Практическое задание 2 по курсу «Байесовский выбор моделей»

Общая информация

- Время сдачи задания: 9е декабря, 21:00 по Москве и это **жесткий дедлайн!!!**;
- Максимальная базовая оценка за задание 100 баллов, так что при желании можно выполнять не всё;
- Оценка автора наилучшей работы удваивается (с учетом баллов сверх 100), но не более, чем до 250 баллов;
- Вопросы и само задание принимаются по почте: aduenko1@gmail.com;
- Тема письма: вопрос по практическому заданию #2 или решение практического задания #2;
- Задания опоздавших **не принимаются** (в силу необходимости проверки заданий к зачету 5го декабря).

Задача (байесовская смесь моделей линейной регрессии). Пусть имеется K поставщиков одного товара, например, новой модели Xphone. У каждого поставщика с номером k есть базовая отпускная цена на этот товар p_k. Магазины в разных городах страны покупают этот товар у одного из поставщиков, причем вероятность выбора поставщика k есть π_k, $\boldsymbol{\pi} = [\pi_k,\ k = 1, \ldots, K]^\mathsf{T}$. Цена продажи поставщика отличается в зависимости от города, для которого магазин покупает товар и меняется с учетом следующих факторов:

- Уровня конкуренции;
- Покупательской способности;
- Уровня арендных ставок и т.д.

За каждую единицу проданного товара магазин получает от производителя фиксированную премию, но магазин на свое усмотрение может установить цену как ниже, так и выше, чем цена покупки у поставщика. Итоговая цена в магазине (которую мы наблюдаем) дается следующей моделью

$$P_i = p_{k_i} + \mathbf{v}_{k_i}^\mathsf{T}\mathbf{x}_i + \varepsilon_i = \mathbf{w}_{k_i}\mathbf{x}_i + \varepsilon_i,$$

где p_{k_i} – базовая цена выбранного поставщика, $\mathbf{x}_i$ – признаковое описание района, где размещен магазин, $\mathbf{w}_{k_i}$ – веса признаков в признаковом описании района, где размещен магазин, для выбранного поставщика, включая константный признак для учета базовой цены p_{k_i}, а ε_i – поправка к цене, устанавливаемая магазином. Считаем, что поправка для каждого магазина выбирается независимо от других магазинов, а также от того, какой поставщик товара был выбран, и в каком районе находится магазин.

Пусть имеется выборка $(\mathbf{X},\ \mathbf{p}) = (\mathbf{x}_i,\ P_i)$, $i = 1, \ldots, m$ описаний разных магазинов, а также информация о цене на Xphone в них. Пусть K – оценка сверху на общее количество поставщиков, которое неизвестно (например, $K = 100$). В качестве априорного распределения на $\boldsymbol{\pi}$ введем распределение Дирихле $p(\boldsymbol{\pi}|\mu) = \mathrm{Dir}(\boldsymbol{\pi}|\mu\mathbf{e})$, где $\mu < 1$ для поощрения разреженности (например, $\mu = 10^{-6}$). На $\mathbf{w}_k$ введем априорное нормальное распределение $\mathbf{w}_k \sim N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})$, $k = \overline{1,\ K}$, где $\mathbf{A}_k$ – диагональная матрица. Шум (поправки к цене, устанавливаемые магазином) считаем нормальным, то есть $\varepsilon_i \sim N(\varepsilon_i|0,\ \beta^{-1})$.

- Выписать совместное правдоподобие $p(\mathbf{p}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \beta, \mu)$ описанной модели в явном виде (4 балла);

- Выписать апостериорное распределение $p(\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{p},\ \mathbf{A},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \beta,\ \mu)$ с точностью до мультипликативной константы и качественно описать, почему не удается указать параметрический вид для этого распределения (4 балла);

- Ввести матрицу скрытых переменных $\mathbf{Z} = \|z_{ik}\|$, где $z_{ik} = 1$, если для i-го магазина выбран поставщик k, и выписать совместное правдоподобие модели со скрытой переменной $p(\mathbf{p},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \beta,\ \mu)$ (4 балла);

- Используя вариационное приближение $q(\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}) = q(\mathbf{Z})q(\boldsymbol{\pi})q(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K)$ для апостериорного распределения $p(\boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{p}, \mathbf{A}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \beta, \mu)$ получить $q(\mathbf{Z})$, $q(\boldsymbol{\pi})$, $q(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K)$ в явном виде с формулами для параметров распределений (35 баллов);

- Используя принцип максимума обоснованности, провести оптимизацию гиперпараметров в смеси моделей путем решения задачи

  $$p(\mathbf{p}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \beta,\ \mu) \to \max_{\mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \beta}$$

  с помощью вариационного EM-алгоритма, считая μ известным и фиксированым (30 баллов). Описать при каких условиях признак j исключается из k-й модели в смеси, то есть происходит отбор признаков (3 балла);

- После выполнения предыдущих пунктов, сообщить об этом на aduenko1@gmail.com и получить индивидуальную выборку данных для анализа (а также описания формата входных и выходных данных). Выборка включает в себя обучающую совокупность $(\mathbf{X}_{\text{train}},\ \mathbf{p}_{\text{train}})$, а также признаковое описание тестовой совокупности $\mathbf{X}_{\text{test}}$. Требуется построить прогноз $\hat{\mathbf{p}}_{\text{test}}$, а также вектор неуверенности в прогнозе $\mathbf{u}_{\text{test}}$ оптимальные в следующем смысле

  $$-\tfrac{1}{2}\sum_{i=1}^{m_2} \log u_i^2 - \tfrac{1}{2u_i^2}(P_i - \hat{P}_i)^2 \to \max,$$

  где суммирование производится по объектам тестовой выборки (60 баллов).