Correcting for misinformation using machine learning

Adem Dugalic† Tram Nguyen‡

Stanford University

December 15, 2016

Abstract

We propose several algorithms to reduce errors of classification created by using imperfect training sets. A classic example is misallocation of scarce funds to poor households due to unobserved income in combination with data misreporting or corruption at the administrative level. Suppose there are several training examples in which the targets are misclassified at a certain rate. Since the errors are imperfectly correlated across training examples, wrongly classified observations in one sample will share common characteristics with observations that are differently classified in other samples. We use this insight to develop several machine learning algorithms to reduce this classification error. We apply the algorithms to the context of aid programs targeting poor households with limited knowledge of their true income and poverty ranking. In our mainline specification of the problem, our main algorithm makes a 37% improvement in the allocation of the funds.

†This paper is an extended version of a project developed in a machine learning course at Stanford.
‡Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305, adugalic@stanford.edu.
§Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305, tram@stanford.edu
1 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 the true poor households. In particular, while there are observable characteristics of households from survey data, income is usually unobserved for the very poor (Glewwe, 1992). An example of a program of this nature is the popular targeting program PROGRESA (Oportunidades) that was implemented widely in Mexico. The program offers conditional cash transfers to the poor to encourage children’s school attendance and preventative health care (Skoufias et al., 2001). While PROGRESA’s targeting methodology does quite well for the extremely poor households and is less effective in identifying less poor households (Skoufias et al., 2001), in other developing countries where targeting programs are smaller in scale and where data is more limited, the level of effectiveness may be much lower.

Among the most common forms of targeting is the community-based approach in which local authorities administer the funds to the beneficiaries (Coady et al., 2004). The list of poor households is obtained from local administrative offices, or community leaders/village chiefs. One problem that arises from this approach is corruption at the local level, as officers might have incentives to misreport who the poorest households are for the sake of benefitting their own relatives and friends (Alatas et al., 2013; Basurto et al., 2017). Ignoring this elite capture problem might lead to a substantial misallocation of funds in such targeting programs. Nevertheless, this decentralized approach of distributing funds comes with its own benefits since local leaders are usually more informed about financial standing of households in their village, and are also held accountable for their decisions (Basurto et al., 2017).

The goal of this paper is to propose a machine learning algorithm that reduces errors of classification created by local administrative offices in community-based targeting programs, without having to eliminate the decentralized nature of the approach. The method
relies on the assumption that reports coming from administrative offices, while suffering from corruption, are still somewhat informative about poverty levels of households in the respective villages/communities. This is because local authorities care about their reputation in their community and could also be held accountable for corruption (Niehaus and Atanassova, 2013). As a result, misreporting is limited. Information from 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, households wrongly classified as poor in one village will share common characteristics with households classified as rich in other villages. We exploit this observation and propose two machine learning algorithms to deal with the issue. The algorithms belong to the class of expectation maximization (EM) algorithms (Dempster et al., 1977): discriminative EM and generative EM. When benchmarked against the performance of a naive algorithm, our better performing algorithm, the discriminative EM, makes a significant 37% improvement in identifying poor households in the population.

The rest of the paper is as follows. Section 2 describes in more details the example problem we create to closely mirror real-world targeting problems and the data used for this experiment. Section 3 describes the two types of algorithms we propose. Section 4 discusses the results. Finally, Section 5 concludes.

2 Problem and Data

2.1 Definition of Problem

Suppose there are $K$ villages with $M_k$ households living in village $k$ for each $k \in \{1, \ldots, K\}$. 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 $x$ 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 misreporting 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 $H = 3,000$.

**2.2 Dataset and Features**

For the purpose of testing how well our algorithms perform, we use the already available US household data extracted from the 2015 American Community Survey. For this paper we randomly draw 50,000 observations from the data that are representative of US households. Household characteristics of interests are the 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 think of these features as characteristics of households that are generally easier to observe that household’s income, but are correlated with income. We also observe income in the dataset, but use it to assess the performance of our algorithms. 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 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 $\beta_k$ is randomly drawn from Beta(2,4); finally, for each village $k$ we randomly switch the values of $z^{(i,k)}$ at rate $\tau_k$ 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 $\bar{y} = 24,036$ as the threshold for a household to be considered “poor”. We simulate the dataset 200 times and assess the performance of our algorithms over the simulated datasets.

