


Learning from Only Positive and Unlabeled Data Using Generalized Stochastic Frontier Model

Liuxia Wang
Sentrana Inc

Joint Statistical Meeting
San Diego, CA
08/02/2012



1. Motivation and Introduction
2. Literature Review
 - 1) PU Classification
 - 2) Stochastic Frontier Model
3. Generalized Stochastic Frontier Model for PU Classification
4. Simulation Studies
5. Real Application - Penetration

1. **Business Goal:** Predict whether a customer needs certain items but didn't purchase from the particular vendor (e.g. Food distributors, retailers) yet so that the vendor should issue a coupon to the customer to induce the customer to purchase. In marketing, this is called **penetration**.
2. **Typical solution** is classification. The training set contains both real need (positive labels) and real no-need (negative labels) of items from a set of customers.
3. **Difficulties**
 - I. Food distribution industry is very competitive. Each restaurant usually has more than two vendors. The biggest food distributor only has about 20% market share and 30% of wallet share on average.
 - II. For each customer, we only have the transactions from this particular seller. Hence we can only know what they need from the history, but we don't know for sure that they don't need the items if they didn't purchase from this particular vendor.
- III. Technically, this means that the training data for classification only contains positive labels and the unlabeled data.

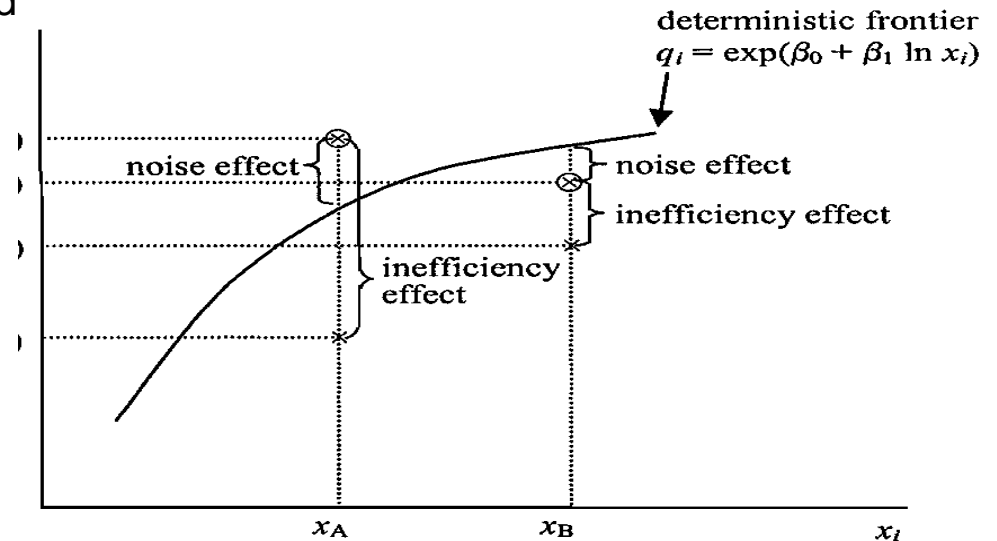
1. Intensive application in real world, e.g. Text mining, Species Detection, Marketing, etc.
2. Some Existing Methods
 - I. **Two-step classification approach** (Liu, et al, 2002; Li and Liu, 2003; Yu et al, 2002)
 - i. Use model to sort out the reliable negative labels from the unlabeled data
 - ii. Build classifier using the true positive labels v.s. the predicted reliable negative labels.
 - II. **Weighted approach:**
 - i. SVM - Add more weights to positive labels, and less on unlabeled data (Liu, et al, 2005; Liu, Z. et al, 2005; Elkan and Noto, 2008). E.g. Biased SVM (Liu, et al 2005).
 - ii. Probabilistic model with weights to control the different sampling rates (Ward, et al., 2009)

1. **Stochastic frontier model (SFM)** was first proposed by Aigner et al (1977) and Meeusen and van den Broeck (1977) independently.
2. SFM is an econometric model to estimate inefficiency effects in production. It adds a one-sided error term to the traditional linear regression to explicitly capture the amount of deviation of actual production from the expected production.
3. Mathematically, the model is formulated

$$y_t = f(x_t) + \epsilon_t \pm u_t, \quad t = 1, \dots, T$$

$$u_t = |U_t| \text{ where } U_t \sim N(0, \sigma_u^2)$$

$$\epsilon_t \sim N(0, \sigma^2)$$



4. To extend the model, Greene (2005) introduced model to capture the **heterogeneity** by allowing the means of the underlying normal distribution of the inefficiency to be non-zero, and the non-zero mean can be explained using other variables

$$U_t \sim N(g(z_t), \sigma_u^2).$$

Generalize Stochastic Frontier Model for PU Classification

1. Generalized SFM to handle non-Gaussian Data. For binary data, the model is formulated as

$$Y_i \sim \text{Bernoulli}(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = X_i'\beta - u_i$$
$$u_i \sim |N(Z_i'\gamma, \sigma^2)|$$

2. Where

- I. The variables in X are used to predict the true output
- II. The variables in Z are used to explain what causes the inefficiency.
- III. Y are the binary observations, with Y = 1 being the observed positive label and Y = -1 being unlabeled observations

3. Interpretation in PU Classification

- I. The true positive and negative labels are explained using the X variables.
- II. The positive and negative labels from the unlabelled observations are differentiated using the inefficiency term u_i , which is explained using Z variables. In another words, the reason that the positive labels from the unlabelled data are not observed are explained by the variables in Z.

