Отчёт по лабораторной работе: Анализ главных компонент (PCA)

Выполнили: Смирнов Олег, Муртазалиев Матвей. Группа J3110
Санкт-Петербург, 2025

1 Цели и задачи

Цель: Реализовать алгоритм PCA и применить его к тестовым данным для анализа влияния шума.

Задачи:

1. Реализовать метод Гаусса для решения систем линейных уравнений (СЛАУ).
2. Реализовать функции центрирования данных и вычисления матрицы ковариаций.
3. Найти собственные значения и векторы матрицы ковариаций.
4. Выполнить проекцию данных и оценить долю объяснённой дисперсии.
5. Исследовать влияние шума на результаты PCA.
6. Визуализировать проекции данных на первые две главные компоненты.

2 Теоретическая часть

PCA основан на следующих этапах:

1. **Центрирование данных**:

$$X_c = X - \bar{X},$$

где X — исходная матрица данных размером $n \times m$, $\bar{X}$ — вектор средних значений признаков.

2. **Вычисление матрицы ковариаций**:

$$\Sigma = \frac{1}{n-1} X_c^T X_c,$$

где X_c — центрированная матрица данных.

3. **Нахождение собственных значений и векторов**: Собственные значения λ_i и векторы v_i матрицы Σ определяют направления главных компонент.

4. **Проекция данных**:

$$X_{\text{proj}} = X_c V_k,$$

где V_k — матрица первых k собственных векторов.

5. **Оценка качества**: Доля объяснённой дисперсии:

$$\gamma = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}.$$

3 Доказательство того, что оптимальные направления в PCA совпадают с собственными векторами матрицы ковариаций

Постановка задачи

Дана центрированная матрица данных $\mathbf{X} \in \mathbb{R}^{n \times m}$, где n — количество объектов, а m — количество признаков, причём каждый столбец имеет нулевое среднее ($\sum_{i=1}^n x_{ij} = 0$). Матрица ковариаций определяется как:

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

где $\mathbf{X}^T$ — транспонированная матрица размером $m \times n$, а $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{m \times m}$.

Цель PCA — найти единичные векторы $\mathbf{v} \in \mathbb{R}^m$, такие что $\|\mathbf{v}\| = 1$, которые максимизируют дисперсию данных, спроецированных на $\mathbf{v}$. Эти векторы называются главными компонентами.

Доказательство

Шаг 1: Дисперсия проекций

Рассмотрим проекцию i-го объекта $\mathbf{x}_i \in \mathbb{R}^{1 \times m}$ (i-я строка матрицы $\mathbf{X}$) на единичный вектор $\mathbf{v}$:

$$\text{proj}_{\mathbf{v}}(\mathbf{x}_i) = \mathbf{x}_i \cdot \mathbf{v}$$

Поскольку данные центрированы, среднее значение проекций равно нулю:

$$\text{Mean}(\text{proj}_{\mathbf{v}}) = \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i \cdot \mathbf{v}) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m x_{ij}v_j = \sum_{j=1}^m v_j\left(\frac{1}{n}\sum_{i=1}^n x_{ij}\right) = 0$$

так как $\sum_{i=1}^n x_{ij} = 0$ для всех j.

Дисперсия проекций определяется как:

$$\text{Var}(\text{proj}_{\mathbf{v}}) = \frac{1}{n-1}\sum_{i=1}^n(\mathbf{x}_i \cdot \mathbf{v})^2$$

В матричной форме сумма квадратов проекций равна:

$$\sum_{i=1}^n(\mathbf{x}_i \cdot \mathbf{v})^2 = (\mathbf{X}\mathbf{v})^T(\mathbf{X}\mathbf{v}) = \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}$$

где $\mathbf{X}\mathbf{v} \in \mathbb{R}^{n \times 1}$ содержит проекции всех объектов. Таким образом, дисперсия:

$$\text{Var}(\text{proj}_{\mathbf{v}}) = \frac{1}{n-1}\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}$$

Подставляя $\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$, получаем:

$$\text{Var}(\text{proj}_{\mathbf{v}}) = \mathbf{v}^T\mathbf{\Sigma}\mathbf{v}$$

Задача сводится к максимизации квадратичной формы $\mathbf{v}^T\mathbf{\Sigma}\mathbf{v}$ при условии $\mathbf{v}^T\mathbf{v} = 1$.

Шаг 2: Свойства матрицы ковариаций

Матрица ковариаций $\mathbf{\Sigma}$ обладает следующими свойствами:

1. **Симметричность**:

$$\mathbf{\Sigma}^T = \left(\frac{1}{n-1}\mathbf{X}^T\mathbf{X}\right)^T = \frac{1}{n-1}(\mathbf{X}^T\mathbf{X})^T = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{\Sigma}$$

   Таким образом, $\mathbf{\Sigma}$ — симметричная матрица.

