

Appendix

Label prediction model (Section 3.2.a, derivation of Eqn. 3)

Label prediction model $P(d_i, c_i | a_i, x_i, z_i)$ is defined as follows:

$$P(d_i, c_i | a_i, x_i, z_i) = P(d_i | c_i, a_i, x_i, z_i) P(c_i | a_i, x_i, z_i)$$

1. Coarse-level label distribution

$$P(c_i | a_i, x_i, z_i) = P_{z_i}^c(c_i | a_i, x_i)$$

where $P_{z_i}^c(c_i | a_i, x_i)$ is a classifier with outputs normalized to 1, i.e., $\sum_{c_i} P_{z_i}^c(c_i | a_i, x_i) = 1$.

2. Detailed-level conditional label distribution

$$P(d_i | c_i, a_i, x_i, z_i) \propto P_{z_i}^d(d_i | a_i, x_i) [c_i = f(d_i)]$$

where $P_{z_i}^d(d | a, x)$ is a classifier with output normalized to 1, i.e., $\sum_d P_{z_i}^d(d | a, x) = 1$ (sum over all the detailed label values), and $[c = f(d)] = 1$ if c is the parent of d in the label hierarchy, and 0 otherwise. In doing so, we share a single detailed classifier across different coarse label classes. The normalizing constant of $P(d_i | c_i, a_i, x_i, z_i)$ can be written as

$$\sum_d P_{z_i}^d(d | a_i, x_i) [c_i = f(d)] = \sum_{d \in s(d_i)} P_{z_i}^d(d | a_i, x_i).$$

where the first sum is over all the detailed label values, and the second sum is over all the siblings of d_i , which all share the same parent, c_i .

3. Detailed-level label distribution can be derived by summing out the coarse-level label variable:

$$P(d_i | a_i, x_i, z_i) = \frac{P_{z_i}^d(d_i | a_i, x_i)}{\sum_{d \in s(d_i)} P_{z_i}^d(d | a_i, x_i)} P_{z_i}^c(c[d_i] | a_i, x_i)$$

Inference algorithm (Section 4, derivation of Eqn. 5)

The joint distribution of the model can be written as

$$P(\mathbf{a}, \mathbf{d}, \mathbf{c}, \mathbf{z}, \theta | \alpha, \mathbf{x}) = P(\theta | \alpha) \prod_i P(d_i, c_i | a_i, x_i, z_i) P(a_i | z_i) P(z_i | \theta).$$

- We can integrate out θ due to the conjugacy property

$$\begin{aligned} P(\mathbf{z}, \mathbf{a}, \mathbf{d}, \mathbf{c} | \alpha, \mathbf{x}) &= \int_{\theta} P(\mathbf{a}, \mathbf{d}, \mathbf{c}, \mathbf{z}, \theta) d\theta \\ &= \prod_i P(d_i, c_i | x_i, a_i, z_i) P(a_i | z_i) \times \frac{\Gamma(\sum_k \alpha_k) \prod_k \Gamma(\alpha_k + \sum_i \delta(z_i, k))}{\prod_k \Gamma(\alpha_k) \Gamma(\sum_k \alpha_k + N)}. \end{aligned}$$

- To use Gibbs sampling, we want to derive $P(z_i | \mathbf{z}_{-i}, \mathbf{a}, \mathbf{x})$, $P(z_i | \mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x})$, and $P(z_i | \mathbf{z}_{-i}, \mathbf{a}, \mathbf{c}, \mathbf{x})$. In the following, we use $P(z_i | \mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x})$ as an example (the other two are similar). Note that

$$\begin{aligned} P(z_i = k | \mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x}) &\propto P(z_i = k, \mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x}) \\ &= C \cdot P(d_i | a_i, x_i, z_i = k) P(a_i | z_i = k) \Gamma(\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k) + 1) \prod_{l \neq k} \Gamma(\alpha_l + \sum_{j \in S \setminus i} \delta(z_j, l)) \\ &= C \cdot P(d_i | a_i, x_i, z_i = k) P(a_i | z_i = k) (\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k)) \prod_l \Gamma(\alpha_l + \sum_{j \in S \setminus i} \delta(z_j, l)) \\ &\propto P(d_i | a_i, x_i, z_i = k) P(a_i | z_i = k) (\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k)) \end{aligned}$$

in which we make use of the property of the Gamma function: $\Gamma(x+1) = x\Gamma(x)$.

Learning label prediction models (Section 5, derivation of Eqns. 7, 9, 10)

In M-step, we have the following likelihood function (ignoring other irrelevant terms):

$$\begin{aligned}\mathcal{L} &= \sum_{n,i} \langle \log P(d_i^n | a_i^n, x_i^n, z_i^n) \rangle_{q^d(z_i^n)} + \gamma \sum_{t,i} \langle \log P(c_i^t | a_i^t, x_i^t, z_i^t) \rangle_{q^c(z_i^t)} \\ &= \sum_{n,i} \{ \langle \log P_{z_i^n}^d(d_i^n | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} + \langle \log P_{z_i^n}^c(c_i^n | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} \\ &\quad - \langle \log \sum_{d' \in c[d_i^n]} P_{z_i^n}^d(d' | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} \} + \gamma \sum_{t,i} \langle \log P_{z_i^t}^c(c_i^t | a_i^t, x_i^t) \rangle_{q^c(z_i^t)}\end{aligned}$$

- Assume that the output of the coarse label classifier can be approximated by the detailed label classifier (i.e., they are consistent during training), we have

$$P_{z_i}^c(f(d_i) | a_i, x_i) \approx \sum_{d \in s(d_i)} P_{z_i}^d(d | a_i, x_i).$$

Then the log likelihood function can be simplified as

$$\mathcal{L} \approx \sum_{n,i} \langle \log P_{z_i^n}^d(d_i^n | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} + \gamma \sum_{t,i} \langle \log P_{z_i^t}^c(c_i^t | a_i^t, x_i^t) \rangle_{q^c(z_i^t)}$$

- Denote the parameters in the k th detailed-label classifier as ν_k , and consider the gradient of \mathcal{L} w.r.t. ν_k :

$$\frac{\partial \mathcal{L}}{\partial \nu_k} = \sum_{n,i} q^d(z_i^n = k) \frac{\partial}{\partial \nu_k} \log P_k^d(d_i^n | a_i^n, x_i^n).$$

- Denote the parameters in the k th coarse-label classifier as μ_k , then the gradient of \mathcal{L} w.r.t. μ_k can be written as

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{t,i} q^c(z_i^t = k) \frac{\partial}{\partial \mu_k} \log P_{z_i^t}^c(c_i^t | a_i^t, x_i^t).$$