Project 2: Credit Analytics

(First discussion: Oct 16; Last questions: Oct 30; Deadline: Nov 6)

Responsible: Jean-Loup Dupret

This project is about credit analytics for consumer loans. The goal is to estimate risk profiles of individuals applying for a loan. For simplicity, we work with artificially generated data and only consider three borrower characteristics: age, monthly income and employment status. In reality, the availability of good data is important, and typically, many more features are taken into account.

1. Let $m = 20000, n = 10000$ and simulate $m+n$ vectors $x^i = (x_1^i, x_2^i, x_3^i) \in \mathbb{R}^3$, $i = 1, \ldots, m+n$, with

   - x_1^i = age in $[18, 80]$ (from the uniform distribution)
   - x_2^i = monthly income in CHF 1000 in $[1, 15]$ (from the uniform distribution)
   - x_3^i = salaried/self-employed in $\{0, 1\}$, where 0=salaried and 1=self-employed (probability of being self-employed is 10%)

   such that x_1^i, x_2^i, x_3^i are independent.

   a) Compute the empirical means and standard deviations of x_1^i, x_2^i and x_3^i over $i = 1, \ldots, m$.

   b) Can you think of additional features (besides age, income, salaried/self-employed) that could be relevant in reality?

2. Let ξ^i, $i = 1, \ldots, m+n$ be independent random variables that are uniformly distributed on $(0, 1)$ and $\psi \colon \mathbb{R} \to (0, 1)$ the logistic (or sigmoid) function given by

$$\psi(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}.$$

   Consider two functions $p_1, p_2 \colon \mathbb{R}^3 \to (0, 1)$ of the form

$$p_1(x) = \psi\left(13.3 - 0.33x_1 + 3.5x_2 - 3x_3\right)$$
$$p_2(x) = \psi\left(5 - 10\left[1_{(-\infty,25)}(x_1) + 1_{(75,\infty)}(x_1)\right] + 1.1x_2 - x_3\right)$$

   and generate two artificial data sets (x^i, y_1^i) and (x^i, y_2^i), $i = 1, \ldots, m+n$, by setting

$$y_1^i = \begin{cases} 1 & \text{if } \xi^i \leq p_1(x^i), \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad y_2^i = \begin{cases} 1 & \text{if } \xi^i \leq p_2(x^i), \\ 0 & \text{otherwise.} \end{cases}$$

   (We use the convention that $y_s^i = 1$ is a good borrower whereas $y_s^i = 0$ is a delinquent borrower. That is, p_1 and p_2 are the conditional probabilities that loans will be paid back in the two data generating regimes.)

   For both data sets, $s = 1, 2$, do the following:

   a) Fit a *logistic regression model* $\hat{p}_s^{\log} \colon \mathbb{R}^3 \to \mathbb{R}$ on the *training data* (x^i, y_s^i), $i = 1, \ldots, m$. Calculate the cross-entropy loss of $\hat{p}_s^{\log}$ on the training and test data. You can use the function sklearn.linear_model.LogisticRegression for this.

b) For SVM classification, we denote by $\hat{\sigma}_j$ the empirical standard deviation of $(x^i_j)_{i=1}^m$ and work with the normalized data $\tilde{x}^i_j = x^i_j/\hat{\sigma}_j$ (for both training *and* evaluation).

(i) Fit a SVM $\hat{f}^{\text{svm}}_s \colon \mathbb{R}^3 \to \mathbb{R}$ of the form

$$\hat{f}^{\text{svm}}_s(x) = \langle w, \Phi(x)\rangle + b$$

with feature map Φ on the *training data* using the hinge loss, kernel $k(x,x') = \exp\big(-\frac{1}{10}\|x-x'\|_2^2\big)$ and regularization parameter $\lambda = \frac{5}{2m}$. You can use the function sklearn.svm.SVC for this (the given choice of λ corresponds to the parameter $C = 1/(2\lambda m) = 0.2$ in sklearn.svm.SVC).

(ii) On top of $\hat{f}^{\text{svm}}_s$, fit a *logistic function* $\hat{g}_s : \mathbb{R} \to \mathbb{R}$ of the form

$$\hat{g}_s(z) = \frac{1}{1+\exp(\alpha z + \beta)} \quad \text{for parameters } \alpha, \beta \in \mathbb{R}$$

so that $\hat{p}^{\text{svm}}_s := \hat{g}_s \circ \hat{f}^{\text{svm}}_s$ predicts conditional probabilities that loans are paid back; see Platt (1999)[1]. To this end, you may simply use the option "probability=True" in the sklearn.svm.SVC function.

(iii) Compute the cross-entropy loss of $\hat{p}^{\text{svm}}_s$, $s = 1, 2$, on both, the training and test data.

c) Generate FDR/TPR-curves and AUC from the test data for $\hat{p}^{\log}_s$ and $\hat{p}^{\text{svm}}_s$.

3. Let us now focus on the second dataset (x^i, y^i_2), $i = 1, \ldots, m+n$. The goal is to find "good investment opportunities" in the *test data set* based on the *features* x^i, $i = m+1, \ldots, m+n$. We here assume that loans are either completely repaid with interest or fully delinquent. In reality, a lender tries to recover parts of delinquent loans.

We compare three different lending strategies:

(i) We give out a loan to every person in the dataset in the amount of CHF 1000 charging an interest rate of 5.5%.

(ii) We only charge an interest rate of 1%, but we selectively choose the applicants who are awarded a loan (in the amount of CHF 1000) using the selection criterion

$$\hat{p}^{\log}_2(x^i) \geq 95\%.$$

(iii) We only charge an interest rate of 1% but we selectively choose the applictants who are awarded a loan (in the amount of CHF 1000) using the selection criterion

$$\hat{p}^{\text{svm}}_2(x^i) \geq 95\%.$$

To estimate the performance of the strategies (i)–(iii) above, we simulate different market scenarios according to the conditional probabilities $p_2(x^i)$, $i = m+1, \ldots, m+n$. Using independent Unif$(0,1)$-distributed random variables $\xi^{i,k}$, $i = 1, \ldots, n$, $k = 1, \ldots, 50000$, generate the $n \times 50000$-matrix $D \in \{0,1\}^{n\times 50000}$ given by

$$D_{i,k} = \begin{cases} 1 & \text{if } \xi^{i,k} \leq p_2(x^{m+i}) \\ 0 & \text{otherwise,} \end{cases}$$

where $D_{i,k} = 1$ means that in scenario k, the i-th loan is paid back with interest. So, the k-th column of D describes which loans are paid back in the k-th scenario.

Now, for each of the strategies (i)–(iii) above ...

a) plot a histogram of the profits & losses over the different market scenarios and estimate the expected profit & loss.

b) estimate the 95%-VaR of the profit & loss distribution (= negative of the 5%-quantile).

[1] `https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf`