Лабораторная работа №3

1. Рассматривается обратимая химическая реакция

$$A \rightleftarrows B$$

Обозначим через $C_A(t), C_B(t)$ – концентрации реагентов A и B в момент времени t. Пусть динамика изменения концентрации описывается следующей системой дифференциальных уравнений

$$\begin{aligned} \frac{dC_A}{dt} &= -k_1 C_A + k_2 C_B \\ \frac{dC_B}{dt} &= k_1 C_A - k_2 C_B \end{aligned}$$

с начальными условиями (предполагаются известными)

$$C_A(0) = C_{A,0}, \qquad C_B(0) = C_{B,0}.$$

В моменты времени $0 < t_1 < t_2 < \cdots < t_n$ производится замер концентрации вещества A: $C_A(t_j),\ j = 1, ..., n$. Требуется найти оценки неизвестных параметров k_1, k_2 в рамках следующей модели наблюдения:

$$y_j = C_A(t_j) + e_j, \qquad e_j \sim \mathcal{N}(0, \sigma^2).$$

В файле data.txt записаны результаты измерения концентрации вещества A на отрезке времени $[0, 3]$, при этом частота дискретизации составляет 1 кГц. Расчеты произвести для следующих значений параметров: $C_{A,0} = 10$; $C_{B,0} = 15$; $\sigma = 0,2$.

2. Рассматривается латентное размещение Дирихле (LDA – Latent Dirichlet Allocation) – вероятностная модель порождения текста, предназначенная для описания текстов с точки зрения их тематик. При этом тема рассматривается как некоторое распределение вероятностей в пространстве слов из общего словаря [1]. Используемые в дальнейшем обозначения приведены в табл. 1.

Модель LDA задается следующим образом:

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

Таблица 1: Основные обозначения

$w \in \{1, ..., W\}$	– номер слова в словаре
$t \in \{1, ..., T\}$	– номер темы
N_d	– число слов в документе d
$\mathbf{w}_d = [w_{d,1}, ..., w_{d,N_d}]$	– слова в документе d, $w_{d,n} \in \{1, ..., W\}$
$\mathbf{z}_d = [z_{d,1}, ..., z_{d,N_d}]$	– темы документа d, $z_{d,n} \in \{1, ..., T\}$
$\boldsymbol{\theta}_d = [\theta_{d,1}, ..., \theta_{d,T}]$	– вероятности тем в документе d
$\boldsymbol{\phi}_t = [\phi_{t,1}, ..., \phi_{t,W}]$	– вероятности слов в теме t
$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_d]^T \in \mathbb{R}^{D \times T}$	– вероятности тем во всех документах
$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_T]^T \in \mathbb{R}^{T \times W}$	– вероятности слов во всех темах
$\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_D\}$	– набор всех слов в корпусе
$\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_D\}$	– разбиение всех слов по темам

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 анизотропный вариант модели Изинга на прямоугольной решетке с системой соседства первого рода, для которой распределение конфигурации $\mathbf{x} = (x_1, ..., x_n)$ (n – число узлов; $x_i \in \{-1, +1\}$) задается следующим соотношением

$$\pi(\mathbf{x}) = \frac{1}{Z} \exp(-U(\mathbf{x})),$$

где Z – нормировочная константа (статистическая сумма), а энергия определяется следующим выражением

$$U(\mathbf{x}) = -\beta_1 \sum_{\boxed{\text{i}}\boxed{\text{j}}} x_i x_j - \beta_2 \sum_{\substack{\boxed{\text{i}} \\ \boxed{\text{j}}}} x_i x_j.$$

Требуется реализовать процедуру генерации конфигураций модели 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