Molecular Classification 2

BIOF2014

Problem

We want to classify tumours based on their molecular characteristics, as determined by RNA expression profiles. Each expression profile consists of the expression levels of selected genes, measured as counts of molecules.

We have previously mathematically derived posterior predictive distribution $p\left(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathbf{X}, \boldsymbol{y}\right)$. for a new class label $\tilde{y}$ given new data $\tilde{\boldsymbol{x}}$.

If we are willing to tolerate computational inefficiency, we can also use an Markov chain Monte Carlo sampler such as Stan to estimate the posterior predictive distribution (or any quantities of interest).

Model

Given training data with expression profiles $\mathbf{X} = [\boldsymbol{x}_i^\top]$ and class labels $\boldsymbol{y} = [y_i]$ for $i \in \{1 \ldots N\}$, we want to predict the unknown label $\tilde{y}$ of a new sample with expression profile $\tilde{\boldsymbol{x}}$.

Each expression profile consists the detected counts of transcript molecules of J genes, so $\boldsymbol{x}_i, \tilde{\boldsymbol{x}} \in \mathcal{N}_0^J$. Suppose there are K classes, so $y_i, \tilde{y} \in \{1 \ldots K\}$.

Define our model as follows:

$$\begin{aligned}
\boldsymbol{X}_i = \boldsymbol{x} \mid Y_i = y, \boldsymbol{\eta}_y &\sim \text{Multinomial}\left(m_i \mid \boldsymbol{\eta}_y\right) \\
\boldsymbol{\eta}_k &\sim \text{Dirichlet}\left(\boldsymbol{c}_k\right) \\
Y_i &\sim \boldsymbol{\theta} \\
\boldsymbol{\theta} &\sim \text{Dirichlet}\left(\boldsymbol{d}\right),
\end{aligned}$$

where $\boldsymbol{c}_k \in \mathcal{R}_{\geq 0}^J$ and $\boldsymbol{d} \in \mathcal{R}_{\geq 0}^K$ are hyperparameters.

Tasks

1. Implement the classification model in Stan and R and sample from $p\left(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathbf{X}, \boldsymbol{y}\right)$.
2. Implement the previously derived $p\left(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathbf{X}, \boldsymbol{y}\right)$.
3. Compare the empirical pmf of from MCMC sampling vs. the mathematically-derived posterior predictive distribution $p\left(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathbf{X}, \boldsymbol{y}\right)$.