CS229 Machine Learning Project: Allocating funds to the right people

Dugalic, Adem adugalic@stanford.edu
Nguyen, Tram tram@stanford.edu

Introduction

Our project is motivated by a common issue in developing economies. In many developing countries, government programs targeting the poor face the problem of identifying who the poor households are. In particular, while there are observable characteristics of the households from survey data, income is usually unobserved for the very poor. A common approach is obtaining a list of poor households from the local administrative office, or the village chief. One problem that arises from this approach is corruption at the local level, as the officers might have incentives to misreport the poorest households to some extent to benefit their own relatives and friends. Ignoring the corruption might lead to a large misallocation of financial resources away from the people who are in real need of those. Our goal is to propose a method that reduces the errors of classification created by the local administrative offices. The method relies on the assumption that the reports coming from administrative offices, while suffering from corruption, are still somewhat informative of the poverty in the village. This is because local authorities also care about their reputation in the community and do not misreport completely. Information from the administrative offices, therefore, might be partially representative of the poverty ranking of households in the community, but not entirely accurate. Given a sample of lists of poor households provided by local authorities with varying corruption rates that are independent across villages, we aggregate this data and use it to obtain a superior identification of poor households. Due to the independence of errors, characteristics of individuals wrongly classified as poor in one village will be similar to those of the rich in other villages. We propose two machine learning algorithms to deal with the issue. Our better performing algorithm makes a significant improvements in identifying poor households in the population.

Methodologies and Algorithms

Definition of the Problem

Suppose there are K villages with M_k households living in village k for each $k \in \{1, ..., K\}$. Thus, the size of the training sample is $M = \sum_{k=1}^{K} M_k$. Each household is characterized by income y and a vector of features x, which which are described in the next section. Let $\mathcal{D}$ be the distribution over (y, x). Suppose that the government has limited resources that it wishes to distribute to H households in the population that are considered poor. The government obtains from each village the list of households that have earnings below income $\bar{y}$, where $\bar{y}$ characterizes the poverty line. The information that each village representative provides to the government is corrupt. In particular, let $z^{(i,k)} = 1$ if household i in village k is reported to have income no higher than $\bar{y}$, and $z^{(i,k)} = 0$ otherwise. We assume that if $y^{(i,k)} \leq \bar{y}$ then $z^{(i,k)} = 1$ with probability $1-\tau_k$ and $z^{(i,k)} = 0$ otherwise, and if $y^{(i,k)} > \bar{y}$ then $z^{(i,k)} = 0$ with probability $1 - \tau_k$ and $z^{(i,k)} = 1$ otherwise, with misreports independent across households. Note that we allow for different villages to have different misreporting rates. We assume that $\tau_k < \frac{1}{2}$ for each $k \in \{1, ..., K\}$, but is otherwise an unknown parameter that needs to be estimated. In addition to $z^{(i,k)}$, the government observes the vector of characteristics $x^{(i,k)}$ for each household. The goal of the exercise is to find H most likely poor households in the population given data $\{(z^{(i,k)}, x^{(i,k)})\}_{(1,1)\leq(i,k)\leq(M_k,K)}$. For this project we set $H = 3,000$.

Dataset and Features

For the purpose of testing how well our algorithms do, we use the already available US household data extracted from the 2015 American Community Survey. For this project we randomly draw $50,000$ training units from the data. Household characteristics of interests are number of household members, the head's age

and education, dwelling characteristics such as stove, fridge, television, etc., mortgage payment, food stamp eligibility and spending on gas, water and fuel. These features are strongly correlated with household income that we also obtain in the dataset. We treat different US states as different "villages" for the purpose of the analysis. Therefore, our dataset consists of 51 "villages" (including Puerto Rico). We then use a threshold of poverty line to construct the variable $z^{(i,k)}$ that indicates whether household i in village k lives below the poverty line. We construct variable $z^{(i,k)}$ in the following way: we initially define $z^{(i,k)}$ to be one 1 if household's income is below the poverty line, and 0 otherwise; we then endow each village k with the probability of misreporting $\tau_k = \frac{\beta_k}{2}$, where β_k is randomly drawn from Beta(2,4); finally, for each village k we randomly switch the values of $z^{(i,k)}$ at rate τ_k independently across households. The U.S. Census Bureau defines the income poverty threshold in 2016 for the family consisting of two adults and two children to be \$24,036, and thus we set $\bar{y} = 24,036$. We simulate the dataset 200 times and assess the performance of our algorithms over the simulated datasets.

