Exercises from Course Slides

NRMSE calculation

Calculate the NRMSE and Evaluate

x_i	y_i
1	0.5
2	2.5
3	2.0
4	4.0
5	3.5
6	6.0
7	5.5

x_i	y_i
0.2	8.2
0.4	8.4
0.6	8.5
0.8	8.6
1.0	8.8
1.2	8.7

$$MSE = \frac{SSE}{n},\ RMSE = \sqrt{\frac{SSE}{n}},\ NRMSE = \frac{RMSE}{\bar{y}}$$

$$S_r = SSE = \sum_{i=1}^{n} |e_i^2| = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

R^2 Calculation

Calculate R^2 and Evaluate

x_i	y_i
1	0.5
2	2.5
3	2.0
4	4.0
5	3.5
6	6.0
7	5.5

x_i	y_i
0.2	8.2
0.4	8.4
0.6	8.5
0.8	8.6
1.0	8.8
1.2	8.7

$$R^2 = 1 - SSE/TSS,\ SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2,\ TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

KNN

Apply KNN to the approximation for the value x=4.4 and for the values k=3, 4 and 5

x_i	y_i
1	0.5
2	2.5
3	2.0
4	4.0
5	3.5
6	6.0
7	5.5

K-Fold Cross Validation

Compare KNN (K=2) and Linear Regression using K-fold Cross Validation (K=2)

x_i	y_i
1	0.5
2	2.5
3	2.0
4	4.0
5	3.5
6	6.0
7	5.5

$$MSE = \frac{SSE}{n},\ RMSE = \sqrt{\frac{SSE}{n}},\ NRMSE = \frac{RMSE}{\bar{y}}$$

$$CV = \frac{\sum_{i=1}^{k} NRMSE_i}{k}$$

$$a_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$a_0 = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

Exercises from Previous Exams

Midterm 2017-2018

Consider the following data set E composed of 6 points defined by their coordinates (x, y). Compare the linear regression method and the k-nearest neighbor method using the cross-validation method.
Divide the set E into two subsets: E_1= {p_1, p_3, p_5}, E_2= {p_2, p_4, p_6}

	x	y
p1	1	0
p2	2	3
p3	3	1
p4	4	2
p5	5	4
p6	6	5

Midterm 2018-2019

One company has started a process of collecting data from a large number of programmers. For the sake of simplicity, we consider 4 observations (programmers) only here.

Observation	Experience	Score	Salary
1	3	2	3
2	7	6	5
3	1	4	2.5
4	5	8	4

The purpose of this exercise is to understand the relationship between the salary of a programmer and the level of experience (number of years of work) and/or his/her score (evaluation of the company).

1. Estimate the salary of a programmer with 2 years of experience using the K-nearest Neighbors Method.
2. Estimate the salary of a programmer with 2 years of experience and 3 score points using the K-Nearest Neighbors method. The Euclidean distance between two observations must be based on the two variables (predictors): Experience and Score.
3. To validate the model of K-Nearest Neighbors on this dataset, the cross-validation method is used. Divide the set into 2 subsets $\{O_1, O_2\}$ and $\{O_3, O_4\}$. Consider both Experience and Score predictors. Calculate the corresponding CV value.

Additional Exercise

A research study has been conducted to determine the loss of activity of a drug. The table below shows the results of the experiment.

Time (in years)	1	2	3	4	5
Activity (%)	96	84	70	58	52

Construct the linear regression model of activity on time.

1. According to the linear model, when will the activity be 80%?
2. When will the drug have lost all activity?