Метод головних компонент з точки зору методів оптимізації

Н. Фордуй, О. Галганов

Опис задачі

Задано набір $\{x_1, ..., x_m\}$ із m точок в $\mathbb{R}^n$. Необхідно для заданого $k < n$ знайти k-вимірну гіперплощину, яка буде найближчою до цих точок в сенсі евклідової норми: тобто, мінімальною має бути різниця між цими точками та їх проекціями на шукану площину.
Ця задача називається **методом головних компонент** (англ. principal component analysis, PCA) і широко застосовується в статистиці та машинному навчанні.

Постановка задачі

Розглянемо задані $\{x_1, ..., x_m\}$ — m векторів з $\mathbb{R}^n$. Відомо, що k-вимірна гіперплощина H_k в $\mathbb{R}^n$ задається k ортонормованими векторами $\{u_1, ..., u_k\}$, які можна доповнити до базису (далі — ОНБ), та вектором зсуву відносно нуля b:
$H_k = \{x = b + c_1 u_1 + ... + c_k u_k : c_1, ..., c_k \in \mathbb{R}\}$.
Нехай $\{u_i\}_{i=1}^n$ — деякий ОНБ в $\mathbb{R}^n$. Тоді $\forall\ i = 1, ..., n : x_i = b + \sum_{j=1}^{n} c_{i,j} u_j$, де $c_{i,j} = (x_i, u_j)$ (це — розклад Фур'є в $\mathbb{R}^n$).
Покажемо, що після застосування деякого перетворення заданих векторів можна вважати $b = 0$.

Допоміжна задача

Для заданих m точок $\{x_1, ..., x_m\}$ в $\mathbb{R}^n$ знайти точку, яка знаходиться найближче до них в сенсі евклідової норми.

Розв'язання допоміжної задачі

Задача має вигляд $F(x) = \sum\limits_{k=1}^{m} \|x - x_k\|^2 \to \min,\ x \in \mathbb{R}^n$.

$F'_x(x^*) = 2\sum\limits_{k=1}^{m}(x^* - x_k) = 0 \Rightarrow x^* = \frac{1}{m}\sum\limits_{k=1}^{m} x_k$ — стаціонарна точка.

Оскільки $F(x)$ — опукла функція, то $x^* = \frac{1}{m}\sum\limits_{k=1}^{m} x_k$ — розв'язок.

Постановка задачі

Таким чином, якщо замінити x_i на $y_i = x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$, то можна вважати $b = 0$, оскільки тепер найближчою точкою $\mathbb{R}^n$ буде точка 0, тому й шукана гіперплощина теж має проходити через 0.

Отже, $\forall\ i = 1, ..., n : y_i = \sum_{j=1}^{n} c_{i,j}u_j$, де $c_{i,j} = (y_i, u_j)$. Проекції на гіперплощину y_i будемо шукати у вигляді $\hat{y}_i = \sum_{j=1}^{k} c_{i,j}u_j,\ k < n$. Введемо вектори похибок $\varepsilon_i = y_i - \hat{y}_i = \sum_{j=k+1}^{n} c_{i,j}u_j$, які зберемо у матрицю:

$$E = \big(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m\big) = \big(u_{k+1}, u_{k+2}, ..., u_n\big) \cdot \begin{pmatrix} c_{1,k+1} & c_{2,k+1} & \cdots & c_{m,k+1} \\ c_{1,k+2} & c_{2,k+2} & \cdots & c_{m,k+2} \\ \cdots & \cdots & \cdots & \cdots \\ c_{1,n} & c_{2,n} & \cdots & c_{m,n} \end{pmatrix}$$

Коротко: $E = UC$, причому $U^TU = I$, бо ці вектори ортонормовані, а $C = Y^TU$.

Розв'язання задачі

Задача

Знайти такі ортонормовані вектори $\{u_i\}_{i=1}^n$, що $\|E\|^2 \to min$, де $\|E\| = \sqrt{\sum_{i,j=1}^n e_{ij}^2}$ — норма Фробеніуса матриці похибок E.

Введемо позначення $Y = (y_1, ..., y_m)$, $F = YY^T$.
$\|E\|^2 = \mathrm{Tr}\left(E^T E\right) = \mathrm{Tr}\left(C^T U^T U C\right) = \mathrm{Tr}\left(C^T C\right) = \mathrm{Tr}\left(U^T Y Y^T U\right) =$
$= \mathrm{Tr}\left(U^T F U\right)$. Оскільки $U = \sum\limits_{j=k+1}^{n} \left(0, ...0, u_j, 0..., 0\right) = \sum\limits_{j=k+1}^{n} U_j$, за лінійністю Tr (слід матриці, сума діагональних елементів) маємо $\mathrm{Tr}\left(U^T F U\right) = \sum\limits_{j=k+1}^{n} \mathrm{Tr}\left(U_j^T F U_j\right)$. В кожній матриці U_j лише один стовпець не рівний нулю, тому $\sum\limits_{j=k+1}^{n} \mathrm{Tr}\left(U_j^T F U_j\right) = \sum\limits_{j=k+1}^{n} (F u_j, u_j)$.

Таким чином, отримуємо задачу умовної оптимізації.