2. **Неотрицательная определённость**: Для любого ненулевого вектора $\mathbf{w} \in \mathbb{R}^m$,

$$\mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \frac{1}{n-1}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = \frac{1}{n-1}\|\mathbf{X}\mathbf{w}\|^2 \geq 0$$

   Так как $\|\mathbf{X}\mathbf{w}\|^2 \geq 0$, $\mathbf{\Sigma}$ неотрицательно определена. Если $\mathbf{X}$ имеет полный ранг $(n \geq m)$, то $\mathbf{\Sigma}$ положительно определена.

3. **Спектральное разложение**: Согласно спектральной теореме для симметричных матриц, $\mathbf{\Sigma}$ раскладывается как:

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

   где $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m]$ — ортогональная матрица собственных векторов ($\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$), а $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ — диагональная матрица неотрицательных собственных значений, упорядоченных как $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. Собственные векторы удовлетворяют:

$$\mathbf{\Sigma}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

Шаг 3: Максимизация квадратичной формы

Необходимо решить задачу:

$$\max_{\mathbf{v}} \mathbf{v}^T\mathbf{\Sigma}\mathbf{v} \quad \text{при условии} \quad \mathbf{v}^T\mathbf{v} = 1$$

По спектральной теореме любой единичный вектор $\mathbf{v}$ можно представить в базисе собственных векторов:

$$\mathbf{v} = \sum_{i=1}^{m} \alpha_i\mathbf{v}_i, \quad \text{где} \quad \sum_{i=1}^{m} \alpha_i^2 = 1$$

Квадратичная форма равна:

$$\mathbf{v}^T\mathbf{\Sigma}\mathbf{v} = \left(\sum_{i=1}^{m} \alpha_i\mathbf{v}_i\right)^T \mathbf{\Sigma} \left(\sum_{j=1}^{m} \alpha_j\mathbf{v}_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j\mathbf{v}_i^T\mathbf{\Sigma}\mathbf{v}_j$$

Используя $\mathbf{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j$ и ортогональность $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$, получаем:

$$\mathbf{v}^T\mathbf{\Sigma}\mathbf{v} = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j\lambda_j\mathbf{v}_i^T\mathbf{v}_j = \sum_{i=1}^{m} \alpha_i^2\lambda_i$$

Поскольку $\lambda_1 \geq \lambda_i$ для всех i

$$\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \sum_{i=1}^{m} \alpha_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{m} \alpha_i^2 = \lambda_1$$

Равенство достигается при $\mathbf{v} = \mathbf{v}_1$, собственном векторе, соответствующем наибольшему собственному значению λ_1. Таким образом, максимальная дисперсия:

$$\mathrm{Var}(\mathrm{proj}_{\mathbf{v}}) = \lambda_1$$

достигается при $\mathbf{v} = \mathbf{v}_1$.

Шаг 4: Последующие главные компоненты

Вторая главная компонента $\mathbf{v}_2$ максимизирует дисперсию при условии ортогональности к $\mathbf{v}_1$: $\mathbf{v}_2^T \mathbf{v}_1 = 0$. Представим $\mathbf{v} = \sum_{i=1}^{m} \alpha_i \mathbf{v}_i$ с $\sum_{i=1}^{m} \alpha_i^2 = 1$. Условие ортогональности:

$$\mathbf{v}^T \mathbf{v}_1 = \sum_{i=1}^{m} \alpha_i \mathbf{v}_i^T \mathbf{v}_1 = \alpha_1 = 0$$

Таким образом, $\mathbf{v} = \sum_{i=2}^{m} \alpha_i \mathbf{v}_i$, и:

$$\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \sum_{i=2}^{m} \alpha_i^2 \lambda_i \leq \lambda_2 \sum_{i=2}^{m} \alpha_i^2 = \lambda_2$$

Максимум достигается при $\alpha_2 = 1$, $\alpha_i = 0$ для $i > 2$, то есть $\mathbf{v} = \mathbf{v}_2$.

Аналогично, k-я главная компонента — это собственный вектор $\mathbf{v}_k$, соответствующий λ_k, ортогональный всем $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$.

Шаг 5: Заключение

Главные компоненты — это собственные векторы матрицы $\mathbf{\Sigma}$, упорядоченные по убыванию собственных значений:

- Первая главная компонента $\mathbf{v}_1$ соответствует λ_1 и задаёт направление максимальной дисперсии.
- Вторая главная компонента $\mathbf{v}_2$ соответствует λ_2 и максимизирует дисперсию в подпространстве, ортогональном $\mathbf{v}_1$, и так далее.

4 Реализация

Реализация выполнена на Python с использованием собственных классов и методов для матричных операций. Все операции производятся над разреженными матрицами. Ниже описаны ключевые функции.

