Доказательство оптимальности главных компонент (PCA)

Автор:

Пользователь

21 апреля 2025 г.

Доказательство оптимальности главных компонент (PCA)

Пусть имеется матрица наблюдений X размера $n \times m$, где n — число наблюдений, m — число признаков. Ковариационная матрица признаков определяется как:

$$X_{\text{cov}} = \frac{1}{n-1} X_{\text{centered}}^T X_{\text{centered}},$$

это симметричная матрица размера $m \times m$, элементами которой являются ковариации признаков. Важное свойство: X_{cov} положительно полуопределена, а её диагональные элементы равны дисперсиям отдельных признаков. Рассмотрим произвольное направление в пространстве признаков, заданное единичным вектором $\mathbf{w} \in \mathbb{R}^m$ ($|\mathbf{w}| = 1$). Проекция центрированных данных на это направление даётся линейной комбинацией признаков $\mathbf{y} = X_{\text{centered}}\mathbf{w}$ (здесь $\mathbf{y}$ — вектор длины n, содержащий координаты всех наблюдений вдоль $\mathbf{w}$). Дисперсия проекций данных на направление $\mathbf{w}$ вычисляется как средний квадрат отклонения $\mathbf{y}$ от нуля (среднее ноль из-за центрирования):

$$\text{Var}(y) = \frac{1}{n-1} \|X_{\text{centered}}\mathbf{w}\|^2 = \frac{1}{n-1} \mathbf{w}^T (X_{\text{centered}}^T X_{\text{centered}}) \mathbf{w} = \mathbf{w}^T X_{\text{cov}} \mathbf{w}.$$

Таким образом, дисперсию проекции на $\mathbf{w}$ можно выразить квадратичной формой $\mathbf{w}^T X_{\text{cov}} \mathbf{w}$. Задача нахождения направления максимальной дисперсии сводится к максимизации этой квадратичной формы при ограничении $|\mathbf{w}| = 1$.

Одномерный случай

Нужно решить задачу нахождения максимума:

$$\max_{\|\mathbf{w}\|=1} \mathbf{w}^T X_{\text{cov}} \mathbf{w}.$$

Поскольку X_{cov} — симметрическая матрица, то существует ортонормированный базис из собственных векторов $e_1, e_2, \dots, e_m \in \mathbb{R}^m$, в котором X_{cov} диагонализуема. Обозначим соответствующие собственные значения и упорядочим их: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$. Тогда

$$X_{\text{cov}} e_i = \lambda_i e_i, \quad i = 1, \dots, m$$

Вектор $\mathbf{w}$ можно разложить по базису $\{e_i\}$:

$$\mathbf{w} = a_1 e_1 + a_2 e_2 + \dots + a_m e_m,$$

где $e_i^T \mathbf{w} = a_1 e_i^T e_1 + \dots + a_i e_i^T e_i + \dots + a_m e_i^T e_m = a_i$ и $|\mathbf{w}| = \sum_{i=1}^m a_i^2 = 1$. Тогда:

$$\mathbf{w}^T X_{\text{cov}} \mathbf{w} = \sum_{i=1}^m \lambda_i a_i^2.$$

Поскольку λ_1 — наибольшее собственное значение, максимум достигается при $a_1^2 = 1$, а $a_2^2 = \dots = a_m^2 = 0$, то есть при $\mathbf{w} = e_1$. Таким образом, оптимальное направление — это e_1.

Обобщение на k измерений: первые k главных компонент

Пусть теперь требуется выбрать не одно направление, а подпространство размерности k (где $1 \leq k \leq m$), на которое данные будут проецироваться. Мы хотим, чтобы суммарная дисперсия проекций на это k-мерное подпространство была максимальной. Иными словами, нужно выбрать k ортонормированных направлений $\mathbf{w}_1, \ldots, \mathbf{w}_k$ (где $\mathbf{w}_i^T \mathbf{w}_j = 0$ при $i \neq j$ и $|\mathbf{w}_i| = 1$), которые максимизируют сумму дисперсий:

$$\sum_{i=1}^{k} \mathrm{Var}(X_{\text{centered}} \mathbf{w}_i) = \sum_{i=1}^{k} \mathbf{w}_i^T X_{\text{cov}}\, \mathbf{w}_i.$$

Разложим каждый из векторов $\mathbf{w}_i$ по базису собственных векторов $e_1, \ldots, e_m$ ковариационной матрицы. Тогда:

$$\mathbf{w}_i = \sum_{j=1}^{m} a_{ij}\, \mathbf{e}_j, \quad \text{и} \quad \mathbf{w}_i^\top X_{\text{cov}} \mathbf{w}_i = \sum_{j=1}^{m} \lambda_j a_{ij}^2.$$

Суммарная дисперсия по всем k направлениям:

$$\sum_{i=1}^{k} \mathbf{w}_i^\top X_{\text{cov}} \mathbf{w}_i = \sum_{i=1}^{k} \sum_{j=1}^{m} \lambda_j a_{ij}^2 = \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{k} a_{ij}^2 \right).$$

Здесь величина $w_j := \sum_{i=1}^{k} a_{ij}^2$ показывает, какая часть вектора $\mathbf{w}_j$ лежит в нашем k-мерном подпространстве. Из ортонормированности $\mathbf{w}_i$ следует, что для каждого j выполняется $0 \leq w_j \leq 1$, а суммарно $\sum_{j=1}^{m} w_j = k$. Теперь видно, как максимизировать сумму $\sum_{j=1}^{m} \lambda_j w_j$ при данных ограничениях на w_j. Поскольку $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, наибольшую отдачу дают веса, помещённые при первых собственных значениях. Чтобы сумма была максимальной, нужно выбрать $w_1 = w_2 = \cdots = w_k = 1$ (т.е. полностью включить первые k собственных векторов в подпространство), а остальные $w_{k+1}, \ldots, w_m = 0$. В этом случае выполняется ограничение $\sum w_j = k$, и сумма равна $\lambda_1 + \lambda_2 + \cdots + \lambda_k$. Таким образом, направления главных компонент PCA — это собственные векторы ковариационной матрицы данных. Ортогональность собственных векторов гарантирует, что главные компоненты независимы и образуют ортонормированный базис в выбранном подпространстве.