Introduction to Statistical Machine Learning: Take-Home Assignment

Fabian Krüger

Start date: June 11, 2018
Submission date: June 21, 2018

Data set

The assignment is based on a data set of credit records provided by Lending Club, a peer-to-peer lending platform (`https://www.lendingclub.com`). The following table describes the variables we consider:[1]

name	type	description
`default`	binary	Did credit holder default?
`loan_amount`	numeric	The listed amount of the loan applied for by the borrower.
`term`	numeric	The number of payments on the loan (values in months, either 36 or 60).
`int_rate`	numeric	Interest rate on the loan.
`installment`	numeric	The monthly payment owed by the borrower if the loan originates.
`grade`	categorical	Loan grade assigned by lending platform (A to G).
`addr_state`	categorical	State provided by the borrower in the loan application
`emp_length`	categorical	Employment length (11 categories).
`home_ownership`	categorical	Borrower's home ownership status.
`annual_inc`	numeric	Annual income (self-reported by the borrower during registration).
`verif_status`	categorical	Indicates if income was verified by lending platform, not verified, or if the income source was verified.
`purpose`	categorical	Category provided by the borrower for the loan request.
`dti`	numeric	Ratio of monthly debt payment to monthly income.
`open_acc`	numeric	Number of open credit lines in the borrower's credit file.
`total_acc`	numeric	Total number of credit lines currently in the borrower's credit file.
`obs_id`	numeric	Observation ID (training data have ID 1 to 50000, test data have ID 50001 to 100000)

I randomly split the data into two parts: A training data set and a test data set, each containing 50000 observations. In the assignment, you'll be asked to develop a statistical prediction

[1] The table is based on the codebook provided by the credit platform, available at `https://resources.lendingclub.com/LCDataDictionary.xlsx`.

model for the `default` variable. The latter variable is contained in the training data set but not the test data set. Both data sets are stored in the file `assignment.Rdata`. Running the command `load(assignment.Rdata)` in R adds both data sets to your workspace.

Problem 1

Use the training data to explore how `default` relates to other variables in the data set, and describe your findings in a short report (2–3 pages, including figures and/or tables to support your statements). This short report need not be exhaustive; instead, please focus on a few key points that you find interesting and relevant.

Problem 2

Formulate a statistical model for predicting defaults. Describe and motivate your specification (1–2 pages, including figures and/or tables to support your statements), and estimate it using the training data set.

Problem 3

Submit predicted probabilities of the `default` variable, for all observations in the **test sample**. *Importantly, make sure that your submission satisfies the format described under 'Rules for submission and grading' below.* The probabilities will be evaluated using the log-likelihood on the test-sample data, given by

$$LL = \sum_{i=n_1}^{n_2} \left\{ \texttt{default}_i \ \log(p_i) + (1 - \texttt{default}_i) \ \log(1 - p_i) \right\}, \tag{1}$$

where $i = n_1, \ldots, n_2$ denotes the test sample observations,

$$\texttt{default}_i = \begin{cases} 1 & \text{default for observation } i \\ 0 & \text{otherwise} \end{cases},$$

and $p_i \in [0, 1]$ is the predicted probability for observation i.

Rules for submission and grading

- The deadline for submission is on **Thursday, June 21, at 4.15 pm**.
- Each submission must contain **exactly three files**:
  - A `.pdf` file containing your solutions for Problems 1 and 2. The file must be named `assignment_[StudentID].pdf`, where `[StudentID]` is your student ID.
  - A `.csv` file named `predictions_[StudentID].csv`, containing your predicted probabilities for Problem 3. The file must contain 50000 rows (one for each observation in the test sample), and two columns. Column 1 must contain the observation ID, and column 2 must contain the associated predicted probability (a number between zero and one).

For example, the first row of your submitted file could read

50001, 0.8097263

and the last row could read

100000, 0.4141398

(I produced the probabilities using a random number generator, so please don't interpret these numbers!)

- An `.R` file named `code_[StudentID].R`, containing your R code for all problems.

- Please **submit all files via Moodle**.

- **Grading:** Every student who submits three files that

  i) satisfy the formal guidelines mentioned above, and

  ii) contain at least satisfactory solutions to Problems 1–3

  will receive a bonus of $0,3$ on their grade from the final exam. As noted earlier, a grade of $4,0$ or better in the final exam is necessary to pass the course.

- **Log-Likelihood evaluation:** The student whose submission for Problem 3 attains the highest log-likelihood value (see Equation 1) will receive a copy of the book *The Signal and the Noise: The Art and Science of Prediction* by Nate Silver.

- **Policy regarding collaboration:** You are welcome to discuss your work with your classmates. Nevertheless, every student must formulate and hand in their work independently.