Зачётный тест 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 (20 баллов). Пусть имеется НОР (i.i.d.) выборка $x_1,\, \ldots,\, x_n$ из нормального распределения $\mathcal{N}(m,\, \sigma^2)$, то есть $x_i \sim \mathcal{N}(m,\, \sigma^2)$. Введем априорные распределения на m и σ^2 вида

$$m \sim \mathcal{N}(m_0,\, \sigma_0^2),\ \tfrac{1}{\sigma^2} \sim \Gamma\left(\alpha,\, \beta\right), \tag{1}$$

где $\alpha,\, \beta,\, m_0,\, \sigma_0^2$ – известные гиперпараметры.
а) Выписать совместное правдоподобие модели $p(\mathbf{x},\, m,\, \sigma^2|\alpha,\, \beta,\, m_0,\, \sigma_0)$ (1 балл);
б) Выписать $p(m,\, \sigma^2|\mathbf{x},\, \alpha,\, \beta,\, m_0,\, \sigma_0)$. Принадлежит ли оно параметрическому семейству нормально-гамма распределений (normal-gamma distributions) и почему? (3 балла);
в) Является ли исходное априорное распределение (2) сопряженным к правдоподобию и почему? (1 балл) Найти и обосновать параметрическое семейство сопряженных распределений (2 балла);
г) Получить вариационное приближение $q(m,\, \sigma^2) = q(m)q(\sigma^2)$ для полного апостериорного распределения $p(m,\, \sigma^2|\mathbf{x},\, \alpha,\, \beta,\, m_0,\, \sigma_0)$ (13 баллов).

Задача 2 (20 баллов). Пусть рассматривается поток посетителей в магазине, и измеряются интервалы между приходом двух последовательных посетителей. Считаем, что интервалы $t_1,\, \ldots,\, t_k,\, \ldots$ между последовательными посетителями независимы в совокупности имеют показательное распределение с параметром λ, то есть

$$p(t_k) = \lambda \exp(-\lambda t_k), t_k \geq 0.$$

а) Выписать правдоподобие модели $p(\boldsymbol{t}|\lambda)$ (2 балла). Ввести априорное распределение на λ $p(\lambda|\boldsymbol{\alpha})$, сопряженное с правдоподобием, где $\boldsymbol{\alpha}$ – вектор гиперпараметров (3 балла). Какое семейство распределений сопряжено с таким правдоподобием? (1 балл)
б) Получить выражения для обоснованности модели $p(\boldsymbol{t}|\boldsymbol{\alpha})$ в явном виде и описать метод поиска $\boldsymbol{\alpha}$ из принципа максимума обоснованности (14 баллов).

Задача 3 (40 баллов). Пусть имеется K математических кружков, в каждом из которых обучается N студентов. Пусть известны результаты решения одинаковых по сложности задач студентами кружков. Там для студента с номером n в кружке с номером k известны два числа: t_{kn} – количество задач, которые студент попробовал решить и s_{kn} – количество успешно решенных задач. В каждом кружке считаем, что вероятность каждого студента решить задачу равна p_k, зависящая от кружка, но не зависящая от студента. Успешность решения разных задач одним студентом, а также успешность решения задач между студентами независимы в совокупности. Считаем, что априорное распределение на p_k есть $p_k \sim \mathrm{Beta}(\alpha,\, \beta)$, где $\alpha,\, \beta$ – неизвестные гиперпараметры.

Обозначим $\mathbf{p} = [p_k,\, k = 1,\, \ldots,\, K]^\mathsf{T}$, $\mathbf{T} = \|t_{kn}\|$, $\mathbf{S} = \|s_{kn}\|$, $k = 1,\, \ldots,\, K$, $n = 1,\, \ldots,\, N$.
а) Выписать в явном виде совместное правдоподобие $p(\mathbf{S},\, \mathbf{p}|\mathbf{T},\, \alpha,\, \beta)$ (5 баллов);
б) Получить апостериорное распределение $p(\mathbf{p}|\mathbf{T},\, \mathbf{S},\, \alpha,\, \beta)$ и описать структуру зависимостей между компонентами $\mathbf{p}$ (10 баллов). Вычислить $\mathbb{E}\mathbf{p}$ по апостериорному распределению (3 балла). Какие выводы можно сделать из полученного результата? (2 балла)
в) Выписать обоснованность $p(\mathbf{S},\, \mathbf{T}|\alpha,\, \beta)$ в явном виде. Описать, как найти оценки гиперпараметров $\alpha,\, \beta$ из принципа максимума обоснованности (20 баллов).

