include all wiggles and curves that are there purely by chance, and not because they are part of the true relationship. This means that such wiggles and curves may not be present in a new set of points, and we may make errors when predicting their presence.
    Generally, the case when a model follows the training points too closely and therefore makes errors for new points, that is called overfitting – there are many examples shown in Lecture 7a for such models.

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
```

As indicated in the text of the exercise, four linear regression models are considered in the LOOCV process. In Model 1, the model was built on data from years 2, 3 and 4, and the validation set is the year 1 data. Model 1 would give the following predicted completion time for the year 1 data:

$$\widehat{\text{Time}} = 100 - 0.923 \times 70 = 35.39.$$

The MSE for Model 1 is the squared difference between the prediction with model 1 and the observed completion time for year 1:

$$MSE_1 = \frac{(20-35.39)^2}{1} = 236.85.$$

Similarly, we can look at the MSE values for the other three models:

$$MSE_2 = \frac{(60-(107.69-1.08\times 30))^2}{1} = 233.78,$$

$$MSE_3 = \frac{(100-(78.57-0.714\times 10))^2}{1} = 816.24,$$

$$MSE_4 = \frac{(20-(107.14-1.29\times 90))^2}{1} = 838.68.$$

The LOOCV estimate of MSE is the average of these values:

$$MSE_{LOOCV} = \frac{MSE_1+MSE_2+MSE_3+MSE_4}{4} = 531.39.$$

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