Quality control when measuring a continuous variable with noisy workers

November 7, 2012

1 Introduction

In crowdsourcing settings, we have access to a large number of people that provide answers to a variety of tasks. A key problem in such settings is that we do not know the quality of the answers provided by the participants. The key solutions proposed over the last few years all revolve around two key ideas: redundancy and gold labels. However, most of the quality control techniques are designed to operate with discrete answers. In this paper, we introduce a technique that, in the presence of noisy workers, and answers that are a continuous variable

- Estimates the most likely answer values
- Estimates the quality of the workers that return the answers

2 Model Assumptions

The basic idea is the following: The true answers $X = \{x_1, \ldots, x_n\}$ are drawn from a latent distribution $F(\theta)$; we cannot observe directly the values of X. Instead, we have a set of noisy labelers $w^1, \ldots, w^k$; each worker w_j observes one or more of the examples in X and returns a noisy estimate of the observed value. For notational purposes, we use $y_j(x_i)$ to mark the value that worker w_j returns after observing the value x_i. We use $G_j(\theta_j)$ to denote the distribution of returned values for worker w_j. The distribution $G_j(\theta_j)$ is independent of $F(\theta)$ if the worker w_j returns random results. If $F(\theta)$ is identical with $G_j(\theta_j)$, the worker is perfect (we can also have cases where $G_j(\theta_j)$ is not identical with $F(\theta)$ and the worker is still perfect). We assume that the distributions $G_j(\theta_j)$ are conditional independent of each other, given $F(\theta)$ (i.e., the workers only observe the "latent" values x_i and the value $y_j(x_i)$ is only influenced by x_i and the quality of the worker w_j, and not by the values assigned by other workers).

The bi-variate Gaussian case: We restrict our current approach to the case where $F(\theta)$ and $G_j(\theta_j)$ form a bi-variate Gaussian distribution; the quality of the worker w_j is the correlation ρ_j between $F(\theta)$ and $G_j(\theta_j)$. Since F and G_j are two Gaussians, their parameters are $\theta = \mu, \sigma$ and $\theta_j = \mu_j, \sigma_j$.

[TODO: Add a picture of the correlated distributions, showing the effect of ρ in the bi-variate distribution]

3 Estimation Algorithm I: Point-to-point

Our goal is to know the quality ρ_j (the covariance) of each worker, and the most likely correct value x_i for each data point.

We start by estimating the characteristics of the distribution G_j of the answers from each worker w_j. Given the points Y_j returned by worker w_j, we estimate the mean μ_j and variance σ_j of the G_j distribution using the usual maximum likelihood estimates:

$$\mu_j = \frac{1}{|Y_j|} \cdot \sum_{y_j(x_i) \in Y_j} y_j(x_i) \tag{1}$$

$$\sigma_j = \frac{1}{|Y_j|} \cdot \sum_{y_j(x_i) \in Y_j} (y_j(x_i) - \mu_j)^2 \tag{2}$$

Now, to compute the reliability ρ_j of each worker, we need to know the matching x_i values for each $y_j(x_i)$ label assigned by the worker. We have:

$$\begin{aligned}
P(X|Y_1 \ldots Y_k) &= \frac{P(Y_1 \ldots Y_k|X) \cdot P(X)}{P(Y_1 \ldots Y_k)} \\
&= \frac{P(X) \cdot \prod_{j=1}^{k} P(Y_j|X)}{\prod_{j=1}^{k} P(Y_j)} \\
&= \frac{P(X) \cdot \prod_{j=1}^{k} P(X|Y_j)P(Y_j)/P(X)}{\prod_{j=1}^{k} P(Y_j)} \\
&\propto \prod_{j=1}^{k} P(X|Y_j)
\end{aligned}$$

Now, we further break $P(X|Y_j)$ by splitting X into its individual points, and now each point x_i gets conditioned on the overall vector Y_j and the estimate $y_j(x_i)$ for x_i given by worker w_j.

