Задачи к лекции и семинару “Метод главных компонент”

1

Сгенерируйте случайную симметричную матрицу A размера 3×3. Сгенерируйте N элементов из нормального распределения $P \propto e^{-\boldsymbol{x}^T A \boldsymbol{x}}$ (получится матрица объект-признак X размерности $N \times 3$). Визуализируйте полученное облако точек (для построения интерактивных трехмерных графиков можно воспользоваться пакетом `ipympl` в системе jupyter). Примените к матрице X метод главных компонент, визуализируйте сингулярные вектора вместе с облаком точек, а также двумерные проекции элементов выборки на плоскости, задаваемые сингулярными векторами.

2

Пусть X — матрица объект-признак (размерность $l \times F$), для которой сингулярное разложение имеет вид $X = V\sqrt{\Lambda}U^T$. После понижения размерности данных с помощью метода главных компонент, в диагональной матрице $\Lambda = \mathrm{diag}\{\lambda_1 \geq \cdots \geq \lambda_F\}$ оставляются только $\tilde{F}$ наибольших сингулярных чисел: $\tilde{\Lambda} = \mathrm{diag}\{\lambda_1 \geq \cdots \geq \lambda_{\tilde{F}}\}$. При этом данные, как правило, можно восстановить только с некоторой ошибкой: $\tilde{X} = V\sqrt{\tilde{\Lambda}}U^T \neq X$. Покажите, что L_2 норма ошибки выражается через сумму по оставшимся сингулярным числам:

$$\frac{1}{l}\left\|X - \tilde{X}\right\|^2 = \sum_{i=\tilde{F}+1}^{F} \lambda_i.$$

3

Покажите, что сингулярный вектор матрицы X, отвечающий наибольшему сингулярному числу, является решением задачи

$$\boldsymbol{u} = \mathrm{argmax}_{||\boldsymbol{u}||=1}(X\boldsymbol{u})^2,$$

где подразумевается матричное умножение X на $\boldsymbol{u}$.

4

Пусть дан набор точек на плоскости (x_i, y_i), для которых выборочные средние x_i и y_i равны нулю. Покажите, что сингулярный вектор для матрицы объект-признак, отвечающий наибольшему сингулярному числу, задает прямую a (проходящую через начало координат), которая является решением следующей задачи оптимизации:

$$L' = \sum_{i=1}^{N} \mathrm{distance}^2[(x_i, y_i); a] \quad \longrightarrow \quad \min_a,$$

где $\mathrm{distance}[(x_i, y_i); a]$ — расстояние от точки (x_i, y_i) до прямой a (равное длине перпендикуляра).

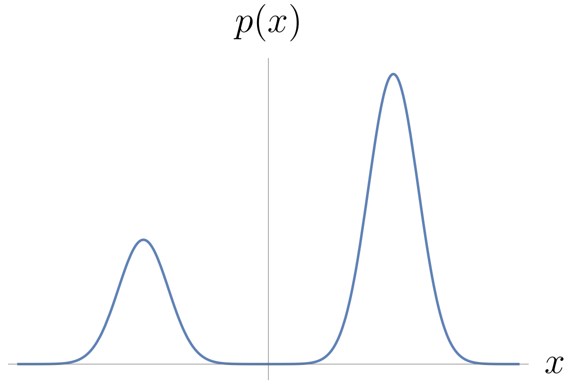
Обратите внимание, что такая задача отличается от задачи МНК, в которой расстояние от точки до аппроксимирующей прямой вычисляется не по перпендикуляру, а вдоль оси y, отвечающей целевой переменной.

5

Пусть дан набор из N точек в трехмерном пространстве $X_{i\alpha}$, $i \in \{1, \ldots, N\}$, $\alpha \in \{1, 2, 3\}$. Покажите, что задача нахождения сингулярных чисел матрицы X эквивалентна нахождению *главных моментов инерции* твердого тела, составленного из набора точечных масс, расположенных в точках (X_{i1}, X_{i2}, X_{i3}) (можно представлять себе, что точечные массы соединены между собой невесомыми и абсолютно жесткими стержнями).

6*

Задача матричного разложения (аппроксимация матрицы произведением двух других матриц меньшего ранга) *с ограничениями* (например, условие положительности элементов) не решается в общем случае с помощью сингулярного разложения. Для решения такой задачи может использоваться EM-алгоритм. Изучим его на примере другой простой модельной задачи.



Пусть дана выборка точек x_i, взятая из *смеси* гауссовых распределений:

$$p(x) = \alpha N_{\mu_1,\sigma_1}(x) + (1-\alpha) N_{\mu_2,\sigma_2}(x).$$

Тогда можно поставить задачу оценки параметров α, μ_1, μ_2,σ_1,σ_2 по выборке $\{x_i\}$.

- Покажите, что задача максимизации обычного правдоподобия $\prod_i p(x_i) \longrightarrow \max_{\alpha,\mu_1,\mu_2}$плохо определена. Какие значения параметров максимизируют такое правдоподобие?

- Сгенерируйте данные (две сгустка точек должны быть хорошо видны при визуализации) и найдите параметры α, μ_1, μ_2,σ_1,σ_2 с помощью EM-алгоритма. Инициализировать параметры можно какими-то случайными значениями.
  EM-алгоритм состоит из двух чередующихся шагов:

  1. M(Maximization)-шаг. Относим каждую точку x_i к первой или второй гауссиане, сравнивая значения правдоподобия для каждой компоненты смеси:

  $$a(x_i) = \begin{cases} 1, & p_1(x_i) > p_2(x_i), \\ 2, & p_2(x_i) > p_1(x_i), \end{cases}$$

  где $p_1(x) = \alpha N_{\mu_1,\sigma_1}(x)$, $p_2(x) = (1-\alpha) N_{\mu_2,\sigma_2}(x)$.

  2. E(Expectation)-шаг. Находим параметры μ_1,σ_1 и μ_2,σ_2, максимизируя правдоподобие (или его логарифм) отдельно по точкам, отнесенным к каждой гауссиане:

  $$\prod_{x_i : a(x_i)=1} p_1(x_i) \longrightarrow \max_{\mu_1,\sigma_1}$$

  $$\prod_{x_i : a(x_i)=2} p_2(x_i) \longrightarrow \max_{\mu_2,\sigma_2}$$

  При нахождении параметра α можно оптимизировать обычное правдоподобие $\prod_i p(x_i)$. Все такие максимизации правдоподобия осуществляются аналитически в общем виде для гауссовых распределений

  Реализуйте EM-алгоритм. Так как метод является итерационным, необходимо выбрать какой-либо критерий остановки, например, прекращать процесс, если относительное изменение каждого параметра при очередном шаге меньше некоторого порога. С какой точностью удалось восстановить α, μ_1, μ_2, σ_1,σ_2?