CASE STUDY: Logistic Regression and Regularized Logistics Regression Applied to Estimating the Probability of Cesarean Section (logexp_sum, polynom_abs, cardn, logistic, crossvalidation)

This case study finds an optimal estimate of the cesarean section rate in a women population. The risk of difficult labor is described by a probabilistic model that depends on measurable demographic factors. We evaluated the effects of demographic factors on the probability of Cesarean section. This case study considers 6 primary factors: age, height, weight, maternal weight gain, gestational age, and birth weight. Background for this case study is described in Chen et al. (2004).

We considered four formulations of the logistic regression optimization problem:

- Problem 1. Maximization of the log-likelihood function ("plain vanilla" logistic regression).
- Problem 2. Maximization of the log-likelihood function minus additional regularization term (regularized logistic regression).
- Problem 3. Maximization of the log-likelihood function subject to constraint on cardinality.
- Problem 4. Cross-Valiadtion applied to Maximization of the log-likelihood function.

Problem 1 was implemented in PSG by maximizing the log-likelihood function which is a standard PSG function `("logexp_sum")`. This problem formulation was considered in Chen et al (2004).

The regularization term in Problem 2 was subtracted from the log-likelihood function to improve the out-of-sample performance of the regression model. The regularization is very popular in data-mining applications, see for instance, Shi et al (2008). For regularization we used the `"polynom_abs"` function, which is a standard function of PSG. Coefficients for this polynomial absolute function were obtained with the steepest descent algorithm which optimizes out-of-sample performance.

The constraint on cardinality in the Problem 3 was used to reduce the number of factors and improve the out-of-sample performance of the regression model.

Problem 4 is the 4-fold Cross-Validation for the Maximization of the log-likelihood (which was done in Problem 1). In each pass we selected ¾ of the data as in-sample dataset on which we calibrated the model. Then we tested the performance of the models on the remaining (out-of-sample) ¼ part of data to observe how the model predicts the probability of Cesarean section.

Initial Data

Total number of factors (features): I= 6
Total number of scenarios: J=12,690

References

Chen, G., Uryasev, S., and T.K. Young (2004): On the prediction of the cesarean delivery risk in a large private practice. American Journal of Obstetrics and Gynecology, 191, 617-25

Shi W., Wahba, G., Wright S, Lee, K., Klein, R, Klein, B. (2008): LASSO-Patternsearch algorithm with application to ophthalmology and genomic data. Stat Interface., 1(1), 137-153.

Notation

I= total number of factors; *i=1,…,I* is the index of the factors;

J= number of scenarios; *j=1,…,J* is the index of the scenarios;

$\vec{x} = (x_0, x_1 \ldots . x_I)$ is the vector of decision variables;

θ_{ij} = value of factor *i* in scenario *j* ;

Logarithms Exponents Sum `(logexp_sum)` function (log-likelihood function in logistic regression):

$$\text{logexp\_sum}\left(\vec{x}\right) = \sum_j p_j \ln\left[\frac{c_j \exp\left\{\sum_i \theta_{ij} x_i\right\}}{c_j \exp\left\{\sum_i \theta_{ij} x_i\right\} + \left(1 - c_j\right)\exp\left\{-\sum_i \theta_{ij} x_i\right\}}\right],\ 0 < c_j \le 1;$$

Polynomial Absolute `(polynom_abs)` regularization function:

$$\text{polynom\_abs}\left(\vec{x}\right) = \sum_i \eta_i \left|x_i\right|;$$

Cardinality function (cardn) function counts the number of scaled nonzero elements of a decision vector with certain precision. It is defined as follows:

$$\boldsymbol{cardn}(\vec{x}, w) = \sum_i \boldsymbol{u}(x_i, w),$$

where

$$\boldsymbol{u}(y, w) = \begin{cases} 0, & \text{if } -w < y < w \\ 1, & \text{otherwise} \end{cases},$$

w is a threshold value.

Card = upper bound on number of scaled nonzero components of a decision vector $\vec{x}$.

Optimization Problem 1

Maximizing log-likelihood

$$\max_x \left(\text{logexp\_sum}\left(\vec{x}\right)\right)$$

calculation of Logistic function

Value:

$$\text{logistic}(\vec{x})$$

Optimization Problem 2

Maximizing regularized log-likelihood

$$\max_x \left(\text{logexp\_sum}\left(\vec{x}\right) - \text{polynom\_abs}\left(\vec{x}\right) \right)$$

calculation of Logistic function

Value:

$$\text{logistic}(\vec{x})$$

Optimization Problem 3

Maximizing log-likelihood

$$\max_x \left(\text{logexp\_sum}\left(\vec{x}\right) \right)$$

subject to

constraint on cardinality

$$\text{cardn}(\vec{x}, w) \leq Card \qquad \text{(CS.4)}$$

calculation of Logistic function

Value:

$$\text{logistic}(\vec{x})$$

Optimization Problem 4

CrossValidation

Maximizing log-likelihood

$$\max_x \left(\text{logexp\_sum}\left(\vec{x}\right) \right)$$

calculation of Logistic function

Value:

$$\text{logistic}(\vec{x})$$