Лабораторная работа 1: Алгоритмы разложения матриц. PCA.

Цель работы: Использование методов матричного разложения в алгоритмах обработки данных (метода главных компонент).

Описание: В этой лабораторной работе сделан акцент на использовании методов матричного разложения для решения прикладной задачи - использования метода главных компонент (PCA) для анализа произвольного массива данных.

Предлагаемые методы: При выполнении данной лабораторной работы предлагается использовать метод главных компонент для анализа массива входных данных. Основной идеей данного алгоритма является снижение размерности анализируемого объекта путём линейного преобразования входных данных в новую координатную систему таким образом, что при помощи меньшего числа измерений можно описать большую дисперсию входных данных. Данное линейное преобразование можно представить следующим образом:

Пусть $X = \mathbf{x}_1, \ ... \ , \mathbf{x}_n{}^T$ - матрица входных данных, где $\mathbf{x}_i$ - вектор длины m, описывающий i-ую запись входных данных. Матрица входных данных должна быть центрированна (по каждому признаку): $x_i'j = x_ij - 1/n \sum_k x_kj$, где i - индекс вектора данных, а j - индекс признака. Линейное преобразование переводит вектор записи входных данных в новую форму $\mathbf{t} = (t_1, \ ... \ t_s)$, где s - параметр, определяющий, на сколько главных компонент проецируются данные, а W - матрица линейного преобразования, размерности $m \times s$.

$$\mathbf{t}_i = \mathbf{x}_i W \tag{1}$$

Задача определения главных компонент допускает использование собственных значений и собственных векторов матрицы ковариации. Требование ортогональности и задание максимальной дисперсии при помощи компонент приводит к тому [для доказательства см. источники, например [2]], что w_j соответствуют собственным векторам матрицы X^TX. Вклад j-ой компоненты в описание дисперсии данных пропорционален отношению сингулярного числа σ_j к суммы сингулярных чисел $\sum_k \sigma_k$ матрицы X^TX.

Использование матричных разложений:

Нахождение главных компонент, описывающих данные, при помощи собственных векторов/значений нормальной матрицы X^TX, можно произвести через матричные разложения.

Подобный подход предполагает сингулярное разложение матрицы данных X, имеющей размерности $n \times m$. Сингулярное разложение является обощением спектрального разложения на прямоугольные матрицы. Матрица измерений представляется через матричное произведение (2).

$$X = U\Sigma V^T \tag{2}$$

Здесь $\Sigma = \mathrm{Diag}\{\sigma_1, \ ... \ , \sigma_p\}, \ p = \min(m, n)$ - прямоугольная диагональная матрица размерности $n \times m$, где на диагонали находятся неотрицательные числа σ_i, называемые сингулярными числами. Сингулярные числа определяются через собственные значения нормальной матрицы XX^T, которые мы обозначим как $\lambda_i(XX^T)$:

$$\sigma_i(X) = \sqrt{\lambda_i(XX^T)} \tag{3}$$

U - квадратная матрица $n \times n$, которая содержит "левые сингулярные векторы" матрицы X, которые соответствуют собственным векторам матрицы XX^T.

Аналогично определяется матрица V, содержащая "правые сингулярные векторы" (соответствуют собственным векторам матрицы X^TX).

Вычисление сингулярных векторов и сингулярных чисел для матрицы X можно производить следующим образом. Для определения сингулярных значений матрицы нужно вычислить собственные значения матрицы XX^T. Левые и правые сингулярные векторы определяются через собственные векторы матриц XX^T и X^TX.

Ход работы и задачи:

1. Реализуйте сингулярное матричные разложения (не используя готовые решения вроде numpy.linalg.svd, или numpy.linalg.eig для поиска собственных векторов и чисел);

2. Выберите произвольный многомерный массив данных, содержащий минимум 5 признаков;

3. Используйте написанный метод сингулярного разложения для получения матрицы преобразований данных к главным компонентам и значения объясняемых дисперсий. Опишите ограничения применимости реализаций SVD к решению данной задачи;

4. Определите достаточное число компонент для описания процесса, визуализируйте данные и проведите анализ полученных компонент;

5. (Дополнительное задание) Реализуйте методы сингулярного разложения матрицы, имеющие меньшую вычислительную сложность, чем классический подход, основанный на собственных векторах и собственных числах матрицы X^TX.

Список литературы

[1] Banerjee, S., Roy, A., Linear Algebra and Matrix Analysis for Statistics (1st ed.), Chapman and Hall/CRC, 2014, https://doi.org/10.1201/b17040.

[2] I. T. Jolliffe, Principal Component Analysis, Springer New York, NY, 2002, https://doi.org/10.1007/b98835.

[3] Menon, Aditya Krishna and Elkan, Charles, "Fast Algorithms for Approximating the Singular Value Decomposition", Association for Computing Machinery, New York, NY, USA, 2011. year = 2011,