4. The model is estimated using **Markov chain Monte Carlo**.

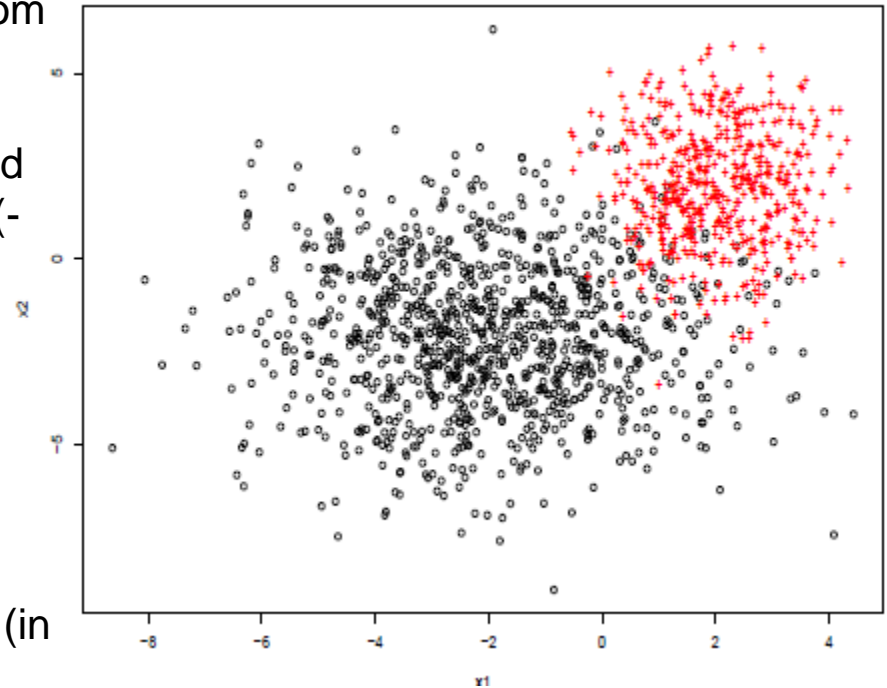
1. Data Generation

I. True Data

- i. 500 positive labels are generated from Normal distribution with mean (2,2) and variance (1, 1.5)
- ii. 1000 unlabelled points are generated from Normal distribution with mean (-2, -2) and variance (2, 2).

II. Observed Data

- i. A certain percent of true positive labels are **randomly** selected., and they joint with the negative labels to form the unlabeled data.
- ii. The percentage takes the value 0% (in which all labels are correct), 50%, 90% and 95%.



2. For each data set, run both SFM (with X matrix being the two dimension of X, and Z matrix being the intercept) and biased SVM to compare the performance. The comparison is based on 1000 repetitions.

Percentage	SFM	Biased SVM
0%	0.9428	0.9408
50%	0.9415	0.9368
90%	0.9262	0.9163
95%	0.9205	0.9024

1. Data Generation

I. True Data

- i. It's data to predict the localization site of protein. Eight predictors are available.
- ii. The number of positive label is 244 and the number of negative labels is 429.

II. Observed Data

- i. A certain percent of true positive labels are **randomly** selected., and they joint with the negative labels to form the unlabeled data.
- ii. The percentage takes the value 0% (in which all labels are correct), 50%, 90% and 95%.

2. For each data set, run both SFM (with X matrix containing eight predictors, and Z matrix being the intercept) and biased SVM to compare the performance. The comparison is based on 1000 repetitions.

Percentage	SFM	Biased SVM
0%	0.7735	0.7635
50%	0.7828	0.7367
90%	0.7302	0.7052
95%	0.5503	0.4646

1. The vendor is food distributor, and their customers are the restaurants.
2. Food distribution industry is well-grown and highly-competitive
3. The biggest food distributor only has about 20% of market share and 30% of wallet share
4. When customer made purchase before, we know for sure that customer needs the items, i.e. positive labels. However, if customer didn't purchase from this vendor, there are two possibilities: either they don't need, i.e. negative labels or they purchased from other vendors, i.e. positive labels.
5. For this competitive industry, the observed true labels can be very limited.
6. There are ~450 product categories.

1. **Response variable** Y: $Y = 1$ if customer purchased fresh beef from the vendor in the past, and $Y = -1$ (unlabeled) otherwise.
2. **Predictors X** to capture the true purchase probability
 - I. Customer Cuisine Code
 - II. Price range
 - III. Service type
 - IV. Purchase incidence of substitute products
 - V. Purchase incidence of complimentary products
 - VI. Demographic information
 - VII. Seasonality
 - VIII. Holiday effect
 - IX. Macro-Economic data, such as food CPI, inflation, etc.
3. **Predictors Z** to capture the inefficiency of purchasing from this vendor
 - I. Customers' wallet share with this vendor
 - II. The salesperson's capability
 - III. The vendor's market share in the local market
4. The **sample size** is ~1,000,000 customers.

1. Testing Data:

- I. Our restaurant experts read the restaurants' menu and made intelligent guess on what they need and what they don't need based on their expertise knowledge.
- II. Sample size is 11 restaurants from 11 different cuisines for each product category

2. Result: The average of F-score is .5784 over 450 categories

- 3. Computation:** The code is written in C and run 6-core machine in parallel. Totally it takes 16days for 450 product category. But biased SVM takes much longer, hence I didn't have comparison result from biased SVM.

1. Extend stochastic frontier model to non-Gaussian data. The model is used to solve the classification problem with only positive and unlabeled data.
2. The model is evaluated using simulated data, and it's at least competitive with some existing approaches.
3. The model is used in practice to address marketing problem: predict if a customer need a particular product or not, then penetrate the account. The model is further tested by clients in real world



1725 Eye St. NW, Suite 900
Washington DC, 20006

OFFICE 202.507.4480

FAX 866.597.3285

WEB sentrana.com

The Science to Lead Markets™