3 Methodologies and Algorithms

3.1 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 an unsupervised learning setting. We do, however, observe an imperfect signal of the variable of interest, $z^{(i)}$, which distinguishes our setup from standard unsupervised learning problems and makes our contribution novel. We apply two types of Expectation-Maximization (EM) algorithms: the discriminative EM and the generative EM. We assess how our algorithms improve upon a naive method that ignores misreporting and randomly selects $H = 3000$ households from the set of households that are reported to be poor.

3.2 Discriminative EM

The discriminative EM algorithm does not require strict assumptions on the distribution of the explanatory variables. On top of the assumptions outlined in the definition of the problem, we merely assume that $p(s = 1) = \frac{1}{1+e^{-\theta^T x}}$ for some parameter vector $\theta \in \mathbb{R}^n$. We
make use of the following conditional probabilities in our algorithms:

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

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

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

\[
l(\theta, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \log p(z^{(i,k)}|x^{(i,k)}; \theta, \tau_k) = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \log \sum_{s=0}^{1} Q_{i,k}(s) \log \frac{p(z^{(i,k)}, s|x^{(i,k)}; \theta, \tau_k)}{Q_{i,k}(s)}
\]

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

(E-step) For each \(k = 1, ..., K\) and \(i = 1, ..., M_k\), set

\[
Q_{i,k}(s) = p(s|z^{(i,k)}, x^{(i,k)}; \theta, \tau_k)
\]

(M-step) Set

\[
(\theta, \tau) := \arg \max_{(\theta, \tau)} \sum_{k=1}^{K} \sum_{i=1}^{M_k} \sum_{s=0}^{1} Q_{i,k}(s) \log \frac{p(z^{(i,k)}, s|x^{(i,k)}; \theta, \tau_k)}{Q_{i,k}(s)}
\]

which simplifies to

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

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

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

Having obtained estimates $\theta$ and $\tau$, we compute the estimated conditional probability that household $i$ in village $k$ earns income below the poverty line, given the report $z^{(i,k)}$ and household characteristics $x^{(i,k)}$:

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

### 3.3 Generative EM

Our feature vector of household characteristics consists of numerous variables, some continuous and some discrete. Making standard Gaussian distributional assumptions on the feature vector is inappropriate in this context. Therefore, 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. Moreover, 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 $x^{i|s} \sim N(\mu_s, \Sigma_s)$, and that $s \sim \text{Bernoulli}(\rho)$. We make use of the following conditional probabilities in our algorithm

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

$$
p(z, s, x; \mu_s, \Sigma_s, \rho, \tau) = p(z, x|s; \mu_s, \Sigma_s, \tau_k)p(s; \rho) = (1 - \tau_k)^{1 - |z - s|} \frac{1}{\tau_k} \phi \left( \Sigma_s^{-\frac{1}{2}} (x - \mu_s) \right) (1 - \rho)^{1-s} \rho^s
$$
The log likelihood is
\[
l(\mu, \Sigma, \rho, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \log p(z^{(i,k)}, x^{(i,k)}; \mu, \Sigma, \rho, \tau) = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \log \sum_{s=0}^{1} p(z^{(i,k)}, x^{(i,k)}, s; \mu_s, \Sigma_s, \rho, \tau_k)
\]

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 \(i = 1, \ldots, M_k\), set
\[
Q_{i,k}(s) = p(s|z^{(i,k)}, x^{(i,k)}; \mu, \Sigma, \rho, \tau)
\]

