Машинное обучение

Методы понижения размерности

Маничева А.С.,
доцент, канд. техн. наук

Постановка задачи по сокращению размерности

Сокращение или понижение размерности – задача обучения без учителя.

Суть задачи понижения размерности – имея данные с большим количеством признаков (столбцов), надо преобразовать их в новую таблицу с меньшим количеством признаков (столбцов). Количество объектов (строк) при этом останется неизменным.

Зачем использовать сокращение размерности

- Сжатие данных
- Ускорение предсказаний
- Визуализация данных
- Более компактное и «правильное» описание объектов
- Повышение интерпретируемости

Методы выделения новых признаков

Методы выделения новых признаков – методы, создающие новые столбцы, которые вычисляются по формулам, зависящим от имеющихся в данных столбцов.

Новые признаки могут быть сложно интерпретируемыми, но могут содержать гораздо больше информации об объектах, чем любой набор исходных признаков того же количества.

Методы снижения размерности пространства признаков:

- **метод главных компонент** (Principal Component Analysis, PCA)
- **топологический анализ данных** (Topological data analysis, TDA)
- **UMAP** (Uniform Manifold Approximation and Projection)

Топологический анализ данных

Топологический анализ данных (Topological data analysis, TDA) – это современное и быстро развивающееся направление анализа данных и компьютерного зрения, которое обеспечивает основу для анализа данных, у объектов которых помимо собственных признаковых описаний существуют некоторые взаимосвязи, описывающие взаимодействие подмножеств этих объектов.

Основные вопросы:

- Как из низкоразмерных представлений получать структуры высоких размерностей.
- Как дискретные единицы складываются в глобальные структуры.

Топологический анализ данных за счет использования методов из топологии и геометрии обеспечивает уменьшение размерности и устойчивость к шуму.

Метод главных компонент

Метод главных компонент (principal component analysis, PCA) – метод, который осуществляет вращение данных с тем, чтобы преобразованные признаки не коррелировали между собой. Часто это вращение сопровождается выбором подмножества новых признаков в зависимости от их важности с точки зрения интерпретации данных.

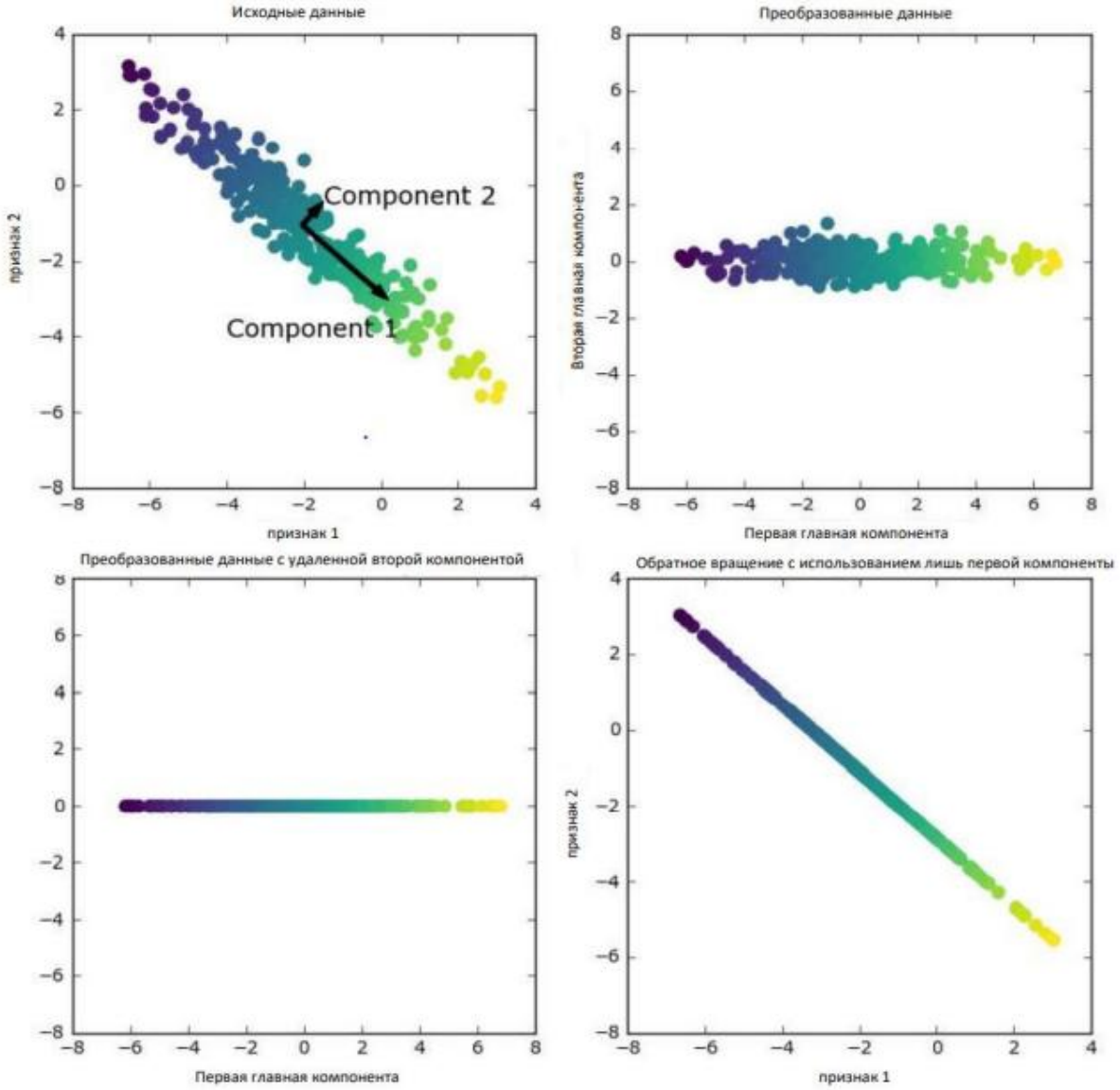
Допущения в задача снижения размерности:

