This post was republished to Yifan&#039;s Blog at 3:42:32 PM 2/4/2016

Comments on Imbalanced GLM

To make it simple, we assume . For logistic regression models unbalanced training data affects **only** the estimate of the model intercept. This only skews your predicted probabilities (), but doesn't change the rank of the predictions, hence have no effect on AUC or other rank based measurements. Here is the reason.

The basic LRM is , or equivalently, . Here  is also known as the log relative risk. The logistic regression is indeed a **probability model** and uses MLE:



Taking derivative on right hand side of this formula [[1]](#footnote-1)[1]leads to:



Besides, the meaning of  could be interpreted as the log-odds ratio corresponding to , i.e. . As we overstated, the logistic regression is indeed a **probability model**, and MLE is a consistent estimator, and meanwhile THE **asymptotically unbiased most powerful test statistic** to . The performance of using LRM is promised.

But  
a. How should we tune the skewness?  
b. How LRM + penalized behaves?

## Exingting approaches

1. Bias Correction method proposed by King and Zeng (<http://gking.harvard.edu/relogit>). This seems to have been very popular with political scientists but it may not be the best approach. However, see their papers for examples of the problem.
2. Penalized Maximum Likelihood Estimation proposed by Firth (Stata program: Joseph Coveney’s firthlogit, available from SSC)
3. **Exact logistic regression** (Cyrus and Nitin, 1995) only works when N is very small (< 200), and works best when covariates are discrete (preferably dichotomous) and the number of covariates is very small. But it requires a great deal of memory and hence usually won’t work with bigger problems.

Only the first two items are covered in this document.

## Correction

Notation:

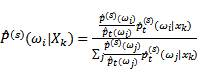
1. : i-th label, ;
2.  (t=training) prediction iii. prediction

Considering a Bayesian setting and updating  using EM, (Patrice, et.al) proposed an algorithm to correct the imbalanced model, for any *k*-th out of  new observation  and each class:

**INITIALIZATION**:



**EM-s-*th*-STEP**:

*E-Step*: 

*M-Step*: 

## Penalized LRM, Penalized-MLE approach

Main reason to use PMLE:

1. thumb of rules: , or ;
2. multicollinearity;
3. Perfect fit fails (no power).

The estimation used in PMLE is indeed: s



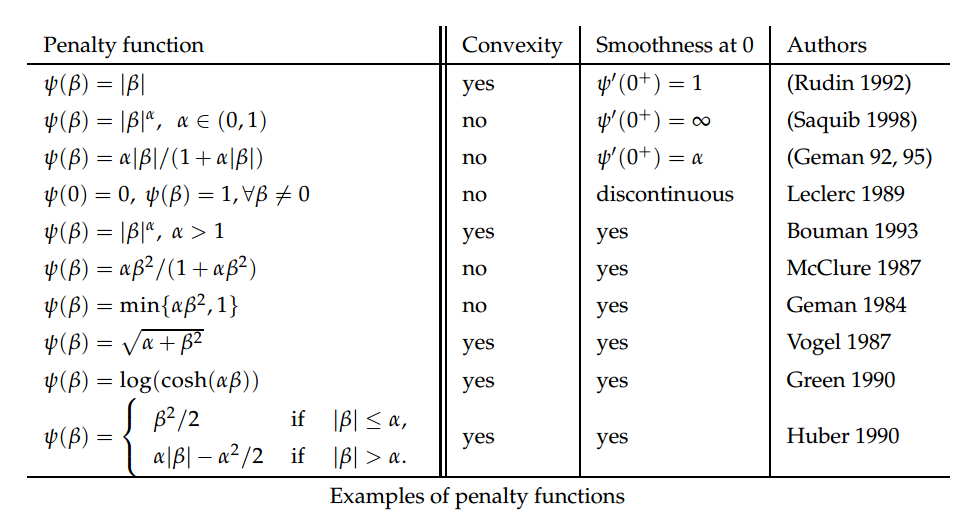
One form of the penalty term is:



Here  are weights. Usually, the penalty term should follow oracle properties (Fan, 2001):

1. unbiasedness
2. sparsity
3. stability

Unfortunately, original LASSO , ) doesn't have such properties. Adaptive LASSO (Zou, 2003) modified the original one by differing . Here is a short summation of different penalty term functions:



## Refs

1. King, Gary, and Langche Zeng. "Logistic regression in rare events data." *Political analysis* 9.2 (2001): 137-163.
2. Zou, Hui. "The adaptive lasso and its oracle properties." *Journal of the American statistical association* 101.476 (2006): 1418-1429.
3. Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96.456 (2001): 1348-1360.
4. Saerens, Marco, Patrice Latinne, and Christine Decaestecker. "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure."*Neural computation* 14.1 (2002): 21-41.
5. Latinne, Patrice, Marco Saerens, and Christine Decaestecker. "Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing." *ICML*. 2001.
6. Mehta, Cyrus R., and Nitin R. Patel. "Exact logistic regression: theory and examples." *Statistics in medicine* 14.19 (1995): 2143-2160.

1. [↑](#footnote-ref-1)