ПСМО ФКН ВШЭ, 3 курс, 2 модуль

Задание 4. Оценивание корреляционной структуры. Матричные разложения

Прикладная статистика в машинном обучении, осень 2018

Время выдачи задания: 8 декабря (суббота).

Срок сдачи: **20 декабря (четверг), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x/PYTHON 3.x.

Правила сдачи

Инструкция по отправке:

1. Решения задач следует присылать единым файлом формата `pdf`, набранным в LaTeX, либо в составе ipython-тетрадки в форматах `ipynb` и `html` (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит тетрадки в формате `ipynb` – а если мы не увидим ваши задачи, мы их не проверим). Отправляйте практические задачи в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

Оценивание и штрафы:

1. Максимально допустимая оценка за работу над основными задачами – 10 баллов.

2. Бонусные баллы (см. конец домашнего задания) и влияют на освобождение от задач на экзамене.

3. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

4. Задание выполняется каждым студентом индивидуально и независимо от других студентов. «Похожие» решения считаются плагиатом и все студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов, причем обнуляются и бонусные баллы. Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

Бонусные задачи

1. (1 балл) Пусть $\boldsymbol{\xi} = (\xi_1, \xi_2)^\intercal$ – гауссовский случайный вектор с математическим ожиданием $\mathrm{E}\,\boldsymbol{\xi} = \boldsymbol{\mu} = (\mu_1, \mu_2)^\intercal$ и ковариационной матрицей
$$\mathrm{E}[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^\intercal] = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$
Подсчитать явное аналитическое выражение для условной плотности $f_{\xi_2|\xi_1}(x_2|x_1)$ распределения случайного вектора ξ_2 при условии $\xi_1 = x_1$.

2. (1 балл) Докажите равенство (trace trick):
$$\boldsymbol{x}^\intercal \boldsymbol{A} \boldsymbol{x} = \mathrm{tr}(\boldsymbol{A} \boldsymbol{x} \boldsymbol{x}^\intercal),$$
где $\boldsymbol{x} \in \mathbb{R}^n$ – вектор, а $\boldsymbol{A} \in (\mathbb{R}^n \to \mathbb{R}^n)$ – квадратная матрица.

3. (1 балл) Докажите, что выборочная ковариационная матрица и выборочное среднее:
$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{\ell} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\intercal,$$
являются оценками максимального правдоподобия для параметров многомерного нормального распределения $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ по выборке $\mathbf{X}^\ell = \{\boldsymbol{x}_i\}_{i=1}^{\ell}$. *Подсказка:* используйте предыдущий пункт про trace trick.

4. (1 балл) Докажите, что если $\widehat{\boldsymbol{\Sigma}} \in (\mathbb{R}^n \to \mathbb{R}^n)$ – оценка ковариационной матрицы $\boldsymbol{\Sigma}$ многомерного нормального распределения $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, построенная по выборке $\mathbf{X}^\ell = \{\boldsymbol{x}_i\}_{i=1}^{\ell}, \boldsymbol{x}_i \in \mathbb{R}^n$:
$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{\ell} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\intercal,$$
и $\ell < n$, то она вырождена (т.е. не существует обратной $\widehat{\boldsymbol{\Sigma}}^{-1}$).

5. (2 балла) Провести анализ внутренней размерности выборки методом MLE, провести разложение на главные (Principal CA) и независимые (Independent CA) компоненты (см. семинар 13).

   Скачайте данные MNIST (https://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html#sphx-glr-auto-example

   В качестве отчета приведите:

   (a) (1 балл) Внутреннюю размерность выборки (численно) и методику ее оценки, приложите график.

   (b) (1 балл) Визуализацию главных и независимых компонент, соответствующие оцененной внутренней размерности выборки.