- Переменные сколько-то существенно различны по информативности и мы принципиально можем выкидывать часть из них без критического ущерба для итоговой информации.
- Какая-то часть переменных зависима между собой и может быть заменена на некоторый интегральный показатель на основе их комбинации (обычно линейной).

Метод главных компонент (PCA)

Главные компоненты (principal components) – направления, найденные с помощью PCA.

Максимально возможное количество главных компонент равно количеству исходных признаков.



Метод главных компонент (PCA)

Пусть N – число исследуемых объектов; n – число признаков; матрица Y порядка $n \times N$ – совокупность всех N наблюдаемых значений всех параметров n после нормализации.

Необходимо описать набор признаков m числом главных компонент $m << n$, обеспечивающих долю дисперсии $\gamma \geq 0{,}95$ и сформировать интегральный показатель оптимальности на основе матрицы A весовых коэффициентов, учитывающих тесноту связи между исходными признаками и главными компонентами:

$$Y = A \cdot F,$$

где матрица F включает совокупность всех N полученных значений всех m главных компонент. Или в развернутом матричном виде:

$$\begin{pmatrix} y_{11} & y_{12} & .. & y_{1N} \\ y_{21} & y_{22} & .. & y_{2N} \\ .. & .. & .. & .. \\ y_{n1} & y_{n2} & .. & y_{nN} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1N} \\ a_{21} & a_{22} & .. & a_{2N} \\ .. & .. & .. & .. \\ a_{n1} & a_{n2} & .. & a_{nN} \end{pmatrix} \cdot \begin{pmatrix} f_{11} & f_{12} & .. & f_{1N} \\ f_{21} & f_{22} & .. & f_{2N} \\ .. & .. & .. & .. \\ f_{n1} & f_{n2} & .. & f_{nN} \end{pmatrix} \qquad (1)$$

Задача сводится к определению матрицы A.

Метод главных компонент (PCA)

Связь между главными компонентами и коэффициентами корреляции для объекта *i* :

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jn} f_{ni}, \quad i = 1,..,N, \qquad (2)$$

где y_{ji} – нормированное значение *j*-го признака для *i*-го объекта;
f_{1i} – значение первой главной компоненты для *i*-го объекта.

Полная дисперсия статистического признака выражается через дисперсию главных компонент:

$$\sigma_o^2 = \frac{1}{N}\sum_{i=1}^{N} y_{ji}^2 = \frac{1}{N}\left[a_{j1}^2 \sum_{i=1}^{N} f_{1i}^2 + a_{j2}^2 \sum_{i=1}^{N} f_{2i}^2 + \ldots + a_{jn}^2 \sum_{i=1}^{N} f_{ni}^2 + \right.$$
$$\left. +2\left(a_{j1}a_{j2}\sum_{i=1}^{N} f_{1i} f_{2i} + a_{j1}a_{j3}\sum_{i=1}^{N} f_{1i} f_{3i} + \ldots + a_{j(n-1)}a_{jn}\sum_{i=1}^{N} f_{(n-1)i} f_{ni} \right) \right]. \qquad (3)$$

Выражение 3 упрощается до (4):

$$\sigma_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jn}^2 = 1. \qquad (4)$$

Слева – полная дисперсия, а справа – доли полной дисперсии, относящиеся к соответствующим главным компонентам.

Метод главных компонент (PCA)

Полный вклад *r*-го фактора в дисперсию всех *n* признаков определяет ту долю общей дисперсии, которую данная главная компонента объясняет. Этот вклад вычисляется как:

$$V_r = \sum_{j=1}^{n} a_{jr}^2, \qquad (5)$$

где *j* – индекс признака (показателя коммерческой эффективности); *r* – индекс главной компоненты.

При помощи *m* первых (наиболее весомых, обеспечивающих долю дисперсии $\gamma \geq 0{,}95$) главных компонент можно объяснить основную часть суммарной дисперсии и эти компоненты используются в дальнейшей работе.

Метод главных компонент (PCA)

Применение PCA:

- **визуализация данных;**
- **сокращение размерности в моделях**;
- **выделение скрытых факторов**;
- **фильтр "шума"**.

UMAP

Uniform Manifold Approximation and Projection (UMAP) – это новый алгоритм машинного обучения, выполняющий нелинейное снижение размерности.

Принцип работы метода:

1) выполняется построение взвешенного графа путем соединения ребрами только тех объектов, которые являются ближайшими соседями;
2) создается граф в низкоразмерном пространстве и приближается к исходному.

Полученное множество из ребер определяет новое расположение объектов и, соответственно, низкоразмерное отображение исходного пространства.

UMAP

Достоинства:

- отсутствие ограничений на размерность исходного пространства признаков, которое необходимо уменьшить,
- быстрее и более вычислительно эффективен, чем существующие методы,
- лучше справляется с задачей переноса глобальной структуры данных в новое, уменьшенное пространство.

Спасибо за внимание!