CASE STUDY: Classification in Loan Application Process (spline_sum, logexp_sum, logistic, pr_pen, bpoe)

Background

Peer-to-peer lending (P2PL) has gained increasing attention in recent years. This case study is trying to reproduce the loan application process of Lending Club, which is the world's largest online company. We used the open data from Lending Club, see https://www.lendingclub.com/info/download-data.action. Problem 0 is the standard logistic regression (used as a benchmark). Features transformation is done using cubic splines (see, Problem 1). Problem 2 is the logistic regression with transformed features. Problem 3 maximizes buffered AUC (bAUC) by minimizing buffered probability of exceedance (bPOE). Problem 4 maximizes AUC by minimizing probability of exceedance (PSG probability function *pr_pen*). We obtained AUC values close to 1 with only 3 features. Results of this case study are described in Wei, Gotoh, and Uryasev paper [2].

Raferences

[1] Norton, M. and S. Uryasev. Maximization of AUC and Buffered AUC in Classification. Research Report 2014-2, ISE Dept., University of Florida, October 2014.

[2] Wei, X., Gotoh, J. and S. Uryasev. Peer-to-Peer Lending: Classification in Loan Application Process. Research Report 2015-5, ISE Dept., University of Florida, September 2015.

Notations (for spline transformation of features)

J=number of observations (scenarios) of independent variable (feature),j= index of observation, j=1,...,J;

x_j = observation of feature, j=1,...,J. Observationsx_j are ordered, i.e., if $j_1 < j_2$ then $x_{j_1} \leq x_{j_2}$;

y_j = observation of of dependent variable corresponding to x_j, j=1,...,J;

$\vec{x} = (x_1, x_2, \dots, x_J)$ = vector of observations of feature;

$\vec{y} = (y_1, y_2, \dots, y_J)$ = vector of dependent variables;

D = degree of spline, $D \geq 0$, integer;

K = number of polynomial pieces in the spline, $K > 0$(integer);

S = smoothing degree of a spline, $0 \leq S \leq D$(integer);

$I = K \cdot (D + 1)$=number of coefficients of polynomial pieces in the spline;

a_{dk}= decision variable = coefficient for degree d in polynomial piece k, $d = 0, \dots, D$, $k = 1, \dots K$;

$\vec{a} = (a_{01}, a_{11}, \dots, a_{D1},\ a_{02}, a_{12}, \dots, a_{D2}, \dots, a_{0K}, a_{1K}, \dots, a_{DK})$ = vector of decision variables = vector of coefficients a_{dk};

$X = \{X_0, X_1, \dots, X_K\}$ = set of points (knots) partitioning segment $[x_1, x_J]$ in sub-segments, $k = 1, \dots K$; every sub-segment $[X_{k-1}, X_k]$ contains at least one observationx_j ($X_0 = x_1, X_K = x_J$);

J_k= sub-set of indexes j=1,...,J corresponding to the sub-segment $[X_{k-1}, X_k]$, $J_k = \{j | x_j \in [X_{k-1}, X_k]\}$;

$L_j(\vec{a}) = y_j - G_j^0(\vec{a}) = y_j - \sum_{d=0}^{D} a_{dk} \cdot x_j^d$ = Loss Functions at corresponding to j, $j \in J_k$, $k = 1, \dots, K$;

$G_j^0(\vec{a}) = \sum_{d=0}^{D} a_{dk} \cdot x_j^d$ = Gain Function with zero scenario benchmark y_j for$j$, $j \in J_k$, $k = 1, \dots, K$;

$\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}) = \{L_1(\vec{a}), L_2(\vec{a}), , \dots, L_J(\vec{a})\}$ = PSG function **Spline_sum** generating a set of loss scenarios $L_j(\vec{a})$, j=1,..., J, using observations;

$\mathbf{logexp\_sum}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$ =$\mathbf{logexp\_sum}(\vec{y}, G_1^0(\vec{a}), G_2^0(\vec{a}), \dots, G_J^0(\vec{a})) =$

$\frac{1}{J}\sum_{j=1}^{J}\left(y_j G_j^0(\vec{a}) - \ln\left(1 + \exp\left(G_j^0(\vec{a})\right)\right)\right)$= PSGfunctionLogarithms Exponents Sum (logistic regression log-likelihood function) applied to Spline_sumfunction. In this case, allvaluesy_j should be 0 or 1.

$\mathbf{logistic}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$= vector with components:

$\exp\left(G_j^0(\vec{a})\right) / \left(1 + \exp\left(G_j^0(\vec{a})\right)\right)$,$j$=1,..., J;

$u_j = \exp\left(G_j^0(\vec{a})\right) / \left(1 + \exp\left(G_j^0(\vec{a})\right)\right)$ = transformed featureat observationx_j,j=1,..., J;

$\vec{u} = (u_1, u_2, \dots, u_J)$ = vector of transformed observations obtained by spline transformation of feature x.

