

# Sample Selection Bias

Yanjun Qi

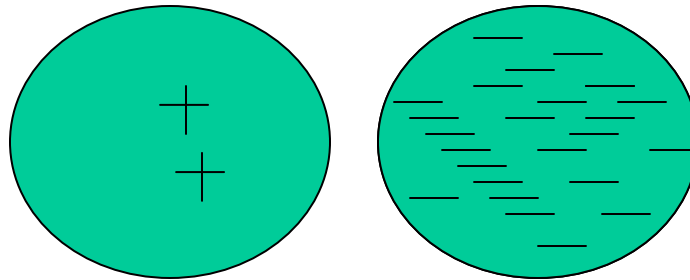
2005.09 BLM Seminar

# Paper covered

- Logistic Regression in Rare Event Data
  - Gary King and Langche Zeng
  - Appeared in *Political Analysis. 2001*
- Learning and Evaluating classifiers under sample selection bias
  - Bianca Zadrozny
  - *ICML 2004*

# Rare Event Data

- Focus on binary classification here
- Dozens to thousands of times fewer **ones** ('events') than **zeros** ('nonevents')



# Rare Event Data

- In literatures, proven to be difficult to predict
- Two problems:
  - Popular statistical procedure, such as logistic regression, sharply underestimate the probability of rare events
  - Commonly used data collection inefficient for rare event data

# Rare Event Data

- For First Problem:
  - Popular statistical procedure, such as logistic regression, sharply underestimate the probability of rare events
  - Some reported methods to make corrections (not cover here)

# Rare Event Data

- For second problem:
  - Commonly used data collection inefficient for rare event data
  - Reason: The fear of collecting data with too few events has led to data collections with
    - huge numbers of observations
    - but relatively few events observations, and poorly measured related features
  - More efficient sampling designs exist for making valid inference

# Rare Event Data

- More efficient sampling designs exist for making valid inference
  - For example: sampling all available events and a tiny fraction of nonevents
  - Enable to save as much as 99% of data collection costs or / and be able to collect much more meaningful (expensive) feature variables

# Sample Selection Bias

- Standard classifier assume
  - Examples  $(x, y)$
  - Commonly we draw independently from a distribution  $D$  with  $(X * Y)$
- Sampling
  - Examples  $(x, y, s)$
  - $S$  controls the selection of examples ( 1 means selected, 0 means not selected )
  - We have only access to  $S=1$  examples

# Four Cases

- Regarding  $s$  dependences on  $(x,y)$ 
  - 1.  $S$  is independent of  $x$  and independent of  $y$ 
    - Just a random sample from  $D$
    - Commonly used in data collection
  - 2.  $S$  is independent of  $y$  given  $x$ 
    - $P(s|x,y)=P(s|x)$
    - Selected examples are biased
    - The biasness only depends on  $x$

# Four Cases

- 3.  $s$  is independent of  $x$  given  $y$ 
  - $P(s|x,y)=P(s|y)$
  - Selected examples are biased
  - The biasness only depends only on label  $y$
  - Corresponding to change in the prior probabilities of labels
  
- 4. No dependence assumption hold between  $x$ ,  $y$  and  $s$ 
  - Selected examples are biased
  - We could not hope to learn a mapping from features to labels
  - Unless: have access to addition features  $X_s$  that control the selection for examples ( even ones with  $s = 0$ )
  - $P(s|x_s, x, y) = P(s|x_s)$

# Case 3<sup>rd</sup> Sampling : choice-based sampling

- $s$  is independent of  $x$  given  $y$ 
  - $P(s|x,y)=P(s|y)$
  - Corresponding to change in the prior probabilities of labels
  - When one of the  $Y$  is rare in population, considerable resources in data collection can be saved by randomly selecting within categories of  $Y$
  - Known in econometrics as **choice-based / endogenous stratified sampling**

# Case 3<sup>rd</sup> Sampling : choice-based sampling

- Select on Y can be consistent and efficient but only with the appropriate statistical corrections
- Commonly use two corrections for choice-based sampling
  - 1. Prior correction
  - 2. Weighting

# Case 3<sup>rd</sup> Sampling : choice-based sampling

- Prior correction
  - Prior about the fraction of ones in population:  $t$
  - Observed fraction of ones in the sample:  $p$
  - For the logit model,  $x \cdot \text{beta\_1} + \text{beta\_0}$
  - $\text{beta\_1}$  statistically consistent
  - $\text{beta\_0}$  need the correction:

$$\hat{\beta}_0 = \ln\left[\left(\frac{1-t}{t}\right)\left(\frac{p}{1-p}\right)\right]$$

