COMP40370 – Practical 6

Polynomial Regression

Prof. Tahar Kechadi

Academic year 2021-2022

To be Graded: NO

Assignment Files

- ./practical06.pdf — assignment questions (this file)
- ./specs/markB_question.csv — data file for Question 1
- ./test_practical6.py — test file to validate your python program.

Question 1: Polynomial Regression

The file ./specs/markB_question.csv contains data about two midterm and final exam grades (out of 100 for each) for a group of students.

1. Generate a simple linear model (using sklearn LinearRegression) that predicts the final term grade based on the MCQ1 grade alone. Use this model to predict the final grade of all the entries in the data file, then save all the predictions in a new column called final_linear. What are the values of the model parameters (coefficient and intercept) and the accuracy of the model (R^2 and RMSE)?

2. Use the same multiple linear regression model with feature transformations (without using sklearn PolynomialFeatures) to develop a degree 2 polynomial model that predicts the final exam grade based on the MCQ1 grade alone. What are the values of the model parameters (coefficient and intercept)? Use this model to predict the final grade of all the entries in the data file, then save all the predictions in a new column called final_quadratic.

3. Generate a group of polynomial models (using sklearn PolynomialFeatures) that predict the final exam grade based on the MCQ1 grade alone. The models should have degrees of 2, 3, 4, 8, and 10. For each model, predict the final grade of all the entries in the data file, then save all the predictions in a new column called final_polyX, where X stands for

the degree of the polynomial model (so, the columns will be called final_poly2, final_poly3 and so on).

4. Find the model accuracies (R^2 and RMSE) for each model and discuss its trend with the increasing degree of the polynomial. Which models are underfitted and which are overfitted? How do you select one as a more generalised model?

5. Save all the predictions, alongside the original data, into a csv file called ./output/question_mcq1.csv. Save the graph in the file ./output/question_mcq1.pdf.

6. Repeat questions 1 and 2 that predicts the final exam grade based on both MCQ1 and MCQ2 grades. What are the values of the model parameters (coefficients and intercept) and model accuracies? Discuss whether MCQ1 and MCQ2 should be normalised or not. Include the numbers in your report, and save the predictions you obtained, alongside the original data, into a csv file called ./output/question_full.csv

Expected output and submission data

Your submission should be a single archive file (zip, tar, tgz, ...) containing one folder called output and the following files and directories:

- ./run.py main Python script
- ./report.pdf your PDF report (2 pages maximum)
- ./output/question_mcq1.csv Output file of Question 1.1 to 1.5
- ./output/question_full.csv Output file of Question 1.6
- ./output/question_mcq1.pdf plot of the models for Q.1.1 to Q1.5
- ./specs/ the original specs folder included in the assignment archive, containing the input data

The final deadline for the submission is **Thursday, 04th of November**, 2021, at **11:55**. You can submit your solution on Brightspace.

Programming requirements and tools

The assignment should be solved in Python, version 3.8 or above (3.9 is recommended). You shall use the following packages for this assignment:

- pandas 1.3+

- matplotlib 3.4+
- sklearn 0.24+

In particular, the following user guides are available for the required algorithms of the assignment:

- Linear Regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear model.LinearRegression.html
- Polynomial Features: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html