In many developing countries, government programs targeting the poor face the problem of identifying who they really are because cause income is usually unobserved for the very poor. A common approach is to obtain a list of poor households from the local administrative office, but the officers might have incentives to misreport who the poorest households are for the sake of benefitting their own relatives and friends. If the government wishes to allocate funds to the poor households in the society, who should receive the funds? We propose two machine learning algorithms to deal with the issue, and more generally, to reduce errors of classification created by using imperfect training sets. Our better performing algorithm makes a significant improvements in identifying poor households in the population.

**Data**

US household data extracted from the 2015 American Community Survey. For this project we draw a sample of 50,000 households from the data that are representative of the US population. We treat US states as different "villages". We construct the variable $z_{ik}$ that indicates whether household $i$ in village $k$ lives below the income poverty threshold $\bar{y}$. The U.S. Census Bureau sets the household income poverty threshold in 2016 at 24,036, so we choose $\bar{y} = 24,036$. We simulate the dataset 200 times and assess the performance of our algorithms over the simulated datasets.

**Features**

Household characteristics from the dataset 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.

We construct the variables $z_{ik}$ in the following way:

- First let $z_{ik} = 1$ if household $i$’s income in village $k$ is below the poverty line, and 0 otherwise.
- Then endow each village $k$ with the probability of misreporting $\tau_k = \frac{\beta_k}{\beta_k}$, where $\beta_k$ is randomly drawn from Beta(2,4).
- Finally, for each village $k$ we randomly switch the values of $z_{ik}$ at rate $\tau_k$, independently across households.

**Method**

Although we do not observe whether a household is poor, $z_{ik} = 1$ ($\bar{y}_z \leq y_{ik}$), we do observe imperfect signal of the variable of interest, $z_{ik}$, which distinguishes our setup from standard unsupervised learning problems.

We apply two types of EM algorithms: discriminative EM and generative EM. The naive method is the method that ignores misreporting and randomly selects H households from the set of households that are reported to be poor.

**Models**

**Discriminative EM**

Assume that $p(z = 1) = \frac{1}{1 + e^{-x}}$ for some parameter vector $\theta \in \mathbb{R}^n$.

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

$$l(\theta, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log p(z_{ik} = 1|z_{ik}; \theta, \tau_k) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log \left( \frac{e^{x_{ik} - \theta \cdot z_{ik}}}{1 + e^{x_{ik} - \theta \cdot z_{ik}}} \right)$$

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

- (E-Step) For each $k = 1, \ldots, K$ and $i = 1, \ldots, N_k$ set $Q_{ik}(s) = p(s|z_{ik}, z_{ik}'; \theta, \tau)$

- (M-Step) Set $\tau = \sum_{i=1}^{K} \sum_{k=1}^{N_k} Q_{ik}(s)$

**Generative EM**

Because some of our features are continuous and some discrete, we find it more appropriate to make Gaussian distributional assumptions on its principal components. We use the first four principal components because they capture almost all of the variation in the data and ease numerical computation. We refer to the transformed feature vector for individual $i$ as $x^{(i)}$. We assume that $x^{(i)} | s \sim N(\mu, \Sigma)$, and that $s \sim Bernoulli(\rho)$. The log likelihood is

$$l(\mu, \Sigma, \rho, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log p(z_{ik} = 1|z_{ik}; \mu, \Sigma, \rho, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log \left( \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \right)$$

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 $k = 1, \ldots, K$ and $i = 1, \ldots, N_k$ set $Q_{ik}(s) = p(s|x^{(i)}; \mu, \Sigma, \rho, \tau)$

- (M-Step) Set $\mu = \sum_{i=1}^{K} \sum_{k=1}^{N_k} Q_{ik}(s) x^{(i)}$ where

$$p(s|x^{(i)}, \mu, \Sigma, \rho, \tau) = (1 - \tau_k) \cdot \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(\frac{-1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)\right) + \tau_k \cdot \rho \cdot \phi(x^{(i)})$$

Having obtained estimates $\beta$ for each method, we compute $\beta^{(i)} = p(z = 1|x^{(i)}; \beta)$ where $\beta = (\theta, \tau)$ for the discriminative EM and $\beta = (\mu, \Sigma, \rho, \tau)$ for the generative EM. We then for both methods find the households with the highest $\beta^{(i)}$ scores of the set $\beta^{(i)}$ for $i = 1 \ldots N_k$, as well as identify households who are classified as poor (i.e. $\beta^{(i)} > 0.5$). We also compare estimated misreporting rates $\tau_{est}$ to $\tau_{true}$ across “villages”.

**Future Research**

The allocation may be improved even further by using a generative EM with different distributional assumptions over the features. The method used in this paper has a potential to improve classification in a wide variety of situations in which labels are imperfectly observed. A practical application would be to take the algorithm to a real targeting program in a developing country and test the performance of our algorithm against the traditional targeting approaches.