Экзамен по курсу «Байесовский выбор моделей»

Время выполнения: 90 минут
Максимальный балл: 150 баллов, поэтому можно выполнять не все задания
Замечание: попавшие в TOP3 по баллам дополнительно получат $\min(X,\, 100 - 30R)$, где X – набранный балл, а R – ранг от 0 до 2.

Задача 1 (10 баллов). а) Что такое дивергенция Йенсена-Шеннона (Jensen-Shannon divergence), что она показывает и когда определена? (2 балла)
б) Докажите, что значение дивергенции Йенсена-Шеннона неотрицательно (3 балла).
в) Пусть у Вас есть две модели логистической регрессии с равномерным априорным псевдораспредлением на параметр $\mathbf{w}$, оцененные на двух разных выборках $(\mathbf{X}_1,\, \mathbf{y}_1)$ и $(\mathbf{X}_2,\, \mathbf{y}_2)$ с одинаковым набором из двух признаков.

Пусть апостериорные распределение для первой выборки $\mathbf{w} \sim \mathcal{N}\left(\mathbf{w} \mid [1,\, 1]^\intercal,\, \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}\right)$, а для второй выборки $\mathbf{w} \sim \mathcal{N}\left(\mathbf{w} \mid [-8,\, -3]^\intercal,\, \begin{pmatrix} 7000 & 0 \\ 0 & 50000 \end{pmatrix}\right)$. Считая, что выборки сгенерированы с помощью модели логистической регрессии, можно ли с уверенностью утверждать, что истинные векторы параметров этих моделей $\mathbf{w}_1$ и $\mathbf{w}_2$ разные? (5 баллов)

Задача 2 (60 баллов). Пусть имеется НОР (i.i.d.) выборка $x_1,\, \ldots,\, x_n$ из смеси нормальных распределений с одинаковым мат. ожиданием и разными дисперсиями, то есть $x_i \sim \sum_{i=1}^{K} \pi_i \mathcal{N}(m,\, \sigma_i^2)$, где K – фиксированная постоянная. Введем априорные распределения на m и σ_i^2 вида

$$m \sim \mathcal{N}(m_0,\, \sigma_0^2),\ \tfrac{1}{\sigma_i^2} \sim \Gamma\left(\alpha_i,\, \beta_i\right),$$

где $\alpha_i,\, \beta_i,\, m_0,\, \sigma_0^2$ – известные гиперпараметры.
а) Выписать совместное правдоподобие модели $p(\mathbf{x},\, m,\, \boldsymbol{\sigma}^2|\boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0,\, \boldsymbol{\pi})$ (2 балла);
б) Выписать $p(m,\, \boldsymbol{\sigma}^2|\mathbf{x},\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0,\, \boldsymbol{\pi})$ с точностью до постоянной. Принадлежит ли оно известному параметрическому семейству (2 балла)?
в) Ввести матрицу скрытых переменную $\mathbf{Z}$ с $z_{jk} = 1$ дающим принадлежность объекта j к компоненте смеси k. Выписать совместное правдоподобие модели $p(\mathbf{x}, m, \boldsymbol{\sigma}^2, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}, m_0, \sigma_0, \boldsymbol{\pi})$ со скрытой переменной (3 балла);
г) Получить вариационное приближение $q(m,\, \boldsymbol{\sigma}^2,\, \mathbf{Z}) = q(m)q(\mathbf{Z})q(\boldsymbol{\sigma}^2)$ для полного апостериорного распределения $p(m,\, \boldsymbol{\sigma}^2,\, \mathbf{Z}|\mathbf{x},\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0,\, \boldsymbol{\pi})$ при известном векторе весов компонент в смеси $\boldsymbol{\pi}$ (13 баллов). Что мешает успешно применить обычный ЕМ-алгоритм (2 балла)?
д) Найти наиболее обоснованные значения весов компонент смеси $\boldsymbol{\pi}$ и гиперпараметров $\boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0$, решив задачу (30 баллов)

$$p(\mathbf{x}|\boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0,\, \boldsymbol{\pi}) \to \max_{\boldsymbol{\alpha},\, \boldsymbol{\beta},\, m_0,\, \sigma_0,\, \boldsymbol{\pi}}.$$

Дает ли полученный $\boldsymbol{\pi}^*$ модель с наибольшей апостериорной плотностью вероятности (1 балл)? Как можно проинтерпретировать полученное выражение для $\boldsymbol{\pi}$ (2 балла)? Что нужно изменить в модели, чтобы «поощрить разреженность», то есть исключение избыточных компонент из смеси (5 баллов)?

Замечание: В формулировке выше векторы $\boldsymbol{\sigma},\, \boldsymbol{\alpha},\, \boldsymbol{\beta}$ комбинируют соответствующие значения по компонентам, например, $\boldsymbol{\sigma} = [\sigma_1,\, \ldots,\, \sigma_K]^\intercal$.

