New Technologies in Banking and Finance
Patrick Cheridito and Stephan Eckstein

ETH Zurich
September 22, 2023

Machine Learning Coding Exercises

1. Choose numbers $J, K \geq 5000$ and simulate $J+K$ vectors $x^j = (x_1^j, x_2^j, x_3^j) \in \mathbb{R}^3$ with

   - x_1^j = age in $[18, 80]$
   - x_2^j = monthly income in CHF 1000 in $[1, 15]$
   - x_3^j = salaried/self-employed in $\{0, 1\}$

   Compute the empirical means of x_1^j, x_2^j and x_3^j over $j = 1, \dots, J$.

   Give two additional features which you believe are relevant for credit risk assessment in reality. Explain your answer.

2. Let ξ^j, $j = 1, \dots, J+K$ be independent random variables that are uniformly distributed on $(0, 1)$ and $\sigma \colon \mathbb{R} \to (0, 1)$ the logistic (or sigmoid) function given by

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

   Consider a function $p \colon \mathbb{R}^3 \to (0, 1)$ of the form

$$p(x) = \sigma\left(a_0 + a_1|x_1 - 50| + a_2(x_2)^{1.25} + a_3 x_3\right)$$

   and generate an artificial data set (x^j, y_1^j), $j = 1, \dots, J+K$ by setting

$$y^j = \begin{cases} 1 & \text{if } \xi^j \leq p(x^j) \\ 0 & \text{otherwise.} \end{cases}$$

   a) Choose $a_0, a_1, a_2, a_3 \in \mathbb{R}$ with $0.25 < |a_i| < 5$ for $i = 1, 2, 3$ such that approximately 5% of all y^j, $j = 1, \dots, J$, are 1. (For the case $a_1 = a_2 = a_3 = 0$, what would be the "best possible" ROC curve?)

   b) "Learn" $\hat{p}_1 \colon \mathbb{R}^3 \to \mathbb{R}$ on the *training data* (x^j, y^j), $j = 1, \dots, J$, with logistic regression. Calculate the total deviance of the regression fit for both the training and test data.

   c) Which fraction of $\left\{y^j : \hat{p}_1(x^j) \in [3\%, 4\%],\ j = J+1, \dots, J+K\right\}$ is 1?

   d) "Learn" $\hat{p}_2 \colon \mathbb{R}^3 \to \mathbb{R}$ from the *training data* (x^j, y^j), $j = 1, \dots, J$, with a neural network. Calculate the total deviance of the regression fit for both the training and test data.

   e) Which fraction of $\left\{y^j : \hat{p}_2(x^j) \in [3\%, 4\%],\ j = J+1, \dots, J+K\right\}$ is 1?

d) Plot the ROC curve and calculate the AUC. Does the AUC change if the labeling convention for y is switched (i.e., we use the dataset $(x^j, \tilde{y}^j)$ with $\tilde{y}^j = 1 - y^j$, $j = 1, \ldots, J+K$)? Explain your answer.

3. Find "good investment opportunities" in the *test data set* based on the *features* x^j, $j = J+1, \ldots, J+K$, to form a portfolio of loans, all in the amount of CHF 1000 with interest rate 3.5%.

   a) Estimate the expected P&L.

   b) Estimate the 95%-VaR of the P&L (= negative of the 5%-quantile)