Methods

The label of interest in our problem is whether a household is poor, $s^{(i)} \equiv 1(y^{(i)} \leq \bar{y})$. We do not observe the labels, which places us in the unsupervised learning setting. We do, however, observe imperfect signal of the variable of interest, $z^{(i)}$, which distinguishes our setup from standard unsupervised learning problems. We apply two types of EM algorithms: discriminative EM and generative EM.

Discriminative EM

On top of the assumptions outlined in the definition of the problem, we merely assume that $p(s = 1) = \frac{1}{1+e^{-\theta' x}}$ for some parameter vector $\theta \in R^n$. We make use of the following conditional probabilities in our algorithms:

$$p(s|z, x; \theta, \tau) = \frac{p(z|s, x; \theta, \tau_k)p(s|x; \theta, \tau_k)}{p(z|x; \theta, \tau_k)} = \frac{(1-\tau_k)^{1-|z-s|}\tau_k^{|z-s|}e^{-\theta' x(1-s)}}{(1-\tau_k)^{1-z}\tau_k^{z}e^{-\theta' x} + (1-\tau_k)^{z}\tau_k^{1-z}}$$

$$p(z, s|x; \theta, \tau_k) = p(z|s, x; \theta, \tau_k)p(s|x; \theta, \tau_k) = (1-\tau_k)^{1-|z-s|}\tau_k^{|z-s|}\frac{e^{-\theta' x(1-s)}}{1+e^{-\theta' x}}$$

Given conditional distributional assumptions on s and z, the log likelihood is,

$$l(\theta, \tau) = \sum_{k=1}^{K}\sum_{i=1}^{M_k} \log p\left(z^{(i,k)}\middle|x^{(i,k)}; \theta, \tau_k\right) = \sum_{k=1}^{K}\sum_{i=1}^{M_k} \log \sum_{s=0}^{1} p\left(z^{(i,k)}, s\middle|x^{(i,k)}; \theta, \tau_k\right)$$

The EM algorithm consists of initializing the vector of parameters (θ, τ) and repeatedly carrying out the following two steps until convergence:

- (E-step) For each $k = 1, ..., K$ and $i = 1, ..., M_k$, set $Q_{i,k}(s) = p(s|z^{(i,k)}, x^{(i,k)}; \theta, \tau)$.
- (M-step) Set $(\theta, \tau) := \arg\max_{(\theta,\tau)} \sum_{k=1}^{K}\sum_{i=1}^{M_k}\sum_{s=0}^{1} Q_{i,k}(s) \log \frac{p\left(z^{(i,k)}, s\middle|x^{(i,k)}; \theta, \tau_k\right)}{Q_{i,k}(s)}$, which simplifies to

$$\tau_k = \frac{1}{M_k}\sum_{i=1}^{M_k}\left(Q_{i,k}(0)z^{(i,k)} + Q_{i,k}(1)(1-z^{(i,k)})\right) \qquad \text{for} \quad k = 1, ..., K$$

$$\theta := \arg\max_{\theta} \sum_{k=1}^{K}\sum_{i=1}^{M_k} \log\left(\frac{e^{-\theta x^{(i,k)}}Q_{i,k}(0)}{1+e^{-\theta' x^{(i,k)}}}\right)$$

Intuitively, τ_k is an expected fraction of misreported households in village k (under probabilities $\{Q_{i,k}\}_{i=1}^{M_k}$). The update for θ is somewhat related to the estimate of the logistic regression, in which targets are Bernoulli distributions with success probabilities $Q_{i,k}(1)$ instead of hard Bernoulli outcomes. The update for θ can be obtained with gradient accent applied to the objective stated above.

