PCA, Principal Components Analysis, Аналіз Принципових Компонент, АПК

З чим він має справу і для чого це?

В результаті N-кратних спостережень одержано N наборів чисел:

$$RD = \{X_1, \ldots, X_N\}, \tag{1}$$

кожний набір має розмірність M:

$$X_n = \{X_{1n}, X_{2n}, \ldots, X_{Mn}\}.$$

Метод АПК має сенс тільки коли $M > 1$. В цьому випадку дані (1) — це набір точок/векторів в M-вимірному просторі. Кожна точка відповідає одному вимірюванню. Точки не співпадають, оскільки вимірювання виконуються на різних об'єктах. Отже, точки (1) певним чином розсіяні в M-вимірному просторі. Якщо об'єкти, на яких проводились вимірювання, належали до декількох класів, то можна очікувати, що точки (1) утворюють в просторі декілька кластерів так, що кожен кластер відповідає одному з класів.

Перша задача АПК — виявлення окремих класів:
Наперед може не бути відомо, що є декілька класів об'єктів, але кластеризація відповідних точок може виявити цей факт.

Друга задача АПК — виявлення приналежності до окремого класу:
Якщо в попередньому аналізі вже виявлено наявність декількох класів, то дані вимірювання нового об'єкту, X', дозволять віднести цей об'єкт до певного класу, якщо X' потрапляє до відповідного кластеру.

Метод АПК полегшує виявлення кластерів.

Знаходження кластрів в багатовимірному просторі може виявитись складним. Можна проектувати множину точок на підпростори нижчої розмірності і шукати кластери в них, але як вибрати такі підпростори найбільш ефективно? Т.т, так, щоб не зменшити, або зменшити мінімально взаємну віддаленість точок і, тим самим зберегти наявну в просторі R^M кластеризацію також і в просторі меншої розмірності. Метод АПК дозволяє знайти такий ортонормований базис в R^M, що точки RD мають найбільше розсіяння в напрямку першої осі цього базису, а в напрямках наступних осей розсіяння понижується. В такому випадку множина векторів RD подається в координатах знайденого базису. Значення першої координати вектора X_n в цьому базисі називається його першою принциповою компонентою, 2-ї — 2-ю ПК і т.д. Часто врахування перших двох ПК виявляється досить для виявлення кластерів, або встановлення приналежності до них.

Очевидно, принципові компоненти окремого вектора данних залежать від всіх інших векторів данних з набору RD.

Як це робиться?

Перш за все, кластеризація не змінюється, якщо над всіма точками одночасно виконати певне переміщення. Перше таке переміщення, яке поміщає набір точок в окіл початку координат в R^M, виконується шляхом віднімання середнього μ_m від кожної координати з номером m:

$$\mu_m = \frac{1}{N} \sum_{1 \le n \le N} X_{mn}, \quad X_{mn} \to (X_{mn} - \mu_m), \quad m = 1, \ldots, M,\ n = 1, \ldots, N.$$

Одержані таким чином нові дані $RD = \{X_1, \ldots, X_N\}$ називають центрованими. При цьому кожна точка даних X_n в просторі R^M зміщується на один і той самий вектор $\mu = \{\mu_1, \ldots, \mu_M\}$, тому кластерна структура множини точок зберігається.

Далі може бути виконане, або не виконане наступне перетворення:

$$X_{mn} \to \frac{X_{mn}}{\sigma_m}, \tag{2}$$

де σ_m — стандартне відхилення по координаті m:

$$\sigma_m = \sqrt{\frac{\sum\limits_{1 \le n \le N} X_{mn}^2}{N}}. \tag{3}$$

Якщо це перетворення виконано, то АПК називають стандартизованим, а нові величини X_{mn} — z-оцінками (z-scores). Отже, надалі ми вважаємо дані центрованими, над якими, можливо, виконане перетворення (2).

Для (центрованого) набору векторів $\{X_1, \ldots, X_N\}$ можна ввести поняття варіації, як сумми варіацій по кожній з координат:

$$\sigma^2 = \sum_{1 \le m \le M} \sigma_m^2,$$

де σ_m дається виразом (3). Останнє можна переписати в наступному вигляді

$$\sigma^2 = \frac{1}{N} \sum_{1 \le n \le N} ||X_n||^2,$$

де $||X_n||^2 = \sum\limits_{1 \le m \le M} X_{mn}^2$ — квадрат довжини вектора X_n. З останнього зрозуміло, що повна варіація σ^2 залишається незмінною при поданні координат векторів X_n в новій системі координат, яка одержується ортогональним перетворенням з старої.