Задача 3 (10 баллов). Что такое независимость и условная независимость пары случайных величин? (3 балла) Верно ли, что из независимости следует условная независимость и почему? (3 балла) Следует ли из условной независимости независимость и почему? (4 балла)

Задача 4 (15 баллов). Какие вероятностные модели называются ориентированными, а какие неориентрованными? (3 балла). Для следующей ориентированной графической модели привести соответствующую ей неориентированную (12 баллов).

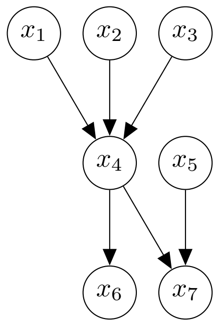


Рис. 1.

Задача 5 (10 баллов). Что такое «марковское одеяло» и какой вид оно имеет для ориентированных (7 баллов) и неориентированных (3 балла) графических моделей и почему?

Задача 6 (10 баллов). Как связаны множества правдоподобий, описываемых без потерь набора условных независимостей для ориентированных и неориентированных графических моделей и почему?

Задача 7 (15 баллов). Пусть моделируется заболеваемость простудой для трех человек: Боба (b), его соседа (a) и коллеги (c). Метка 0 соответствует тому, что человек здоров, а 1 – болезни.

$$p(a,\ b,\ c) = \frac{1}{Z}\psi_{ab}(a,\ b)\psi_{bc}(b,\ c),\ a,\ b,\ c \in \{0,\ 1\}.$$

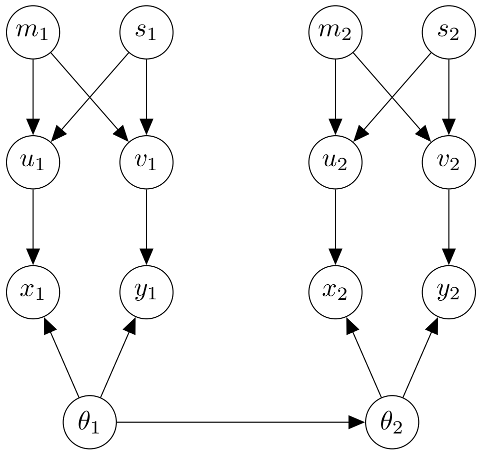
$$\psi_{a,\,b}(a,\ b) = \begin{cases} 20,\ a=0,\ b=0, \\ 4,\ a=1,\ b=1, \\ 2,\ a=0,\ b=1, \\ 1,\ a=1,\ b=0. \end{cases},\ \psi_{b,\,c}(b,\ c) = \begin{cases} 5,\ b=0,\ c=0, \\ 0.7,\ b=1,\ c=1, \\ 5,\ b=0,\ c=1, \\ 0.1,\ b=1,\ c=0. \end{cases}.$$

Какие наборы условных независимостей есть в такой вероятностной модели (3 балла)? Вывести следующие вероятности:

- $p(a=1 \mid b=0,\ c=1)$ (3 балла);
- $p(c=1 \mid b=1,\ a=0)$ (3 балла);
- $p(c=0 \mid a=1,\ b=1)$ (3 балла);
- $p(c=0,\ a=1,\ b=1)$ (3 балла);

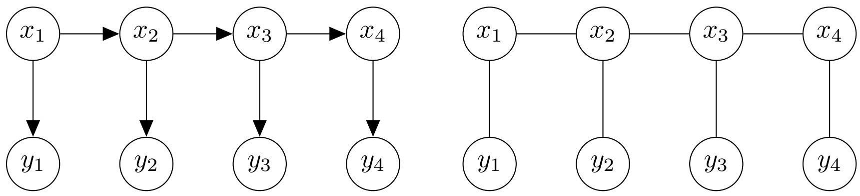
Задача 8 (20 баллов). Верны ли следующие утверждения касательно приведенной графической модели?

- $x_1 \perp x_2$ (3 балла);
- $p(x_1, y_1, \theta_1, u_1) \propto \varphi_1(x_1, \theta_1, u_1)\varphi_2(y_1, \theta_1, u_1)$ для некоторых неотрицательных функций $\varphi_1(.)$, $\varphi_2(.)$ (7 баллов);
- $v_2 \perp \{u_1, v_1, u_2, x_2\} \mid \{m_2, s_2, y_2, \theta_2\}$ (5 баллов);
- $\mathbb{E}[m_2|m_1] = \mathbb{E}[m_2]$ (5 баллов).



Задача 9 (10 баллов + 10 баллов). Задают ли приведенные графические модели одно и то же семейсто распределений и почему?

а)



б)