Розв'язання задачі

$$\begin{cases} F(u_{k+1}, ..., u_n) = \sum\limits_{j=k+1}^{n} (Fu_j, u_j) \to \min \\ \|u_j\|^2 = 1,\ j = k+1, ..., n \\ \{u_{k+1}, ..., u_n\} \text{ — лінійно незалежні} \end{cases}$$

Ця задача є регулярною, бо градієнти функцій, що задають обмеження, є лінійно незалежними за умовою.

Маємо функцію Лагранжа:

$$\mathcal{L}(u_{k+1}, ..., u_n, \lambda_{k+1}, ..., \lambda_n) = \sum_{j=k+1}^{n} \left((Fu_j, u_j) + \lambda_j \left(\|u_j\|^2 - 1 \right) \right)$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial u_j} = 2Fu_j + 2\lambda_j u_j = 0 \\ (j = k+1, ..., n) \\ \|u_j\|^2 = 1,\ j = k+1, ..., n \\ \{u_{k+1}, ..., u_n\} \text{ — лінійно незалежні} \end{cases}$$

Оскільки $2Fu_j + 2\lambda_j u_j = 0 \Leftrightarrow Fu_j = -\lambda_j u_j$, то розв'язком системи будуть u_j — власні вектори F одиничної норми. Але з умови мінімізації цільової функції та лінійної незалежності $\{u_{k+1}, ..., u_n\}$ ці власні вектори мають відповідати найменшим власним числам $\mu_j = -\lambda_j$.

Розв'язання задачі

$F^T = (YY^T)^T = YY^T$, $F \geq 0$, бо $\forall\, x \in \mathbb{R}^n : (Fx, x) = (YY^T x, x) = (Y^T x, Y^T x) \geq 0$. Таким чином, всі $\mu_j = -\lambda_j \geq 0$.

Оскільки цільова функція є нескінченно зростаючою, то отримані u_j є розв'язками задачі. Зрозуміло, що інші складові ОНБ $\{u_i\}_{i=1}^n$ можна покласти рівними іншим власним векторам матриці F, розташувавши всі отримані у порядку спадання відповідних власних значень, причому векторам $u_1, ..., u_k$ відповідатимуть k найбільших власних чисел (нагадаємо, що за цими векторами розкладалися наближення $\hat{y}_i$).

Залишилося згадати, що $y_i = x_i - \frac{1}{m}\sum\limits_{i=1}^{m} x_i$. Позначимо $X = (x_1, ..., x_m)$, тоді $Y = X - \left(\frac{1}{m}\sum\limits_{i=1}^{m} x_i\right) \cdot \underbrace{(1, 1, ..., 1)}_{m}$. Тепер обчислимо матрицю YY^T.

Розв'язання задачі

Введемо позначення $\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$.

$YY^T = (X - \overline{x}\cdot(1,1,...,1))\cdot\left(X^T - (1,1,...,1)^T\cdot\overline{x}^T\right) =$

$= XX^T - \overline{x}\cdot(1,1,...,1)\cdot X^T - X\cdot(1,1,...,1)^T\cdot\overline{x}^T + \overline{x}\cdot(1,1,...,1)\cdot$
$(1,1,...,1)^T\cdot\overline{x}^T = XX^T - m\cdot\overline{x}\cdot\overline{x}^T - m\cdot\overline{x}\cdot\overline{x}^T + m\cdot\overline{x}\cdot\overline{x}^T =$
$XX^T - m\cdot\overline{x}\cdot\overline{x}^T =$

$= XX^T - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)\cdot\left(\sum_{i=1}^{m} x_i^T\right)$

Таким чином, знайдені вектори u_j є власними векторами матриці $XX^T - m\cdot\overline{x}\cdot\overline{x}^T$.

Відповідь

Шуканою k-вимірною гіперплощиною є $\overline{x} + L(u_1, ..., u_k)$.

Тут $\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$, а $L(u_1, ..., u_k)$ — лінійна оболонка власних векторів матриці $XX^T - m \cdot \overline{x} \cdot \overline{x}^T$ ($X = (x_1, ..., x_m)$), які відповідають k найбільшим власним числам. Зазначимо, що при цьому похибка (сума квадратів евклідових норм векторів ε_i) рівна $\sum_{j=k+1}^{n} \mu_j$ — сумі $n - k$ найменших власних чисел цієї матриці.

Для практичного застосування є корисною формула для обчислення проекцій $y_i = x_i - \overline{x}$ на $L(u_1, ..., u_k)$: $\mathrm{pr}(y_i) = (u_1, ..., u_k)^T \cdot y_i$, або в матричному вигляді: $\mathrm{pr}(Y) = (u_1, ..., u_k)^T \cdot Y$

Додаток: диференціювання за векторним аргументом

У розв'язанні задачі було використано похідні скалярної функції з декількома векторними аргументами. Під частинною похідною за векторним аргументом вважається вектор з частинних похідних цієї функції за координатами вектора. Доречно навести виведення використаної формули для похідної від квадратичної форми.

$F(x) = (Ax, x)$, $x \in \mathbb{R}^n$, A — дійсна симетрична $n \times n$ матриця.

$F(x+h) - F(x) = (Ax + Ah, x + h) - (Ax, x) = (Ax, x) + (Ax, h) + (Ah, x) + (Ah, h) - (Ax, x) = \left[A = A^T\right] = (2Ax, h) + (Ah, h)$

Отже, лінійна частина приросту рівна $2Ax$, звідки $F'(x) = 2Ax$. Зокрема, похідна квадрата норми $\|x\|^2$ рівна $2x$.