4.1 Метод Гаусса

Метод Гаусса с частичным выбором главного элемента решает СЛАУ $Ax = b$.

4.2 Центрирование данных

Центрирование вычитает среднее значение по каждому признаку.

4.3 Матрица ковариаций

Матрица ковариаций вычисляется по центрированным данным.

4.4 Собственные значения и векторы

Собственные значения находятся методом бисекции, а векторы — решением однородной СЛАУ.

4.5 Полный алгоритм PCA

Основная функция PCA объединяет все шаги.

4.6 Анализ шума

Функция `add_noise_and_compare` добавляет гауссов шум и сравнивает результаты PCA.

5 Результаты

Для тестирования использовалась матрица X размером 4×2:

$$X = \begin{bmatrix} 1 & 2 \\ 4 & 3 \\ 5 & 6 \\ 8 & 7 \end{bmatrix}$$

Применён шум с уровнем 0.2. Результаты функции `add_noise_and_compare`:

- Доля объяснённой дисперсии (без шума): 1.
- Доля объяснённой дисперсии (с шумом): 1.
- Среднеквадратичная ошибка проекций: 1.24.

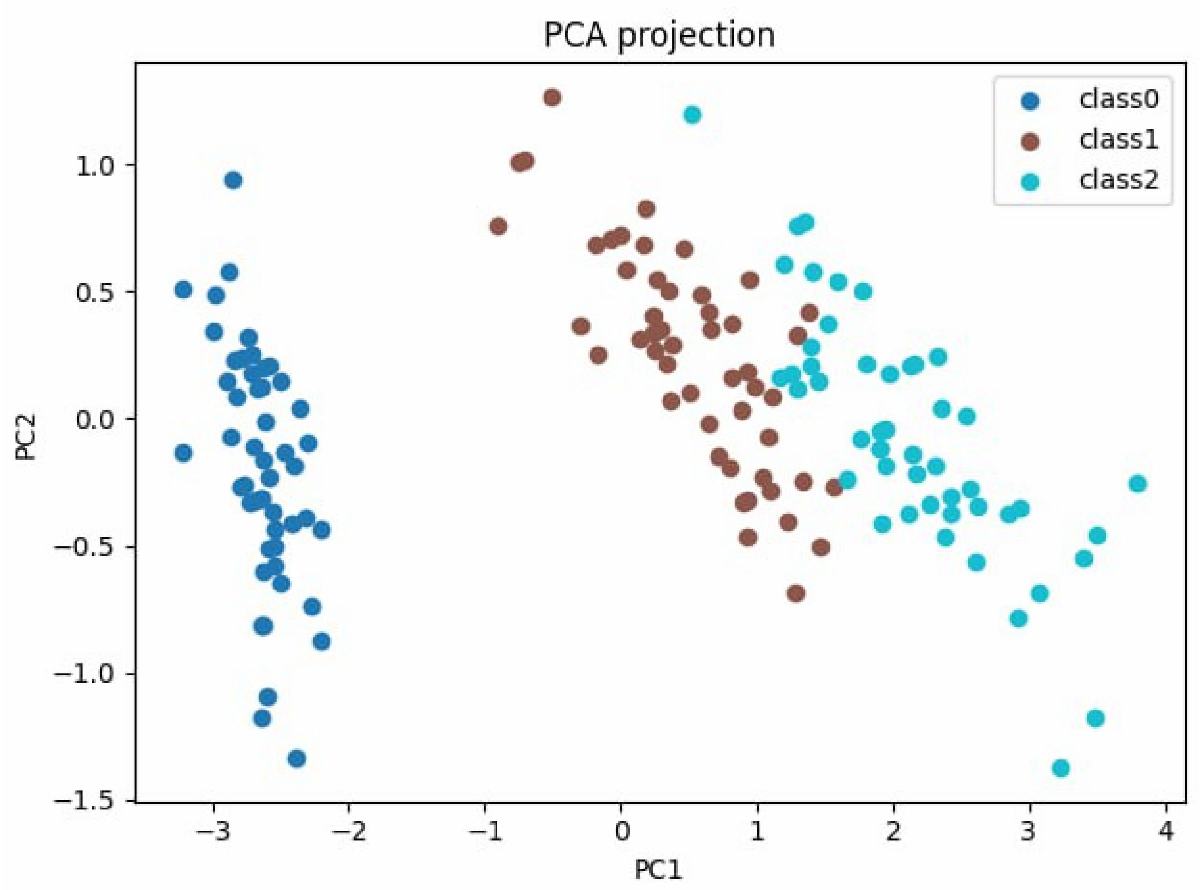

Рис. 1: Визуализация проекций на первые две главные компоненты для датасета iris.