IN-STK-5000, Credit Project

Christos Dimitrakakis `chridim@ifi.no`

September 4, 2018

1 Introduction

Before you start, make sure you have

- Joined one of the project groups in Piazza.
- Forked the code in `https://github.com/olethrosdc/ml-society-science`
- All questions should go through the QA platform in Piazza.

2 Project: Credit risk for mortgages

Consider a bank that must design a decision rule for giving loans to individuals. In this particular case, some of each individual's characteristics are partially known to the bank. We can assume that the insurer has a linear utility for money and wishes to maximise expected utility. Assume that the t-th individual is associated with relevant information x_t, sensitive information z_t and a potential outcome y_t, which is whether or not they will default on their mortgage. For each individual t, the decision rule chooses $a \in \mathcal{A}$ with probability $\pi(a_t = a \mid x_t)$.

As an example, take a look at the historical data in `data/credit/german.data-mumeric`, described in `data/credit/german.doc`. Here there are some attributes related to financial situation, as well as some attributes related to personal information such as gender and marital status.

A skeleton for the project is available at `https://github.com/olethrosdc/ml-society-science/tree/master/src/project-1`. Start with `random_banker.py` as a template, and create a new module `name_banker.py`. You can test your implementation with the `TestLending.py` program.

For ensuring progress, the project is split into two parts:

2.1 Deadline 1: September 14

The first part of the project focuses on a baseline implementation of a banker module.

1. Design a policy for giving or denying credit to individuals, given their probability for being credit-worthy. Assuming that if an individual is credit-worthy, you will obtain a return on investement of $r = 0.5\%$ per month. Take into account the length of the loan to calculate the utility through `NameBanker.expected_utility()`. Assume that the loan is either fully repaid at the end of the lending period n, or not at all to make things simple. If an individual is not credit-worthy you will lose your investment of m credits, otherwise you will gain $m[(1-r)^n - 1]$. Ignore macroenomic aspects, such as inflation. In this section, simply assume you have a model for predicting creditworthiness as input to your policy, which you can access `NameBanker.get_proba()`.

2. Implement `NameBanker.fit()` to fit a model for calculating the probability of creditworthiness from the german data. Then implement `NameBanker.predict_proba()` to predict the probability of the loan being returned for new data. What are the implicit assumptions about the labelling process in the original data, i.e. what do the labels represent?

3. Combine the model with the first policy to obtain a policy for giving credit, given only the information about the individual and previous data seen. In other words, implement `Namebanker.get_best_action()`.

4. Finally, using `TestLending.py` as a baseline, create a jupyter notebook where you document your model development. Then compare your model against `RandomBanker`.

2.2 Deadline 2: September 28

The second part of the project focuses on issues of reproducibility, reliability, privacy and fairness. That is, how desirable would it be to use this model in practice? Here are some sample questions that you can explore, but you should be free to think about other questions.

1. Is it possible to ensure that your policy maximises revenue? How can you take into account the uncertainty due to the limited and/or biased data? What if you have to decide for credit for thousands of individuals and your model is wrong? How should you take that type of risk into account?

2. Does the existence of this database raise any privacy concerns? If the database was secret (and only known by the bank), but the credit decisions were public, how would that affect privacy?

3. Choose one concept of fairness, e.g. balance of decisions with respect to gender. How do you ensure that your policy is fair? How can you measure it?

Submit a final report about your project, either as a standalone PDF or as a jupyter notebook.