Summary of "Bayesian Sets"

Matt Dickenson `mcd31`
STA571/CS590.01 Due: 26 March, 2014

Ghahramani and Heller (2005) describe a method for responding to a query, consisting of a few items, with additional items in the same set. More formally, a user provides a small subset of items $\mathcal{D}_c \subset \mathcal{D}$ that are assumed to represent a cluster, and the algorithm provides a completion set $\mathcal{D}'_c \subset \mathcal{D}$. For a large inventory of potential items, computing the completion set can be understood as computing the probability than an item x also belongs to $\mathcal{D}_c$, $p(x|\mathcal{D}_c)$. Incorporating a prior on x, we can compute a score:

$$\begin{aligned} \text{score}(x) &= p(x|\mathcal{D}_c) \\ &= \frac{p(x, \mathcal{D}_c)}{p(x)p(\mathcal{D}_c)} \\ &\propto p(\mathcal{D}_c|x), \end{aligned}$$

using Bayes rule and the fact that the prior probability of D_c is a multiplicative constant independent of x. Higher scores suggest that x should be included in the completion set.

How does score computation scale to large inventories of potential items? Ghahramani and Heller show that computing scores for a full inventory reduces to one sparse matrix multiplication. The entire data set $\mathcal{D}$ can be put into one large matrix $\mathbf{X}$ with J columns with Bernoulli entries and a Beta(α_j, β_j) prior on each column. For this setting, the vector s of (logged) scores can be computed as

$$\begin{aligned} s &= c + \mathbf{X}q \\ c &= \sum_j \log(\alpha_j + \beta_j) - \log(\alpha_j + \beta_j + N) + \log \tilde{\beta}_j - \log \beta_j \\ q &= \log \tilde{\alpha}_j - \log \alpha_j - \log \tilde{\beta}_j + \log \beta_j, \end{aligned}$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ refer to the posterior α and β, respectively.

The model is robust to prior specifications, returning results competitive with Google Sets with very little tuning of the prior. When naive subjects were supplied with set completions from Google and the Bayesian Sets method, a supermajority of subjects preferred the Bayesian set completions. Of course, this ranking method is only appropriate when there is sufficient reason to assume that the items in the query truly represent a single class/concept. One promising next step for the method is allowing the researcher to set the size of the response set. A threshold probability for setting this size should follow naturally from the Bayesian framework put forth here.