BRIEF ARTICLE

THE AUTHOR

1. Mixture Modesls

Following Shalizi (ref here):

We say that the distribution f is a mixture of K component distributions $f_1, \ldots, f_K$ if

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x; \theta_k)$$

where λ_k are the mixing weights, such that $\lambda_k > 0$, $\sum_k \lambda_k = 1$. This means that the data can generated to the following procedure:

$$\begin{aligned} Z &\sim \text{Multinomial}(\lambda_1, \ldots, \lambda_K) \\ X|Z &\sim f_Z \end{aligned}$$

where Z is a discrete random variable that says which component X is drawn from.

This is useful for record linkage using the FS approach because we assume there are two latent populations corresponding to matches (M) and non-matches (U), that are represented in the population of comparisons with proportions p_M and $p_U = 1 - p_M$ respectively.

Given $(i, j) \in M$ or $(i, j) \in U$, the comparison vector between two files i and j is

$$\gamma(i, j) \sim f_k, \quad k \in \{M, U\}$$

Assuming independent samples, the log likelihood for a generic mixture model for observations $(x_1, \ldots, x_n)$ is:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta) \tag{1}$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \lambda_k f_k(x_i; \theta_k) \tag{2}$$

The overall parameter vector of the model is thus $\theta = (\lambda_1, \ldots, \lambda_{K-1}, \theta_1, \theta_2, \ldots, \theta_K)$.

In record linkage, $x_i \overset{i.i.d.}{\sim} X$ and $y_j \overset{i.i.d.}{\sim} Y$, so $\gamma(x_i, y_j) \perp\!\!\!\perp \gamma(x_{i'}, y_j)$ for all $j, i' \neq i$, yet, $\gamma(x_i, y_j)$ may not be independent from $\gamma(x_i, y_{j'})$. This gives the likelihood:

$$\ell(\theta) \tag{3}$$

1.1. **Mixture Model Estimation.** As shown in Shalizi (ref), maximizing the likelihood for a mixture model is like doing a weighted likelihood maximization, where the weight of each observation x_i depends on the cluster. This is seen by taking the derivative of (**??**) with respect to one parameter θ_j,

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_j} &= \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \lambda_k f_k(x_i;\theta_k)} \lambda_j \frac{\partial f_j(x_i;\theta_j)}{\partial \theta_j} \\ &= \sum_{i=1}^{n} \underbrace{\frac{\lambda_j f_j(x_i;\theta_j)}{\sum_{k=1}^{K} \lambda_k f_k(x_i;\theta_k)}}_{w_{ij}} \frac{\partial \log f_j(x_i;\theta_j)}{\partial \theta_j}\end{aligned}$$

The weight is the conditional probability that observation i belongs to cluster j :

$$w_{ij} = \frac{\lambda_j f_j(x_i;\theta_j)}{\sum_{k=1}^{K} \lambda_k f_k(x_i;\theta_k)} = \frac{P(Z=j, X=x_i)}{P(X=x_i)} = P(Z=j|X=x_i)$$

So if we try to estimate the mixture model, we're doing weighted maximum likelihood, with weights given by the posterior cluster probabilities (which depend on parameters $\lambda_1, \ldots, \lambda_K$ that we are trying to estimate). The EM algorithm (ref here for Rubin, etc.) makes estimation possible:

(1) Start with guesses about the mixture components $\theta_1, \theta_2, \ldots, \theta_K$ and the mixing weights $\lambda_1, \ldots, \lambda_K$.
(2) Until nothing changes very much:
    (a) Using the current parameter guesses, calculate the weights w_{ij} (E-step)
    (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates (M-step)
(3) Return the final estimates for $\theta_1, \ldots, \theta_K, \lambda_1, \ldots, \lambda_K$ and clustering probabilities.