A Bayesian analysis of Wasserman's censoring problem

Thomas P. Minka

June 4, 1999

A bunch of patients walk into a doctor's office. The doctor takes a measurement x and, conditional on that measurement, decides whether to also measure y. His decision is decided by a biased coin with success probability $g(x)$. The resulting data set is a bunch of (x, y) pairs and a bunch of lone x values. From this data, we want to determine

$$E[y|x] = \int \int yp(y|x)dy\; p(x)dx \tag{1}$$

For simplicity, assume $p(x)$ is known. Then the lone x values provide no useful information and we can focus on the (x, y) pairs. Let y be binary so that $p(y|x)$ is Bernoulli. Let's make the most general assumption possible about $p(y|x)$: that it is a separate Bernoulli distribution for every possible value of x. In the following, we will assume x discrete but this is not essential. The conditional distribution can be parameterized by a vector θ as follows:

$$p(y = 1|x = j, \theta) = \theta_j \tag{2}$$

The quantity of interest is therefore

$$E[y|x] = \sum_j \theta_j p(x = j) \tag{3}$$

Since θ is unknown, we have to estimate $E[y|x]$, which means deciding on a loss function. Let the loss function be squared error so that the problem reduces to computing $E[\theta_j]$ given the training data.

To proceed further, we must invoke prior knowledge. Suppose our knowledge is very vague so that $p(\theta_j)$ is uniform between 0 and 1. Given the training data, the posterior for θ_j is therefore Beta with parameters $(N_{0j} + 1, N_{1j} + 1)$, where

$$N_{ij} = \text{number of occurrences of } (y = i, x = j) \tag{4}$$

So

$$E[\theta_j] = \frac{N_{1j} + 1}{N_{0j} + N_{1j} + 2} \tag{5}$$

The quantity $N_{0j} + N_{1j}$ is the number of times we observed $x = j$ *along with* a y value. On average, we expect this quantity to be close to $Ng(j)p(x = j)$, where N is the total number of patients. So $E[y|x]$ is approximately

$$\sum_j p(x = j) \frac{N_{1j} + 1}{Ng(j)p(x = j) + 2} \approx \sum_j \frac{N_{1j}}{Ng(j)} \tag{6}$$

This approximation is exactly the Horowitz-Thompson estimate. Note that it has no dependence on $p(x)$.