Project 1: Credit Analytics

(First discussion: Mar 22; Deadline: May 17)

This project is about credit scoring for consumer loans. The goal is to estimate the default risk of individuals applying for a loan. For simplicity, we work with artificially generated data and only consider three borrower features: 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. **Dataset features generation.**

   Let $m = 20000$ be the number of samples in the training set and $n = 10000$ be the number of samples in the test set. 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 from the uniform distribution on $[18, 80]$,
   - x_2^i = monthly income (in CHF 1000) from the uniform distribution on $[1, 15]$,
   - x_3^i = salaried/self-employed in $\{0, 1\}$, where 0=salaried and 1=self-employed (probability of being self-employed is 10%),

   in such a way that x_1^i, x_2^i, x_3^i are independent.

2. **Dataset labels generation.**

   Let $\psi\colon \mathbb{R} \to (0,1)$ be the logistic function (also known as the sigmoid function) given by

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

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

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

   The functions p_1 and p_2 are the default probabilities of a borrower with characteristics $x = (x_1, x_2, x_3)$ in two different data generating regimes.

   We generate two artificial data sets (x^i, y_1^i) and (x^i, y_2^i), $i = 1, \ldots, m+n$ by sampling independently each label according to the following distribution:

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

3. **Model implementations.**

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

   a) Fit a *logistic regression*[1] (LR) model $\hat{p}_s^{\mathrm{LR}}\colon \mathbb{R}^3 \to \mathbb{R}$ on the *training data* (x^i, y_s^i), $i = 1, \ldots, m$.

      Compute the cross-entropy loss of $\hat{p}_s^{\mathrm{LR}}$ on the training and test data.

[1] You can use the function `sklearn.linear_model.LogisticRegression` for this.

b) Fit a *neural network*[2] (NN) model $\hat{p}_s^{\mathrm{NN}} \colon \mathbb{R}^3 \to \mathbb{R}$ on the *training data* (x^i, y_s^i), $i = 1, \ldots, m$. To make the network output the conditional default probabilities, choose the sigmoid function ψ as activation function for the last layer and train the network by minimizing the cross-entropy.

   You should experiment with different network architectures and hyperparameters, until you find a network that performs well. You can start with two hidden layers of 50 neurons each with ReLU activation function and train for 100 epochs with batch size 1024 and learning rate 0.01.

   Compute the cross-entropy loss of $\hat{p}_s^{\mathrm{NN}}$ on the training and test data.

c) For both data generating regimes, plot the models' ROC curves and compute their AUC scores on the test data.

4. **Comparison of lending strategies.**

   Let us now focus on the second dataset (x^i, y_2^i), $i = 1, \ldots, m+n$. The goal is to find good investment opportunities in the *test data set* based on the features x^i, $i = m+1, \ldots, m+n$.

   Here we assume that each borrower either repays the loan in full (with interest) or defaults with zero recovery. In practice, 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 100 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 100) using the selection criterion

   $$\hat{p}_2^{\mathrm{LR}}(x^i) \leq 5\%.$$

   (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 100) using the selection criterion

   $$\hat{p}_2^{\mathrm{NN}}(x^i) \leq 5\%.$$

   To estimate the performance of the strategies (i), (ii) and (iii) above, we simulate 1000 different market scenarios according to the conditional probabilities $p_2(x^i)$, $i = m+1, \ldots, m+n$.

   Generate a matrix D of size $n \times 1000$, by sampling its entries independently according to the following distribution:

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

   where $D_{i,k} = 1$ means that in scenario k the i-th borrower defaults, while $D_{i,k} = 0$ means that in scenario k the i-th borrower pays back the loan (with interest).

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

   a) plot a histogram of the profits and losses (P&L) over the different market scenarios and estimate the expected P&L,

   b) estimate the 95%-VaR of the loss distribution (i.e, the distribution of $-$P&L).

[2] You can implement it using `Keras`.