$$P(X|Y_1 \ldots Y_k) \propto \prod_{j=1}^{k} \prod_{i=1}^{n} P(x_i|y_j(x_i), Y_j) \tag{3}$$

To estimate $P(x_i|y_j(x_i), Y_j)$ we rely on the fact that X and Y_j (and therefore x_i and $y_j(x_i)$) are drawn from a bivariate normal distribution. So, the distribution of the conditional $P(x_i|y_j(x_i), Y_j)$ is a Gaussian distribution with mean and variance given below:

$$
\begin{aligned}
P(x_i|y_j(x_i), Y_j) &= \mathcal{N}(x_i; \widehat{\mu_j}, \widehat{\sigma_j}) \\
\mathcal{N}(x; \mu, \sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp - \frac{(x-\mu)^2}{2\sigma^2} \\
\widehat{\mu_j} &= \mu + \rho_j \frac{\sigma}{\sigma_j}(y_j(x_i) - \mu_j) \\
\widehat{\sigma_j} &= \sqrt{1-\rho_j^2} \cdot \sigma
\end{aligned}
$$

We now identify the most likely values in X, given the label assignments by the workers. For that, we compute the log-likelihood of the probability:

$$
\begin{aligned}
L = -\ln\left(P(X|Y_1 \dots Y_k)\right) &\propto \sum_{j=1}^{k}\sum_{i=1}^{n} \ln(P(x_i|y_j(x_i), Y_j)) \\
&\propto \sum_{j=1}^{k}\sum_{i=1}^{n} \left(\frac{\ln(2\pi)}{2} + \ln(\widehat{\sigma_j}) + \frac{(x_i - \widehat{\mu_j})^2}{2\widehat{\sigma_j}^2} \right)
\end{aligned}
$$

Given that the values x_i are generated in an i.i.d. fashion, we can find the X that maximizes the likelihood by taking the partial derivative for each x_i and setting it to zero. This will return the "most likely" x_i values, given the values returned by the noisy workers.

$$
\begin{aligned}
\frac{\partial L}{\partial x_i} = \sum_{j=1}^{k} \frac{\partial}{\partial x_i} \left(\frac{(x_i - \mu - \rho_j \frac{\sigma}{\sigma_j}(y_j(x_i) - \mu_j))^2}{2\widehat{\sigma_j}^2} \right) &= 0 \\
\sum_{j=1}^{k} \left(\frac{x_i - \mu - \rho_j \frac{\sigma}{\sigma_j}(y_j(x_i) - \mu_j)}{\widehat{\sigma_j}^2} \right) &= 0 \\
\sum_{j=1}^{k} \left(\frac{x_i - \mu - \rho_j \frac{\sigma}{\sigma_j}(y_j(x_i) - \mu_j)}{\widehat{\sigma_j}^2} \right) &= 0 \\
\sum_{j=1}^{k} \left(\frac{x_i - \mu - \rho_j \frac{\sigma}{\sigma_j}(y_j(x_i) - \mu_j)}{(1-\rho_j^2) \cdot \sigma^2} \right) &= 0 \\
\sum_{j=1}^{k} \left(\frac{\frac{x_i - \mu}{\sigma} - \rho_j \frac{y_j(x_i) - \mu_j}{\sigma_j}}{(1-\rho_j^2)} \right) &= 0 \qquad (4) \\
& \qquad\qquad (5)
\end{aligned}
$$

To simplify the notation, we now set:

$$
z_i = \frac{x_i - \mu}{\sigma} \qquad (6)
$$

$$z_i^j = \frac{y_j(x_i) - \mu_j}{\sigma_j} \quad (7)$$

$$\beta_j = \frac{1}{1 - \rho_j^2} \quad (8)$$

$$\rho_j = \sqrt{1 - \frac{1}{\beta_j}} \quad (9)$$

Using the equations above, and substituting in Equation 4, we have:

$$\sum_{j=1}^{k} \beta_j \cdot \left(z_i - \rho_j \cdot z_i^j\right) = 0$$

Solving for z_i, we get:

$$z_i = \frac{\sum_{j=1}^{k} \beta_j \cdot \rho_j \cdot z_i^j}{\sum_{j=1}^{k} \beta_j} \quad (10)$$

As it becomes clear from thiq equation, the best point estimate that we have for z_i is a weighted average of the $\rho_j \cdot z_i^j$ values, and the weight of each contribution is equal to ρ_j (see Equation 8). It is worth noting that the $\rho_j \cdot z_i^j$ is the maximum likelihood estimate for the value of z_i when we have a single measurement from a worker with correlation ρ_j who returns a measurement equal to z_i^j.