Якщо позначити $Y_n = \{Y_{1n}, \ldots, Y_{Mn}\}$, $n = 1, \ldots, N$ координати даних спостереження в новій системі координат, то задача відшукання першої принципової компоненти полягає в відшуканні такої системи координат, що сумма

$$\sigma_1^2 = \frac{1}{N} \sum_{1 \le n \le N} Y_{1n}^2 \tag{4}$$

дає найбільший можливий внесок в повну варіацію σ^2. В цьому випадку перша координата векторів Y_n називається їх старшою/першою принциповою компонентою. Для знаходження другої принципової компоненти слід розглянути набір даних на одиницю нижчої розмірності $\{Y_1^1, \ldots, Y_N^1\}$, який одержується з $\{Y_1, \ldots, Y_N\}$ відкиданням першої координати в кожному векторі Y_n і відшукати для нового набору данних старшу ПК так само, як в попередньому випадку. В знайденій цього разу системі координат варіація першої координати, σ_2^2, буде давати певний, додатковий до (4), внесок в повну дисперсію, σ^2, початкових даних. Якщо ставиться на меті, щоб варіація принципових компонент складала не менше ніж $\alpha\sigma^2$, де $0 < \alpha < 1$, то процедура повторюється k разів так, щоб

$$\sum_{1 \le i \le k} \sigma_i^2 \ge \alpha\sigma^2.$$

Після цього в просторі R^k утворюється множина точок $Z = \{Z_1, \ldots, Z_N\}$ координати яких — це знайдені перші k принципових компонент. Множина Z називається представленням даних через перші k принципових компонент.

Техніка знаходження старшої ПК

Для знаходження старшої ПК розглянемо матрицю коваріації данних:

$$C_{km} = \frac{1}{N} \sum_{1 \le n \le N} X_{kn} X_{mn}, \quad k, m = 1, \ldots, M,$$

або

$$C = X \cdot X^*,$$

де X — матриця, стовпці якої — це дані X_n, X^* — матриця, спряжена до X, а точка означає множення матриць. Повна варіація данних, σ^2, одержується додаванням діагональних елементів матриці C.

Матриця C самоспряжена: $C^* = C$, отже, існує ортогональна система координат в якій вона діагоналізується. Матриця C також позитивно напіввизначена: $(x, Cx) \ge 0$ для будь-якого $x \in R^M$. Отже, власні значення C — невід'ємні числа. Їх сумма і буде повною варіацією данних σ^2. При цьому власний вектор матриці C, який відповідає найбільшому власному значенню, визначає в просторі данних R^M координату, яка є старшою принциповою компонентою для множини RD.

Маючи матрицю коваріації можна діяти по різному. Якщо M невелике, то можна знайти повний набір власних значень C: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$ і відповідних нормованих власних векторів $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M$, після чого обмежитись першими k. Якщо це зроблено, то для знаходження i-тої принципової компоненти ПК_i данних кожного вектора X_n слід обчислити проекцію X_n на $\mathbf{e}_i$:

$$\text{ПК}_i = (X_n, \mathbf{e}_i). \tag{5}$$

Якщо M — велике число, а шукається тільки перша ПК, то більш ефективним може виявитись відшукання тільки найбільшого власного значення λ_1 матриці коваріації C і відповідного нормованого власного вектора $\mathbf{e}_1$. Представлення даних через першу ПК одержується за формулою (5). Якщо потрібна ще наступна ПК, то дані $\{X_1, \ldots, X_N\}$ перетворюються в нові дані $\{X'_1, \ldots, X'_N\}$ так, щоб виключити вже знайдену ПК:

$$X'_n = X_n - (X_n, \mathbf{e}_1)\mathbf{e}_1, \quad n = 1, \ldots, N.$$

Одержані дані $\{X'_1, \ldots, X'_N\}$ будть векторами розмірності M, але відповідна їм матриця коваріації C' буде мати своїм найбільшим власним числом друге за величиною число матриці C. Після знаходження його і відповідного власного вектора друга ПК одержується за формулою (5) в якій замість X_n можна підставити X'_n, хоча необов'язково. Процедуру можна повторити необхідне число разів.

Використані тут поняття дисперсії/варіації, напівпозитивно визначеної, спряженої і самоспряженої матриць, скалярного добутку, ортогонального перетворення системи координат, власного значення та власного вектора можна знайти в книгах:

Гнєденко Б.В. Курс теорії імовірностей. Радянська Школа, 1949.

Мальцев А.И. Основы линейной алгебры. ГИТТЛ, 1956.

Белманн Р. Введение в теорию матриц. Наука, 1969.

Калужнін Л.А., Вишенський В.А., Шуб Ц.А. Лінійні простори. Вища Школа, 1971.

За матеріалами Інтернет уклав О.К.Відибіда