Project 2: Credit Analytics

(First discussion: Oct 15; Last questions: Oct 29; Deadline: Nov 5)

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 $x_i = (x_{i1}, x_{i2}, x_{i3}) \in \mathbb{R}^3$, $i = 1, \dots, m + n$, with
   - x_{i1} = age in $[18, 80]$ (from the continuous uniform distribution)
   - x_{i2} = monthly income in CHF 1000 in $[1, 15]$ (from the continuous uniform distribution)
   - x_{i3} = salaried/self-employed in $\{0, 1\}$, where 0=salaried and 1=self-employed (probability of being self-employed is 10%)

   such that x_{i1}, x_{i2}, x_{i3} are independent.

   a) Compute the empirical means and standard deviations of x_{i1}, x_{i2} and x_{i3} over $i = 1, \dots, m$.

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

2. Let ξ_i, $i = 1, \dots, m + n$ 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

$$\begin{aligned} p_1(x_i) &= \psi\left(13.3 - 0.33 x_{i1} + 3.5 x_{i2} - 3 x_{i3}\right) \\ p_2(x_i) &= \psi\left(5 - 10\left[1_{(-\infty,25)}(x_{i1}) + 1_{(75,\infty)}(x_{i1})\right] + 1.1 x_{i2} - x_{i3}\right) \end{aligned}$$

   and generate two artificial data sets $(x_i, y_i^{(1)})$ and $(x_i, y_i^{(2)})$, $i = 1, \dots, m + n$, by setting

$$y_i^{(1)} = \begin{cases} 1 & \text{if } \xi_i \le p_1(x_i), \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad y_i^{(2)} = \begin{cases} 1 & \text{if } \xi_i \le p_2(x_i), \\ 0 & \text{otherwise.} \end{cases}$$

   (We use the convention that $y_i^{(s)} = 1$ is a good borrower whereas $y_i^{(s)} = 0$ 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* $\hat{p}_s^{\log}\colon \mathbb{R}^3 \to (0, 1)$ on the *training data* $(x_i, y_i^{(s)})$, $i = 1, \dots, m$. 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 $(x_{ij})_{i=1}^m$ and work with the normalized data $\tilde{x}_{ij} = x_{ij}/\hat{\sigma}_j$ (for both training *and* evaluation).

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

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

with feature map Φ on the *training data* using the hinge loss, kernel $k(x, x') = \exp\left(-\frac{1}{10}\|x - x'\|_2^2\right)$ 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 $\hat{f}_s^{\text{svm}}$, fit a *logistic function* $\hat{g}_s : \mathbb{R} \to (0, 1)$ 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 $\hat{p}_s^{\text{svm}} := \hat{g}_s \circ \hat{f}_s^{\text{svm}}$ 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 $\hat{p}_s^{\text{svm}}$, $s = 1, 2$, on both the normalized training and test data.

(iv) Would the results change if we used standardized data $\tilde{z}_{ij} = (x_{ij} - \hat{\mu}_j)/\hat{\sigma}_j$ instead of the normalized data $\tilde{x}_{ij} = x_{ij}/\hat{\sigma}_j$, with $\hat{\mu}_j$ the empirical mean of $(x_{ij})_{i=1}^m$. Explain why or why not.

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

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

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

(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

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

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

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

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

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).