- TODO: Baseline error: The baseline absolute error can be computed as the expected abso
- TODO: The expected error of z_i, given a distribution of ρ_j values with pdf $f(\rho)$, is $\int f(\rho)$.....
- TODO: Estimate the variance of z_i for a given set of ρ_j values.
- TODO: Estimate the variance of z_i for a probability distribution of ρ_j values.

Knowing the (normalized) values of the latent variables z_i and the assigned labels z_i^j from the workers,[1] we can now estimate the quality ρ_j of each worker as:

$$\rho_j = \frac{\sum_{i=1}^{n} z_i \cdot z_i^j}{\sqrt{\sum_{i=1}^{n} (z_i)^2 \cdot \sum_{i=1}^{n} (z_i^j)^2}} \quad (11)$$

At this point, we have estimated the quality of the different labelers, and we have the *standardized* values z_i from the original distribution. Note that, at this

[1]Notice that the z_i and z_i^j variables have a zero mean, so in Equation 11, there is no need to subtract the mean and the equation from the z values, and the equation is simplified. Of course, someone can always take the extra step of computing the empirical mean and standard deviation for the z_i and z_i^j values.

point, we can use *any* value μ and σ to generate the values x_i from the latent distribution F, and the solution will be mathematically fine. (In other words, the model is not purely "identifiable.")

As a workaround, we assume that the crowd *after quality-adjustment*, does not have a systematic bias. So, we can set the mean value μ for F, as the combination of the mean values from the G_j distributions, adjusted for the reliability of each worker, in the same way that we estimate x_i's by combining the values provided by the different workers.

$$\mu = \frac{\sum_{j=1}^{k} \beta_j \cdot \sqrt{1 - \frac{1}{\beta_j}} \cdot \mu_j}{\sum_{j=1}^{k} \beta_j} \tag{12}$$

Similarly for standard deviation σ:

$$\sigma = \frac{\sum_{j=1}^{k} \beta_j \cdot \sqrt{1 - \frac{1}{\beta_j}} \cdot \sigma_j}{\sum_{j=1}^{k} \beta_j} \tag{13}$$

Finally, we transform now each z_i value into the most likely estimate x_i:

$$x_i = \sigma \cdot z_i + \mu \tag{14}$$

Note: These equations give equal weight to labelers with same quality, independent of the number of labeled examples. This can be both good and bad. Good because no single labeler can influence the outcome, bad because we trust equally the μ_j and σ_j estimates from workers with large and with small number of submissions, while the estimates derived from large number of submissions are by definition more robust.

3.1 Algorithm

1. For each worker w_j, with submissions $Y_j = \{y_j(x_i)\}$, compute the mean μ_j using Equation 1, and the standard deviation σ_j using Equation 2.

2. For each worker w_j, transform the submissions $y_j(x_i)$ into standardized (zero-mean, standard deviation one) z-scores, using Equation 7.

3. Assume equal quality ρ_j for all workers[2] and then compute the coefficients β_j using Equation 8.

4. Use Equation 10, and compute the most likely standardized score z_i for each labeled example.

5. Using the computed z_i values, re-compute the quality ρ_j of each worker using Equation 11 and recompute the coefficients β_j.

6. Repeat Steps 4 and 5 until convergence.

[2] Any value $0 < \rho_j < 1$ should work as an initial condition, but *not* the degenerate values 0 or 1.

7. Use Equations 12, 13, and 14 to transform the z_i values into the most likely estimates x_i of the latent variables.