Traveler's 2020 modeling competition

Insurance modeling

Jun Jin, Ruofan Chen

Department of Statistics
University of Connecticut

December 4, 2020

Table of contents

To handle the unbalanced data

- The failure of Accuracy. AUC or Gini index instead.

		Truth	
		0 majority	1minority
	0 majority	990	10
Predict	1minority	0	0

- Two general solution: Sampling or weighted loss.

$$1 : 100 \to 100 : 100$$
$$\to 1 : 1$$

$$\ell_T(x, y) = w_0 \ell(x, y \,|y = 0) + w_1 \ell(x, y \,|y = 1)$$
$$w_0 : w_1 = \#\{y = 1\} / \#\{y = 0\}$$

UCONN
UNIVERSITY OF CONNECTICUT

Combined model of frequency and severity

In the chapter 6 of the book

Predictive modeling applications in actuarial science

author gives a modeling pattern for the cost:

$$\begin{aligned}
&frequency = claim_count/exposure \\
&severity = loss/claim_count \\
&frequency \sim \quad P(\lambda) \quad or \quad NB(n, m, \mu) \\
&severity \sim G(\alpha, \beta)
\end{aligned}$$

Tweedie regression

In the article

- An index which distinguishes between some important exponential families

by Maurice Tweedie, the author proposed a tweedie distribution constructed by

$$Cost = X_1 + \cdots + X_N$$
$$X_i \overset{iid}{\sim} G(\alpha, \beta)$$
$$N \sim P(\lambda)$$

Comparison between Freq-Severity and Tweedie:

- Loss Cost Modeling vs. Frequency and Severity Modeling, Jun Yan

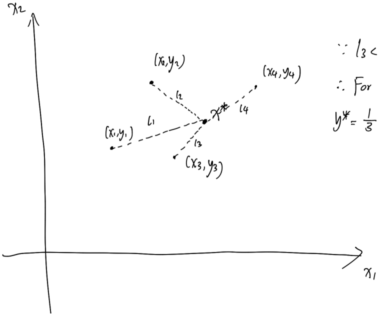
Link is here.

Sample based matching from the external data

K-nearest-neighbor (KNN), with $\left\{\vec{x}_i, y_i\right\}_1^n$ as the samples, for a new given $\vec{x}^*$, we set

$$y^* = \frac{1}{k}\sum_{j\in K} y_j$$

$$K = \arg\min_{T,|T|=k} \sum_{i\in T} \left\|\vec{x}_i - \vec{x}^*\right\|^2$$



$\because l_3 < l_4 < l_2 < l_1$

$\therefore$ For 3-NN,

$y^* = \frac{1}{3}(y_2 + y_3 + y_4)$.

Sample based matching from the external data

The data scource: The dataset "dataCar" under the package "insuranceData".

Possible enhancement: Using kernel trick. In K-NN, the sample with different distance have same weights, while we can use the kernel function like Gaussian kernel to assign them with weights "proportional" to the distance.

$$y^* = \sum_{j \in K} w_j y_j$$

$$K = \underset{T, |T| = k}{\arg\min} \sum_{i \in T} \left\| \vec{x}_i - \vec{x}^* \right\|^2$$

$$\textstyle\sum_j w_j = 1$$

$$w_j \propto \exp\left(- \frac{\left\| \vec{x}_i - \vec{x}^* \right\|^2}{2\sigma^2} \right)$$

Sufficient dimension reduction (SDR) and SDR with categorical data

With the continuous x.

$$y \perp x \,\Big|\, \beta^T x$$

Sliced inverse regression for dimension reduction, Li, Ker-Chau

With the continuous x and categorical W.

$$y \perp x \,\Big|\, \left(\beta^T x, W\right)$$

Sufficient dimensions reduction in regressions with categorical predictors, Li, Bing

UCONN UNIVERSITY OF CONNECTICUT

Special Correlation measures

Distance correlation is hot topic in stats recently

Measuring and testing dependence by correlation of distances, Szekely, Gabor J

It is used to measure the distance between vector, so it can be adopted in the testing issue related to features and target. We test:

$$T^* = \frac{ndCov\left(X_i, Y\right)}{\frac{1}{n^2}\sum_{k.l=1}^{n}\left\|X_{ik} - X_{il}\right\|_p \frac{1}{n^2}\sum_{k.l=1}^{n}\left\|Y_k - Y_l\right\|_p} > \left(\Phi^{-1}\left(1 - \alpha/2\right)\right)^2$$

Advantage: Regardless of the continuous or discrete type of variables

Useful info

Git: Git here.

Our department website is Department of Statistics.

Our website for statistical data science lab at Uconn is Data Science Lab