Лабораторная работа №3

1. Данные (файл time_series.dat) представляют собой временной ряд наблюдений количества регистрируемых частиц излучения, поступающего из туманности Ориона (Orion Nebula Cluster). Пусть Y_i – наблюдаемое значение за временной интервал измерений i. Рассматривается следующая статистическая модель:

$$
\begin{aligned}
Y_i|k,\theta,\lambda \sim & \text{Poisson}(\theta), \qquad i=1,\dots,k \\
Y_i|k,\theta,\lambda \sim & \text{Poisson}(\lambda), \qquad i=k+1,\dots,n
\end{aligned}
$$

Априорные распределения выбираются следующим образом:

$$
\begin{aligned}
\theta|b_1 \sim & \text{Gamma}(0.5, b_1) \\
\lambda|b_2 \sim & \text{Gamma}(0.5, b_2) \\
b_1 \sim & \text{IG}(0.01, 1) \\
b_2 \sim & \text{IG}(0.01, 1) \\
k \sim & \text{Uniform}(2,\dots,n-1),
\end{aligned}
$$

где

- Poisson(λ) обозначает распределение Пуассона;
- Gamma(α,β) обозначает гамма-распределение с плотностью распределения
$$\text{Gamma}(x,\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta};$$
- IG(α,β) обозначает обратное гамма-распределение с плотностью распределения
$$\text{IG}(x,\alpha,\beta) = \frac{e^{-\frac{1}{\beta x}}}{\Gamma(\alpha)\beta^\alpha x^{\alpha+1}}.$$

Требуется реализовать схему Гиббса для оценки параметров $(\theta,\lambda,k,b_1,b_2)$.

Указания:

(a) Записать выражение для совместного распределения модели:

$$\pi(\mathbf{Y},\theta,\lambda,k,b_1,b_2) \propto \prod_{i=1}^{n} \pi(Y_i|\theta,\lambda,k,b_1,b_2)\pi(\theta|b_1)\pi(\lambda|b_2)\pi(b_1)\pi(b_2)\pi(k)$$

(b) Для реализации схемы Гиббса из совместного распределения находятся полные условные распределения:

$$\pi(\theta|k,\lambda,b_1,b_2,\mathbf{Y}),$$

$$\pi(\lambda|k, \theta, b_1, b_2, \mathbf{Y}),$$

$$\pi(k|\theta, \lambda, b_1, b_2, \mathbf{Y}),$$

$$\pi(b_1|k, \theta, \lambda, b_2, \mathbf{Y}),$$

$$\pi(b_2|k, \theta, \lambda, b_1, \mathbf{Y}).$$

(c) Для генерации реализаций из обратного гамма-распределения можно воспользоваться тем, что если с. в. $\xi \sim \mathrm{Gamma}(\alpha, \beta)$, то $1/\xi \sim \mathrm{IG}(\alpha, \beta^{-1})$.

(d) Для генерации k из распределения $\pi(k|\theta, \lambda, b_1, b_2, \mathbf{Y})$ можно выполнить одну итерацию алгоритма Метрополиса-Гастингса со вспомогательным распределением $q(k'|k) \sim \mathrm{Uniform}(2, ..., n-1)$.

2. Модель LDA (Latent Dirichlet Allocation) задается следующим образом:

$$p(\mathbf{W}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{\Phi}|\alpha, \beta) = \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\beta) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N_d} p(w_{d,n}|z_{d,n}, \mathbf{\Phi}) p(z_{d,n}|\boldsymbol{\theta}_d),$$
$$p(\boldsymbol{\phi}_t|\beta) = \mathrm{Dir}(\boldsymbol{\phi}_t|\beta), \quad p(\boldsymbol{\theta}_d|\alpha) = \mathrm{Dir}(\boldsymbol{\theta}_d|\alpha),$$
$$p(w_{d,n}|z_{d,n}, \mathbf{\Phi}) = \mathbf{\Phi}_{z_{d,n}, w_{d,n}}, \quad p(z_{d,n}|\boldsymbol{\theta}_d) = \mathbf{\Theta}_{d, z_{d,n}},$$

где $\mathrm{Dir}(\cdot|\gamma)$ означает распределение Дирихле. Требуется реализовать схему Гиббса для маргинального распределения $p(\mathbf{Z}|\mathbf{W}, \alpha, \beta)$ (так называемый collapsed Gibbs sampling, см. [1, 2, 3]).

В файлах 'test1.dat' и 'test2.dat' записаны данные в виде таблицы: первый столбец – номер документа, второй столбец – номер слова из словаря, третий столбец – сколько раз текущее слово встречается в данном документе. Для первого тестового примера задать следующие значения параметров: $T = 3$; $\alpha = 1$; $\beta = 1$; для второго примера: $T = 20$; $\alpha = 0.1$; $\beta = 0.1$.

3. Рассматривается модель анализа клеточной структуры некоторого участка биоткани, который представляет собой прямоугольную область с равномерным разбиением на ячейки. Каждой ячейке ставится в соответствие вершина графа $G = (V, \mathcal{E})$, ребра которого определяют систему соседства. Предполагается стандартная прямоугольная система соседства, в которой каждая клетка за исключением расположенных на границе области имеет четыре соседних. Каждой вершине графа $i \in V$ соответствует случайная величина X_i, принимающая значения из множества $\{1, 2, 3\}$ (значения соответствуют трем типам клеток:

стромальным, плазматическим и раковым). Рассматривается случайный вектор $X = (X_i,\, i \in V)$, совместное распределение которого задается согласно клеточной модели Поттса:

$$P(x) = \frac{1}{Z} \exp\left(\sum_{(i,j)\in\mathcal{E}} \beta(x_i, x_j) \right), \quad x \in \{1,2,3\}^{|V|}, \tag{1}$$

где Z – нормировочная константа, $\beta = ||\beta(i,j)||_{i,j=1,2,3}$ – заданная матрица коэффициентов взаимодействия между различными типами клеток.

Требуется реализовать процедуру генерации реализаций случайного вектора $X = (X_i,\, i \in V)$ с помощью методов MCMC.

Литература:

[1] URL: Лекция «Латентное размещение Дирихле (LDA)» http://www.machinelearning.ru/wiki/images/8/82/BMMO11_14.pdf

[2] G. Heinrich. Parameter estimation for text analysis. Tech. report, 2005. http://www.arbylon.net/publications/text-est2.pdf

[3] T. L. Griffiths, M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 2004, 101 (suppl 1) 5228-5235. https://www.pnas.org/content/pnas/101/suppl_1/5228.full.pdf