Schema inference

Sam Gershman

May 4, 2021

1 Generative model

This section describes the probabilistic generative model that we impute to the subject.

At each time step t, a new state $s_t \in \{1, \dots, S\}$ is sampled from a transition distribution $T(s_t|s_{t-1}, z_{t-1})$ conditional on the current state s and current schema z_t. The schema is sampled from a sticky CRP:

$$P(z_t = k|\mathbf{z}_{1:t-1}) \propto \begin{cases} N_{tk} + \beta\delta[z_{t-1}, k] & \text{if } k \text{ is old} \\ \alpha & \text{if } k \text{ is new} \end{cases} \tag{1}$$

where N_{tk} is the number of times schema k has been sampled prior to t, $\alpha \geq 0$ is a concentration parameter, and $\beta \geq 0$ is a stickiness parameter. Note that "new" here refers to the *first* unused schema (in theory there's an infinite number of unused schemata). Finally, the transition distribution is sampled from a symmetric Dirichlet distribution:

$$T(\cdot|s, k) \sim \text{Dir}(\lambda). \tag{2}$$

The sparsity parameter $\lambda \geq 0$ controls the shape of the prior. When $\lambda = 1$, the prior is uniform. When $\lambda < 1$, the prior has symmetric peaks at 0 and 1, meaning that the deterministic transition distributions are favored. When $\lambda > 1$, the prior is peaked at $1/S$, favoring a uniform transition distribution.

2 Bayesian inference

Exact Bayesian inference over T and $\mathbf{z}$ is intractable, because the number of possible schema histories explodes exponentially. So we will adopt the "local maximum *a posteriori*" (local MAP) approximation. We will also marginalize over the transition distribution rather than update a point estimate, $\hat{\mathbf{z}}$, defined below.

Schema inference is given by:

$$P(z_t|\mathbf{s}_{1:t}, \hat{\mathbf{z}}_{1:t-1}) \propto P(s_t|\mathbf{s}_{1:t-1}, z_t, \hat{\mathbf{z}}_{1:t-1})P(z_t|\hat{\mathbf{z}}_{1:t-1}), \tag{3}$$

where the second term on the right hand side is the sticky CRP given above, and the first term is the marginal likelihood, obtained by marginalizing over the transition distribution:

$$P(s_t = j|s_{t-1} = i, \mathbf{s}_{1:t-2}, z_t = k, \hat{\mathbf{z}}_{1:t-1}) = \frac{\lambda + M_{tkij}}{S\lambda + \sum_{j'} M_{tkij'}}, \tag{4}$$

where M_{tkij} is the number of $i \to j$ transitions observed prior to t when the active schema was k. The predictive distribution over the next state is given by:

$$P(s_t|\mathbf{s}_{1:t-1}, \hat{\mathbf{z}}_{1:t-1}) = \sum_k P(s_t|\mathbf{s}_{1:t-1}, z_t = k, \hat{\mathbf{z}}_{1:t-1})P(z_t = k|\hat{\mathbf{z}}_{1:t-1}). \tag{5}$$

The point estimate for the schema history is updated as follows:

$$\hat{z}_t = \operatorname*{argmax}_k P(z_t = k|\mathbf{s}_{1:t}, \hat{\mathbf{z}}_{1:t-1}). \tag{6}$$

In other words, we "freeze" the schema history to be the locally optimal point estimate.