Having obtained estimates θ and τ, we compute

$$\hat{p}^{(i,k)} \equiv p(s^{(i,k)} = 1|z^{(i,k)}, x^{(i,k)}; \theta, \tau_k) = \frac{(1-\tau_k)^{1-|z^{(i,k)}-s^{(i,k)}|}\tau_k^{|z^{(i,k)}-s^{(i,k)}|}e^{-\theta' x^{(i,k)}(1-s^{(i,k)})}}{(1-\tau_k)^{1-z^{(i,k)}}\tau_k^{z^{(i,k)}}e^{-\theta' x^{(i,k)}} + (1-\tau_k)^{z^{(i,k)}}\tau_k^{1-z^{(i,k)}}}$$

Generative EM

Our feature vector consists of numerous variables, some continuous and some discrete, and we find making standard Gaussian distributional assumptions on the feature vector inappropriate. We firstly transform the feature vector to its principle components. We use the first four principal components because they capture almost all of the variation in the data (using more than four principal components causes numerical convergence issues). Abusing the notation, we refer to the transformed feature vector for individual i as $x^{(i)}$. We thus assume that $x^i|s \sim \mathcal{N}(\mu_s, \Sigma_s)$, and that $s \sim \text{Bernoulli}(\rho)$. We make use of the following conditional probabilities in our algorithm

$$p(s|z, x; \mu, \Sigma, \rho, \tau) = \frac{(1-\tau_k)^{1-|z-s|}\tau_k^{|z-s|}\phi\left(\Sigma_s^{-\frac{1}{2}}(x-\mu_s)\right)(1-\rho)^{1-s}\rho^s}{(1-\tau_k)^{1-z}\tau_k^z\phi\left(\Sigma_0^{-\frac{1}{2}}(x-\mu_0)\right)(1-\rho) + (1-\tau_k)^z\tau_k^{1-z}\phi\left(\Sigma_1^{-\frac{1}{2}}(x-\mu_1)\right)\rho}$$

$$p(z, s, x; \mu_s, \Sigma_s, \rho, \tau) = p(z, x|s; \mu_s, \Sigma_s, \tau_k)p(s; \rho) = (1-\tau_k)^{1-|z-s|}\tau_k^{|z-s|}\phi\left(\Sigma_s^{-\frac{1}{2}}(x-\mu_s)\right)(1-\rho)^{1-s}\rho^s$$

The log likelihood is

$$l(\mu, \Sigma, \rho, \tau) = \sum_{k=1}^{K}\sum_{i=1}^{M_k}\log p\left(z^{(i,k)}, x^{(i,k)}; \mu, \Sigma, \rho, \tau_k\right) = \sum_{k=1}^{K}\sum_{i=1}^{M_k}\log\sum_{s=0}^{1} p\left(z^{(i,k)}, x^{(i,k)}, s; \mu_s, \Sigma_s, \rho, \tau_k\right)$$

The EM algorithm consists of initializing vector of parameters $(\mu, \Sigma, \rho, \tau)$ and repeatedly carrying out the following two steps until convergence:

- (E-step) For each $k = 1, ..., K$ and $i = 1, ..., M_k$, set $Q_{i,k}(s) = p(s|z^{(i,k)}, x^{(i,k)}; \mu, \Sigma, \rho, \tau)$.
- (M-step) Set $(\mu, \Sigma, \rho, \tau) := \arg\max_{(\mu,\Sigma,\rho,\tau)} \sum_{k=1}^{K}\sum_{i=1}^{M_k}\sum_{s=0}^{1} Q_{i,k}(s)\log\frac{p\left(z^{(i,k)}, s, x^{(i,k)}; \mu_s, \Sigma_s, \rho, \tau_k\right)}{Q_{i,k}(s)}$ which simplifies to

$$\tau_k = \frac{1}{M_k}\sum_{i=1}^{M_k}\left(Q_{i,k}(0)z^{(i,k)} + Q_{i,k}(1)(1-z^{(i,k)})\right) \qquad \text{for } k = 1, ..., K$$

