Machine Learning and Data Mining

Assignment 1

Raphael Menges

raphaelmenges@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: 16.11.2020, CEST 23:59

The exercises in this assignment are of theoretical nature and may not be solved by execution of high-level Python commands but through manual step-by-step calculations which must be included in submissions. For this assignment it is also allowed to upload a single .pdf file generated from LATEX code or scanned and compressed (!) handwritten solutions.

1 Statistics 18 points

1.1 6 points

1.2 6 points

2 Error Calculation 12 points

You are given many computed outputs y_i and desired outcomes $\hat{y}_i$.Provide the following error measures in regards to y_i and $\hat{y}_i$ by writing down their formula and a short description about their characteristics, i.e., the behavior in regard to the difference between computed and desired outcome.

2.1 Sum of Square Error (SSE)

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

y_i : computed output
$\hat{y}_i$: desired output

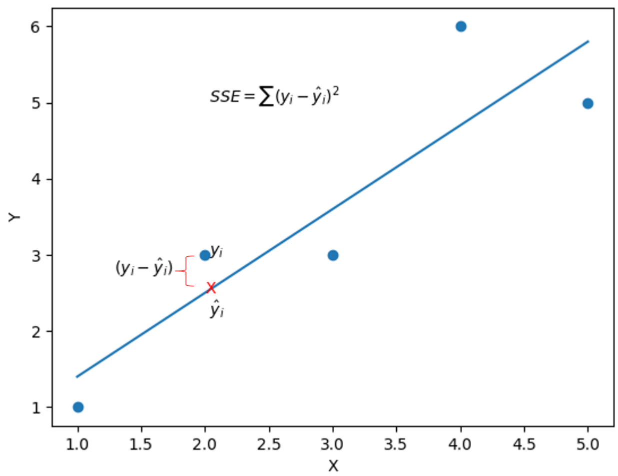

Figure 1: sum of square error.

- The error will be the summation of differences squared between y_i (computed value) and $\hat{y}_i$ (actual value).We compute how far away is our predicted value from actual value.
- Positive terms cancelling out negative terms is avoided by squaring the difference.
- We can differentiate SSE loss function at all points which serves a greater advantage for mathematical optimisations(we obtain optimum point by differentiating the function and equating it to zero)

2.2 Mean Square Error(MSE)

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

- MSE is the quadratic loss function, which squares and subsequently averages the various errors
- In MSE, squaring the error gives more weight to larger errors than smaller ones and thereby penalising them.

2.3 Root Mean Square Error (MSE)

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

- RMSE is the standard deviation of residuals from the best fit line (regression). It gives us an overview of how concentrated these residual points are around the line of best fit.
- The triangle inequality is satisfied by RMSE, which is required for a distance function metric
- RMSE penalises large errors.

2.4 Mean Absolute Error (MAE)

$$MAE = \sum_{i=1}^{n} \frac{1}{n}|y_i - \hat{y}_i|$$

- The models performance is not accurately reflected by MAE metric when dealing with large error values, thus it does not necessarily penalise large errors.
- When the dataset is small and we have large number of outliers, a loss function which is less sensitive to outliers like MAE could be chosen.

3 Research tasks 20 points

.

3.1 Collision

3.2 IPv6

4 Routing Table 20 points