# Case 3<sup>rd</sup> Sampling : choice-based sampling

- Weighting
  - Maximize weighted MLE
    - Prior about the fraction of ones in population:  $t$
    - Observed fraction of ones in the sample:  $p$

$$\text{weighted\_LL} = \omega_1 * \sum_{y_i=1} p(y_i | x_i) + \omega_0 * \sum_{y_i=0} p(y_i | x_i)$$

$$\omega_1 = t / p$$

$$\omega_0 = (1-t) / (1-p)$$

# Case 3<sup>rd</sup> Sampling : choice-based sampling

- Weighting vs. Prior correction
  - Weighting outperform prior correction when
    - both a large sample available
    - and functional form is mis-specified
  - Weighting is asymptotically less efficient than prior correction
    - Effect that can be seen in small samples
    - The differences are not large
  - In most cases, weighting preferable when t is available

# Case 2<sup>nd</sup> Sampling

- **S** is independent of **y** given **x**
  - $P(s|x,y)=P(s|x)$
  - The biasness only depends on **x**
  - Called exogenous stratified sampling in econometrics
  - In order to make  $P(s|x,y)=P(s|x)$  valid in practice, the input to the classifier has to include all the variables that affect the sampling selection

# Case 2<sup>nd</sup> Sampling

- The paper separate classifier learners into two kinds:
  - Local: the output of the learner depends asymptotically only on  $P(y|x)$ 
    - Case 2<sup>nd</sup> sampling does not affect this kind of learner
  - Global: the output of the learner depends asymptotically both on  $P(x)$  and  $P(y|x)$ 
    - Case 2<sup>nd</sup> sampling affects this kind of learner, because the bias change  $P(x)$

# Case 2<sup>nd</sup> Sampling - Bayesian classifier

- Bayesian classifier
  - By using the biased sample as training data

$$\frac{P(x | y, s = 1)P(y | s = 1)}{P(x | s = 1)} = P(y | x, s = 1) = P(y | x)$$

- The biased sample contains
  - More examples in parts of the features space where  $P(s=1|x)$  is high
  - Less examples in parts of the features space where  $P(s=1|x)$  is low

# Case 2<sup>nd</sup> Sampling

- Bayesian classifier
  - As long as  $P(s=1|x)$  is greater than 0 for all  $x$
  - As sample size increase, the results on a selected sample will asymptotically approach the results on a random sample

# Case 2<sup>nd</sup> Sampling - NB

- Naïve Bayes Classifier

$$\frac{P(x_1 | y, s = 1) \dots P(x_n | y, s = 1) P(y | s = 1)}{P(x | s = 1)}$$

- There are no independence relationship between each  $x_i$ ,  $y$  and  $s$
- **→** The estimation affected by the 2<sup>nd</sup> sampling bias

# Case 2<sup>nd</sup> Sampling – Logistic Regression

- Logistic regression

$$P(y = 1 | x, s = 1) = P(y = 1 | x) = \frac{1}{1 + \exp(\beta * x)}$$

- Not affected by 2<sup>nd</sup> sampling bias
- Asymptotically, as long as  $P(s=1|x)$  is greater than 0 for all  $x$ ,
  - Results on a selected samples approach the results on a random sample

# Case 2<sup>nd</sup> Sampling – DT

- Decision Tree

- Splitting criteria different for different tree learners

- CART: GINI index

$$1 - \sum_y P(y|t)$$

- C4.5: Information Gain

$$- \sum_y P(y|t) \log(P(y|t))$$

- All splitting criteria is maximal when examples equally distributed among the classes and minimal when all examples belong to one class

# Case 2<sup>nd</sup> Sampling – DT

- Decision Tree

- The splitting criteria are dependent on

$$P(y | t)$$

- In general:  $P(y | t, s = 1) \neq P(y | t)$

- The DT learner is affected by the 2<sup>nd</sup> sampling bias

# Case 2<sup>nd</sup> Sampling – SVM

- Support Vector Machine
  - For hard margin case:
    - The classes do not have overlapping
    - Assume the selection probability  $P(s=1|x)$  is greater than zero for all  $x$
    - The 2<sup>nd</sup> sample bias does not affect the boundary
  - For soft margin case:
    - The classes do have overlapping
    - 2<sup>nd</sup> Bias does affect the boundary because changing in  $P(x)$

# Case 2<sup>nd</sup> Sampling – Correcting bias

- The paper provided a bias correction method for all classifiers
  - Provided we have a model for the selection probability  $P(s=1|x)$
  - The method works by
    - correcting the distribution of examples through resampling
    - and then applying classifiers on the corrected samples

# Case 2<sup>nd</sup> Sampling – Correcting bias

- The method bears resemblance to
  - Weighting (Statistics literature)
  - Costing (machine learning literature)
- Details in paper