(M-step) Set
\[
(\mu, \Sigma, \rho, \tau) := \arg \max_{(\mu, \Sigma, \rho, \tau)} \sum_{k=1}^{K} \sum_{i=1}^{M_k} \sum_{s=0}^{1} Q_{i,k}(s) \log \frac{p(z^{(i,k)}, s, x^{(i,k)}; \mu_s, \Sigma_s, \rho, \tau_k)}{Q_{i,k}(s)}
\]

which simplifies to
\[
\tau_k = \frac{1}{M_k} \sum_{s=1}^{M_k} \left( Q_{i,k}(0)z^{(i,k)} + Q_{i,k}(1)(1 - z^{(i,k)}) \right) \text{ for } k = 1, \ldots, K
\]
\[
\mu_s = \frac{\sum_{k=1}^{K} \sum_{i=1}^{M_k} Q_{i,k}(s)x^{(i,k)}}{\sum_{k=1}^{K} \sum_{i=1}^{M_k} Q_{i,k}(s)}
\]
\[
\Sigma_s = \frac{\sum_{k=1}^{K} \sum_{i=1}^{M_k} Q_{i,k}(s)(x^{(i,k)} - \mu_s)(x^{(i,k)} - \mu_s)^T}{\sum_{k=1}^{K} \sum_{i=1}^{M_k} Q_{i,k}(s)}
\]
\[
\rho = \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M_k} Q_{i,k}(1)
\]

After obtaining the estimates of \(\mu, \Sigma, \rho, \tau_1, \ldots, \tau_K\) once the algorithm converges, we again compute the estimated probability that household \(i\) in village \(k\) earns income below the
poverty line, given the report $z^{(i,k)}$ and household characteristics $x^{(i,k)}$:

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

3.4 Evaluation of performance of algorithms

For both methods we find the households with the highest $H = 3000$ 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 assess the effectiveness of the tests in several ways: by finding the fraction of poor households among the 3000 households selected by our algorithms; by finding the fraction of poor households among the households predicted to be poor; and by comparing estimated misreporting rates $\tau_1, \ldots, \tau_K$ with the true misreporting rates across "villages". We also assess how our algorithm improves upon a naive method that ignores misreporting and randomly selects $H = 3000$ households from the set of households that are reported to be poor. The naive method serves as a benchmark for us to evaluate the performance of the two EM algorithms.

4 Results and Discussion

Figure 1 shows the estimated misreporting rates with true misreporting rates $\tau$ for each discriminative and generative EMs.
Table 1 reports mean square errors and correlation coefficients between true misreporting rates and the estimated rates $\tau$ generated by each algorithm.

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MSE</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminative EM</td>
<td>0.00022062</td>
<td>0.94214</td>
</tr>
<tr>
<td>Generative EM</td>
<td>0.010148</td>
<td>0.60304</td>
</tr>
</tbody>
</table>

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

It can be seen from Table 1 and Figure 1 that the discriminative EM algorithm 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 reports the fraction of poor households among those $H = 3000$ selected house-
holds for three different methods over 200 simulations.

![Figure 2: Fraction of poor households among selected 3000 households (200 simulations)](image)

The mean rates in Figure 2 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)](image)

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 results 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 are in need of financial aid receive it if the discriminative EM is used to allocate aid. The second noteworthy observation is that the generative EM does very poorly, substantially worsening the allocation achieved by the naive method. The main reason is that the generative EM algorithm relies on the underlying assumption that the features are normally distributed. This assumption is not accurate given the nature of the data we use in this paper. The distribution over the first several principal components exhibits very fat tails.

5 Conclusion

In this project we build and assess the performance of the EM algorithms in correcting for misallocation of scarce financial resources to poor households due to potential corruption at the local levels in community-based targeting programs. While the naive method that ignores corruption allocates about 58% of resources to poor households, our best performing algorithm, the discriminative EM algorithm, brings up this number to 95%: a vast improvement over the naive approach. We also demonstrate that the standard, generative EM algorithm performs very poorly, which is due to the fact that features are not normally distributed but rather come from a distribution with fat tails. Thus, we recommend caution when applying standard normality assumptions on features in generative EM algorithms.

The method used in this paper has a potential to improve classification in a wide variety of situations in which the labels are imperfectly observed. The classification problems may be improved even further by developing a better generative EM method which uses different sets of assumptions over the distributions of the observables. Outside of the technicalities, the next step of 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.
References