Задача 4 (40 баллов). Пусть имеется обучающая выборка $(\mathbf{X},\, \mathbf{y})$, $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{y} \in$

$[-1,\ 1]^m$, полученная из модели генерации данных с совместным правдоподобием

$$p(\mathbf{y},\ \mathbf{w}|\mathbf{A},\ \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}) \prod_j p(y_j|\mathbf{x}_j,\ \mathbf{w}),$$

где $p(y_j|\mathbf{x}_j,\ \mathbf{w})$ дается моделью логистической регрессии, то есть

$$\mathbb{P}(y_j = 1) = \frac{1}{1 + \exp(-\mathbf{w}^\intercal \mathbf{x}_j)} = \sigma(\mathbf{w}^\intercal \mathbf{x}_j),\ \mathbf{A} = \mathrm{diag}(\alpha_j).$$

Используем принцип максимума обоснованности для отбора признаков

$$\mathbf{A} = \arg\max_{\mathbf{A}} p(\mathbf{y}_1|\mathbf{X}_1,\ \mathbf{A}).$$

а) Выписать совместное правдоподобие модели $p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A})$ в явном виде (1 балл);
б) Какое значение α_j соответсвует тому, что признак j незначим, и не используется в модели? (1 балл)
в) Использовать вариационную нижнюю оценку для сигмоидной функции (см. лекцию 7) для получения нижней оценки на совместное правдоподобие (требуется привести подробный вывод) $p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) \geq L(\mathbf{w},\ \mathbf{A},\ \boldsymbol{\xi})$, где $\boldsymbol{\xi} \in \mathbb{R}^m$ – вектор дополнительных переменных (8 баллов);
г) Для нижней оценки на обоснованность

$$p(\mathbf{y}|\mathbf{X},\ \mathbf{A}) = \int p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) d\mathbf{w} \geq \int L(\mathbf{w},\ \mathbf{A},\ \boldsymbol{\xi}) d\mathbf{w} = \tilde{L}(\mathbf{A},\ \boldsymbol{\xi})$$

получить формулы ЕМ-алгоритма для решения задачи ее максимизации

$$\tilde{L}(\mathbf{A},\ \boldsymbol{\xi}) \to \max_{\mathbf{A},\ \boldsymbol{\xi}} \text{ (30 баллов)}.$$

Задача 5 (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

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

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, $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_K]^\mathsf{T}$.

Задача 6 (70 баллов). Пусть рассматривается поток посетителей в магазине, и измеряются интервалы между приходом двух последовательных посетителей. Считаем, что распределение интервала t_n между последовательными посетителями описывается смесью K показательных распределений с показателями $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^\mathsf{T}$ и весами $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^\mathsf{T}$, то есть

$$p(t_n) = \sum_{k=1}^{K} \pi_k \lambda_k \exp(-\lambda_k t_n),$$

причем наблюденные интервалы $\boldsymbol{t} = [t_1, \ldots, t_N]^\mathsf{T}$ независимы в совокупности.
Введем на $\boldsymbol{\pi}$ априорное симметричное распределение Дирихле $\boldsymbol{\pi} \sim \mathrm{Dir}(\mu \mathbf{e})$, где $\mu < 1$ – параметр распределеления Дирихле (считается фиксированным и известным), а $\mathbf{e}$ – единичный вектор.
Введем также априорные распределения на λ_k вида

$$\lambda_k \sim \Gamma(\alpha, \beta), \ k = 1, \ldots, K,$$

где α, β – неизвестные параметры гамма-распределения.
а) Выписать совместное правдоподобие модели $p(\boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{\lambda}|\alpha, \beta, \mu)$ (5 баллов);
б) Ввести матрицу скрытых переменных $\mathbf{Z}$ принадлежности объекта компоненте смеси и выписать совместное правдоподобие модели со скрытой переменной $p(\boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \mathbf{Z}|\alpha, \beta, \mu)$ (5 баллов);
в) Воспользовавшись вариационным ЕМ-алгоритмом, получить аппроксимацию $q(\boldsymbol{\pi}, \boldsymbol{\lambda}, \mathbf{Z}) = q(\boldsymbol{\pi})q(\boldsymbol{\lambda})q(\mathbf{Z})$ для истинного апостериорного распределения $p(\boldsymbol{\pi}, \boldsymbol{\lambda}, \mathbf{Z}|\boldsymbol{t}, \mu, \alpha, \beta)$ (35 баллов), а также задачи для нахождения оценок максимума обоснованности для α, β

$$\alpha^*, \ \beta^* = \arg\max_{\alpha, \beta} p(\boldsymbol{t}|\mu, \ \alpha, \ \beta)$$

и описать способ решения полученных оптимизационных задач (25 баллов).