HW3: Generalized Linear Models

Part 1 (Grade 6): Model Implementation

Implement multinomial logistic regression and ordinal logistic regression as classes (`MultinomialLogReg`, `OrdinalLogReg`) that provide a method `build(X, y)`, which returns the fitted model as an object, whose `predict(X)` method returns the predicted probabilities of given input samples.

Multinomial logistic regression and ordinal logistic regression should be implemented as described in the lecture notes. For both, implement the (log-)likelihood of the model and an algorithm that fits the model using maximum likelihood estimation and can make predictions for new observations. For optimization, you may use any third-party optimization library that allows for box-constraints (for example, `fmin_l_bfgs_b` from `scipy`). Optimization with numerical gradients is sufficient.

Test your implementation with unit tests. Combine all tests in a class named `MyTests`.

Part 2 (Grades 7–8): Application

2.1 Application of the Multinomial Regression

You are provided with `dataset.csv`, which contains data from over 5024 basketball shots in actual games. A detailed description is available in the Methods section of this paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128885.

Your goal is to use multinomial regression to provide insights into the relationship between *shot type* (our target variable) and the other variables. Hypothesize a reasonable practical explanation for any relationships you find.

Note: Regression coefficients, like any estimate, contain uncertainty. And interpreting any estimate without taking into account the associated uncertainty, is a bad idea.

2.2 Application of the Ordinal Regression

Come up with a data-generating process (DGP) and a dataset generated by it, where ordinal logistic regression outperforms multinomial logistic regression. You can implement the process as a generator function named `multinomial_bad_ordinal_good`, which produces IID observations from your DGP but it is enough to explain how the DGP works.

Part 3 (Grades 9-10): GLM Diagnostics

This part references the book Dunn, P. K., & Smyth, G. K. (2018). Generalized linear models with examples in R (Vol. 53, p. 16). New York: Springer.
Study Chapter 8 up to and including 8.8 (Outliers and Influential Observations).
Implement linear regression and the following diagnostic plots (refer to Fig. 8.10):

- Normal Q-Q plot,
- residuals vs fitted values (pick appropriate residuals),
- Cook's distance.

You may use linear algebra primitives, but the rest must be your own code. Demonstrate your implementation on predicting distance from angle in the basketball dataset (simple linear regression, no transformations).

General

- In addition to these instructions, the homework includes a `dataset.csv` file and a chapter from the referenced book.
- Submit a pdf report (all parts combined into a single pdf; no more than 1 page per part!)
- Submit your code in a single Python 3.8-compatible file.
  - The code should only execute functionality under `if __name__ == "__main__"`.
  - The code must conform to the unit tests provided in `test_lr.py`.
- Feel free to use any tools, including LLMs and collaboration with others, but keep in mind that our goal is to understand what we are doing and not merely to do. Your work will be graded based on your understanding of your code, report, and the subject matter in general.