$$\mu_s = \frac{\sum_{k=1}^{K}\sum_{i=1}^{M_k} Q_{i,k}(s)x^{(i,k)}}{\sum_{k=1}^{K}\sum_{i=1}^{M_k} Q_{i,k}(s)}$$

$$\Sigma_s = \frac{\sum_{k=1}^{K}\sum_{i=1}^{M_k} Q_{i,k}(s)(x^{(i,k)}-\mu_s)(x^{(i,k)}-\mu_s)^T}{\sum_{k=1}^{K}\sum_{i=1}^{M_k} Q_{i,k}(s)}$$

$$\rho = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M_k} Q_{i,k}(1)$$

Having obtained estimates μ, Σ, ρ, τ_1,...,τ_K, we compute

$$\begin{aligned}\hat{p}^{(i,k)} \equiv & p(s^{(i,k)} = 1 | z^{(i,k)}, x^{(i,k)}; \mu, \Sigma, \rho, \tau) \\ = & \frac{(1-\tau_k)^{z^{(i,k)}} \tau_k^{1-z^{(i,k)}} \phi\left(\Sigma_1^{-\frac{1}{2}}(x^{(i,k)} - \mu_1)\right) \rho}{(1-\tau_k)^{1-z^{(i,k)}} \tau_k^{z^{(i,k)}} \phi\left(\Sigma_0^{-\frac{1}{2}}(x^{(i,k)} - \mu_0)\right)(1-\rho) + (1-\tau_k)^{z^{(i,k)}} \tau_k^{1-z^{(i,k)}} \phi\left(\Sigma_1^{-\frac{1}{2}}(x^{(i,k)} - \mu_1)\right) \rho}\end{aligned}$$

We then for both methods find the households with the highest H order statistics of the set $\{\hat{p}^{(i,k)} : 1 \leq i \leq M_k, 1 \leq k \leq K\}$, as well as identify households who are classified as poor (i.e. $\hat{p}^{(i,k)} > 0.5$). Since we do observe information on reported earnings in our sample we can assess the effectiveness of the tests in several ways: by finding the fraction of poor households among the selected 3000 households; by finding the fraction of poor households among the households predicted to be poor; by comparing estimated τ_1,...,τ_K with the true misreporting rates across villages. We also assess how our algorithm improves upon a naive method that ignores misreporting and randomly selects H households from the set of households that are reported to be poor.

Results and Discussion

Figure 1 shows the estimated misreporting rates with true misreporting rates for each discriminative and generative EMs.

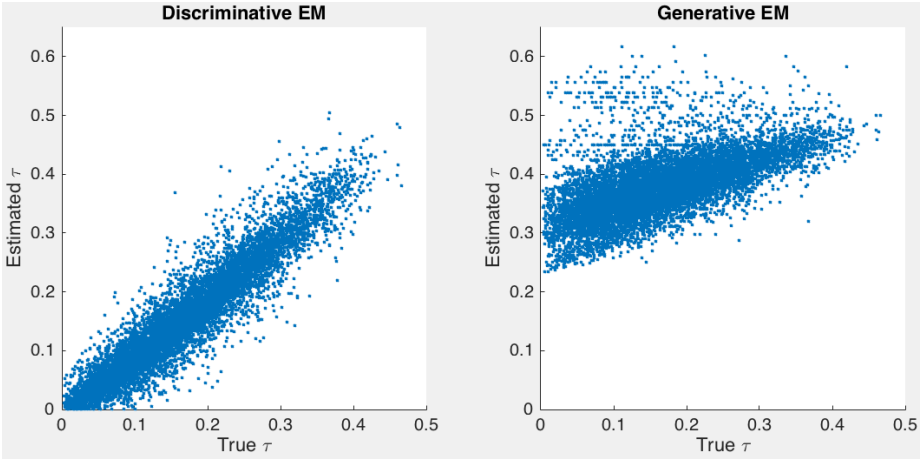

Figure 1: Estimated versus true misreporting rates

Table 1 reports mean square errors and correlation coefficients between estimated and true misreporting rates for each method.

