Classification Tasks Using Logistic Regression and KNN

Dr. Jishan Ahmed

Spring 2024

Figure 1: Data Scientist Sherlock Holmes

Due Dates

Completed Assignment: Due April 1, 2024.

Instructions:

Part 1: Classical Statistics Approach

Annual financial data are collected for bankrupt firms approximately two years prior to their bankruptcy and financially sound firms at about the same time. The data on four variables, $X_1 = \frac{\text{CF}}{\text{TD}}$ (cash flow)/(total debt), $X_2 = \frac{\text{NI}}{\text{TA}}$ (net income)/(total assets), $X_3 = \frac{\text{CA}}{\text{CL}}$ (current assets)/(current liabilities), and $X_4 = \frac{\text{CA}}{\text{NS}}$ (current assets)/(net sales) are given in `bankruptcy.csv`. This is a binary classification problem. Take bankrupt firm as "+" response (indicated as 0 in the data) and non-bankrupt firm as "-" response (indicated as 1 in the data). Use all the data as training data.

a) Using the `statsmodels` Python API, fit logistic regression to the training data using all the predictors. Look at the summary of this fit and decide which predictors might be dropped. Verify your decision by comparing the Akaike information criterion (AIC) scores of the reduced model and full model.

b) Estimate $\exp(\hat{\beta})$ with 95% confidence interval and interpret the values.

c) Compute the error rate, F1 score, and G-mean score for your final model.

d) Find the equation of the decision boundary.

e) Produce a scatter plot of the predictors in your final model (there should be two such predictors) and superimpose the decision boundary. Color the points according to the value of the response. Comment on what you observe.

Part 2: Machine Learning Approach

You can download the dataset and see its description at Default of Credit Card Clients Dataset.

a) Visualize the univariate distribution of each continuous feature, and the distribution of the target.

b) Split data into training and test set. Use `ColumnTransformer` and `pipeline` to encode categorical variables. Evaluate logistic regression, and k-nearest neighbors (KNN) classifiers using 5-fold cross-validation. How different are the results? How does scaling the continuous features with `StandardScaler` influence the results?

c) Tune the hyperparameters using `GridSearchCV`. Do the results improve? Visualize the performance as function of the parameters for all two models.

d) Change the cross-validation strategy from `stratified k-fold` to `kfold` with shuffling. Do the parameters that are found change? Do they change if you change the random seed of the shuffling? Or if you change the random state of the split into training and test data?

Part 3: Model Comparison

Finally, compare the classical statistical approach and the machine learning approach by:

a) Calculating error rate, F1 score, and G-mean score for all models.

b) Discussing the strengths and limitations of each approach in the context of this analysis.

Discussion

Answer the following reflection questions:

- What do I believe I did well on this assignment?
- What was the most challenging part of this assignment?
- What would have made this assignment a better experience?
- What do I need help with?

Group Dynamics

Ensure equitable distribution of work among group members.