0513 report
Kurosu Ryo

1. Collect or identify a data set.

I use red Wine Quality Data Set.
https://archive.ics.uci.edu/ml/datasets/Wine+Quality

2. Implement and apply (multivariate) regression.

I used 80% of the data for training and the remaining 20% for testing.

3. Write a report that clearly discusses the following (preferably as a Jupyter notebook).

- Explanation of the data you used. Clearly state the objective (e.g., predict B from A).

The purpose of my experiment is to predict
(A) the quality of wine from alcohol content, acidity, pH, etc..
(B) the alcohol content from volatile acidity and density

- Explanation of regression analyses you implemented and applied with math.

I analyzed the data using regression analysis, with (1) linear functions, (2) polynomial functions as basis functions (n: 1~4).

- Evaluation results including quantitative analysis with adequate visualization(i.e., graphs).

Here are my scores.

(A)
n=1 (Linear)
train_score: 0.3654519616206866
test_score: 0.3283887639580232
n=2
train_score: 0.4511434209497146
test_score: 0.2884223239785493
n=3
train_score: 0.6662629684201462
test_score: -0.1955593471833006
n=4
train_score: 0.9994146232396502
test_score: -55978.55543358805
n=5
train_score: 0.999975790614722
test_score: -3048091.478041317

(B)
n=1 (Linear)
train_score: 0.2731849993614881
test_score: 0.31642617294963526
n=2
train_score: 0.37893351206607173
test_score: 0.4797700863260561
n=3
train_score: 0.39656240733865744
test_score: 0.48522395946041186
n=4
train_score: 0.4243948413207551
test_score: 0.5167557850406443
n=5
train_score: 0.42935357368023563
test_score: 0.5188074974965116

The result of A was not so good, but B was good.

Linear regression scored best for (A), which examined wine quality from all feature vectors, while polynomial was better for (B), which examined alcohol concentration.
I also examined cases where n was greater than 6, but they became worse than when n=5.

4. Submit, via PandA, the data, code, and report. If your data is large, please include a link to it in your report instead of uploading it.

Please see attached in a separate file.

- Blue dot represents train data
- Red dot represents test data
- “file name” = “img_{(A) quality or (B) alcohol}_n={1,2,..,5}”