Optimization Problem 1: Spline Transformation of Feature

maximizing Logarithms Exponents Sum for building spline

$$\max_{\vec{a}} \mathbf{logexp\_sum}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

calculation of Logistic to get transformed feature $\vec{\boldsymbol{u}}$

calculate

$$\vec{\boldsymbol{u}} = \mathbf{logistic}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a})) \tag{1}$$

Notations (forclassification problems)

I = number of features;

$\vec{u} = (u^1, u^2, \dots, u^I)$ = vector of transformed features;

y = dependent variable;

J= number of observations;

y_j = j-th observation of dependent variable y, j=1,...,J;

$\vec{u}^i = (u_1^i, u_2^i, \dots, u_J^i)$= vector of transformed$i$-th feature calculated according to (1),i=1,..., I;

$\vec{y} = (y_1, y_{2,\dots,} y_J)$ = vector of all observations of dependent variable$y$, $y_j \in \{0,1\}$,j=1,...,J;

$\vec{\theta}_u = \begin{pmatrix} 1, u_1^1, \dots, u_1^I, y_1 \\ 1, u_2^1, \dots, u_2^I, y_2 \\ \dots\dots\dots\dots\dots\dots.. \\ 1, u_J^1, \dots, u_J^I, y_J \end{pmatrix}$ = extended design matrix(matrix of scenarios) for transformed features;the last column is the scenario benchmark;

$\vec{\omega} = (\omega_0, \omega_1, \dots, \omega_I)$ = vector of decision variables (linear classifier);

$G_j^0(\vec{\omega}, \vec{\theta}_u) = \omega_0 + \sum_{i=1}^{I} \omega_i u_j^i =$ Gain function for scenario *j*,with zero scenario benchmark, corresponding to transformed feature,*j*=1,...,*J*;

logexp sum$(\vec{\omega}, \vec{\theta}_u) = \frac{1}{J}\sum_{j=1}^{J}\left(y_j G_j(\vec{\omega}, \vec{\theta}_u) - \ln\left(1 + \exp\left(G_j(\vec{\omega}, \vec{\theta}_u)\right)\right)\right)$= PSG Logarithms Exponents Sum Function (logistic regression log-likelihood function) fortransformed features;

$J_1 = \{j: y_j = 1\}$= the set of scenarios indices, which correspond to the observation value $y_j = 1$;

$J_0 = \{j: y_j = 0\}$= the set of scenarios indices, which correspond to the observation value $y_j = 0$;

$L_j(\vec{\omega}, \vec{\theta}_u^1) = -\omega_0 - \sum_{i=1}^{I} \omega_i u_j^i$ = Loss functionfor scenario *j*, $j \in J_1$;

$L_j(\vec{\omega}, \vec{\theta}_u^2) = -\omega_0 - \sum_{i=1}^{I} \omega_i u_j^i$ = Loss functionfor scenario *j*,$j \in J_0$;

$L(\vec{\omega}, \vec{\theta}_u^1)$ = random Loss function with scenarios $\{L_j(\vec{\omega}, \vec{\theta}_u^1), j \in J_1\}$;

$L(\vec{\omega}, \vec{\theta}_u^2)$ = random Loss function with scenarios $\{L_j(\vec{\omega}, \vec{\theta}_u^2), j \in J_0\}$;

$\boldsymbol{pr\_pen}\left(0, L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2)\right) = P\{L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2) \geq 0\}$ = PSG function **Probability of Exceedance**;

$$\boldsymbol{AUC}(\vec{\boldsymbol{\omega}}) = 1 - P\{L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2) \geq 0\} = 1 - \boldsymbol{pr\_pen}\left(0, L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2)\right)$$

$\boldsymbol{bPOE}\left(0, L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2)\right)$ = **Buffered Probability of Exceedance**(see Nortonand Uryasev);

$\boldsymbol{bAUC}\left(L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2)\right) = 1 - \boldsymbol{bPOE}\left(0, L(\vec{\omega}, \vec{\theta}_u^1) - L(\vec{\omega}, \vec{\theta}_u^2)\right)$= buffered AUC(see Norton, and Uryasev).

Optimization Problem 0: Standard Logistic Regression (here $\vec{\theta}$ is the vector of observation of original untransformed features, with notations similar to the transformed feature vector $\vec{\theta}_u$)

maximizing Logarithms Exponents Sum

$$\max_{\vec{\omega}} \textbf{logexp\_sum}(\vec{\omega}, \vec{\theta})$$

calculation of Probability of Exceedanceat the optimal point$\vec{\omega}^*$(*needed for calculation of AUC*)

calculate

$$\boldsymbol{pr\_pen}\left(0, L(\vec{\omega}^*, \vec{\theta}^1) - L(\vec{\omega}^*, \vec{\theta}^2)\right)$$

Optimization Problem 2

maximizing Logarithms Exponents Sum using transformed features

$$\max_{\vec{z}} \textbf{logexp\_sum}(\vec{\omega}, \vec{\theta}_u)$$

calculation of Probability of Exceedanceat the optimal point$\vec{\omega}^*$(*needed for calculation of AUC*)

calculate

$$\boldsymbol{pr\_pen}\left(0, L\left(\vec{\omega}^{*}, \vec{\theta}_u^1\right) - L\left(\vec{\omega}^{*}, \vec{\theta}_u^2\right)\right)$$

Optimization Problem 3

minimizing buffered probability of exceedance (bPOE) with transformed features(equivalent to maximization of bAUC)

$$\min_{\vec{\boldsymbol{\omega}}} \boldsymbol{bPOE}\left(0, L\left(\vec{\boldsymbol{\omega}}, \vec{\theta}_u^1\right) - L\left(\vec{\boldsymbol{\omega}}, \vec{\theta}_u^2\right)\right)$$

calculation of Probability of Exceedanceat the optimal point $\vec{\omega}^{*}$(*needed for calculation of AUC*)

calculate

$$\boldsymbol{pr\_pen}\left(0, L\left(\vec{\boldsymbol{\omega}}^{*}, \vec{\theta}_u^1\right) - L\left(\vec{\boldsymbol{\omega}}^{*}, \vec{\theta}_u^2\right)\right)$$

Optimization Problem 4

minimizingProbability of Exceedanceusing transformed features (equivalent to maximization of AUC)

$$\min_{\vec{\boldsymbol{\omega}}} \boldsymbol{pr\_pen}\left(0, L\left(\vec{\boldsymbol{\omega}}, \vec{\theta}_u^1\right) - L\left(\vec{\boldsymbol{\omega}}, \vec{\theta}_u^2\right)\right)$$