Noisy channel generative model

Emily Kellison-Linn, Tim O'Donnell, Elias Stengel-Eskin

1 Model hyperparameters

We define a set of top-level PLUs U_T, a set of bottom-level PLUs U_B, a number of HMM states per bottom PLU n_h, a number of Gaussian components per HMM state n_g, and a dimensionality for the Gaussian distributions n_d.

2 Model parameters

1. Draw distributions E over edit operations conditioned on the next top level PLU q for each top PLU from a Dirichlet distribution with parameters $\boldsymbol{\alpha_q}$:

   $E_q \sim D(\boldsymbol{\alpha_q})\ \forall q \in U_T$

2. Draw distributions C over Gaussian component selection for each HMM state s for each bottom-level PLU from a Dirichlet distribution with parameters $\boldsymbol{\alpha_s}$:

   $C_s \sim D(\boldsymbol{\alpha_s})\ \forall s \in all\_HMM\_states$

3. Draw n_d-dimensional Gaussian distributions with mean μ and covariance matrix Σ for each Gaussian component c for each HMM state for each bottom-level PLU from a Normal-Gamma distribution:

   $\mu_c, \Sigma_c \sim NormalGamma(\mu'_c, \lambda_c, \alpha_c, \beta_c)$

3 Generative process

1. Start with a given sequence of of top-level PLUs, $a_1, a_2, ...a_{N_{top}}$, where $a_i \in U_T \forall i$.
2. For i in range $1...N_{top}$:
   (a) Sample an edit operation e from E_{a_i}.
   (b) If $e = insert\_bottom(r)$ for some $r \in U_B$, append r to the list of bottom PLUs.
   (c) if $e = insert\_top(a_i)$, set $i = i + 1$.
   (d) If $e = substitute(a_i, r)$ for some $r \in U_B$, append r to the list of bottom PLUs and set $i = i + 1$.

   The result is a sequence of bottom-level PLUs $b_1, ...b_{N_{bot}}$ where $b_i \in U_B \forall i$.

3. For each bottom PLU b_i in the bottom-level sequence:

   (a) Sample an HMM state sequence $s_1, ...s_{N_{states}}$through b_i with all initial probability on the first state $s = 1$ and transition matrix M, where $m_{x,y} = P(s_{t+1} = y|s_t = x)$ and $|M| = (n_h, n_h)$:

   $$M_{x,x} = 0.5$$
   $$M_{x,x+1} = 0.5$$

4. For each HMM state s in the HMM state sequence:

   (a) Sample a Gaussian component c from C_s

   (b) Sample an n_d-dimensional vector from the Gaussian distribution with mean μ_c and covariance matrix Σ_c.