	MSE	Correlation
Discriminative EM	0.00022062	0.94214
Generative EM	0.010148	0.60304

Table 1: Mean square errors and correlation coefficients between estimated and true misreporting rates

As it can be seen, the discriminative EM does significantly better than the generative EM. For the discriminative EM, the mean square error is 0.00022062, and the correlation coefficient is 0.94214. For the generative EM, the numbers are 0.010148 and 0.60304, respectively.

Figures 2 shows the fraction of poor households among those $H = 3000$ selected households for three different methods over 200 simulations.

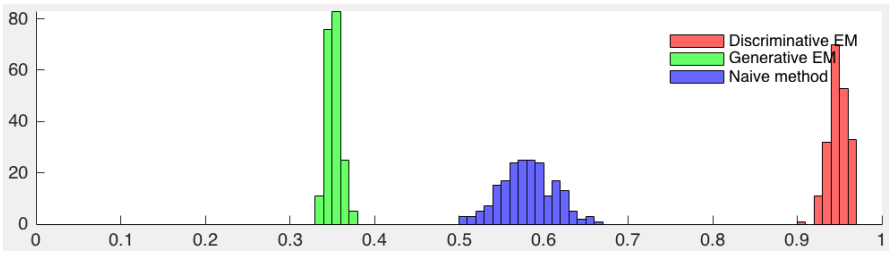

Figure 2: Fraction of poor households among selected 3000 households (200 simulations)

The mean rates corresponding to discriminative EM, generative EM and naive method are, 0.9483, 0.3518 and 0.5807, respectively.

Figures 3 shows the fraction of poor households among the households who are predicted to be poor for three different methods over 200 simulations.

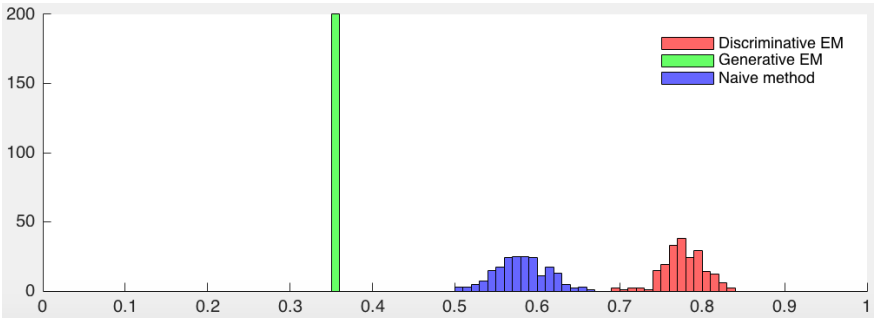

Figure 3: Fraction of poor households among those predicted to be poor (200 simulations)

The mean rates corresponding to discriminative EM, generative EM and naive method are, 0.7771, 0.3520 and 0.5807, respectively.

There are two important things to note. First, the naive method that ignores corruption allocates 58% of funds on average to poor households (Figure 2). The discriminative EM improves the allocation to about 95%. Thus, additional $37\% \times 3000 = 1110$ households who desperately need aid receive it if the discriminative EM is used to allocate aid. The second noteworthy thing is that the generative EM does very poorly, substantially worsening the allocation of the naive method. The main reason is that the underlying assumption that the features are normally distributed is wrong. The distribution over the first several principal components exhibits very fat tails.

Conclusion

In this project we build and assess performance of EM algorithms in fighting corruption and allocation of scarce financial aid to the poorest households in the society. While the naive method that ignores corruption allocates about 58% of resources to poor households, our best performing algorithm, discriminative EM, brings up this number to 95% - a huge success! We also demonstrate that the standard, generative EM performs very poorly, which is due to the fact that features are not normally distributed, but rather come from a distribution with fat tails. The method used in this paper has a potential to improve classification in a wide variety of situations in which labels are imperfectly observed. Further research could also improve upon the generative EM by making alternative assumptions over the distribution of the features.

References

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B. 39 (1): 1-38. 1977

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data, Second Edition.* John Wiley Sons, Inc. 2002