Лекция 10. Методы снижения размерности данных
Тема: PCA, LDA, t-SNE, UMAP

---

1. Введение

В эпоху больших данных аналитики часто сталкиваются с проблемой высокой размерности. Когда количество признаков (атрибутов, переменных) в наборе данных слишком велико, модели машинного обучения становятся сложными, медленными и подверженными переобучению. Этот эффект известен как **«проклятие размерности»**.

Для борьбы с ним применяются **методы снижения размерности (Dimensionality Reduction)** — техники, которые позволяют представить данные в более компактной форме, сохраняя при этом их основную структуру и информативность.

---

2. Цели снижения размерности

Основные задачи этих методов:

- уменьшить количество признаков без потери важной информации;
- улучшить визуализацию данных (особенно в 2D или 3D);
- ускорить обучение моделей;
- устранить шум и коррелированные признаки;
- повысить интерпретируемость данных.

Существует два основных подхода:

1. **Отбор признаков (Feature Selection)** — выбор подмножества наиболее информативных признаков.
2. **Извлечение признаков (Feature Extraction)** — создание новых признаков как комбинаций исходных.

---

3. Метод главных компонент (PCA — Principal Component Analysis)

PCA — один из самых распространённых методов снижения размерности.
Он преобразует исходные признаки в набор **главных компонент** — новых осей, ориентированных вдоль направлений наибольшей дисперсии данных.

Основная идея:
PCA находит такие ортогональные векторы (компоненты), при проекции на которые разброс данных максимален.

Этапы PCA:

1. Центрирование данных (вычитание среднего).
2. Вычисление ковариационной матрицы.
3. Нахождение собственных значений и векторов.

4. Сортировка компонент по величине собственных значений.
5. Выбор первых *k* компонент и проекция данных.

Применение:

- визуализация данных в 2D и 3D;
- удаление шумов;
- предварительная обработка перед классификацией или кластеризацией.

---

4. Линейный дискриминантный анализ (LDA — Linear Discriminant Analysis)

LDA — метод, ориентированный на **классификацию**.
В отличие от PCA, который не учитывает метки классов, LDA использует их, чтобы найти направления, **максимизирующие разделение между классами** при минимизации разброса внутри классов.

Основные шаги LDA:

1. Вычисление средних для каждого класса.
2. Определение внутриклассовой и межклассовой дисперсий.
3. Решение обобщённой задачи собственных значений.
4. Выбор направлений, которые дают наилучшее разделение классов.

Пример применения:
распознавание лиц, диагностика по медицинским данным, биометрия.

---

5. t-SNE (t-distributed Stochastic Neighbor Embedding)

t-SNE — нелинейный метод снижения размерности, разработанный для визуализации сложных данных высокой размерности в 2D или 3D.
Он не сохраняет глобальную структуру, но очень хорошо отображает **локальные группы и кластеры**.

Идея:
t-SNE моделирует вероятностное распределение сходства между точками в исходном пространстве и пытается воспроизвести это распределение в пространстве меньшей размерности.

Преимущества:

- отличная визуализация кластеров;
- работает с любыми нелинейными структурами.

Недостатки:

- высокая вычислительная сложность;
- сложно интерпретировать оси;

- чувствителен к параметрам (например, perplexity).

---

6. UMAP (Uniform Manifold Approximation and Projection)

UMAP — современный метод, разработанный как более быстрый и устойчивый аналог t-SNE.
Он основан на теории **многообразий** и **топологии**, строя граф соседей и затем приближая его в низкоразмерном пространстве.

Преимущества UMAP:

- высокая скорость работы;
- сохранение как локальной, так и глобальной структуры данных;
- возможность использовать для обучения моделей и кластеризации.

Применение:
геномика, обработка изображений, анализ текста и визуализация многомерных признаков.

---

7. Сравнение методов

Метод	Тип	Сохраняет структуру	Учитывает классы	Преимущества	Недостатки
PCA	Линейный	Глобальная	Нет	Простота, скорость	Потеря нелинейных зависимостей
LDA	Линейный	Межклассовая	Да	Хорошо разделяет классы	Только для классификации
t-SNE	Нелинейный	Локальная	Нет	Отличная визуализация	Дорогой по времени
UMAP	Нелинейный	Локальная и глобальная	Нет	Быстрый, гибкий	Зависит от параметров

---

8. Заключение

Методы снижения размерности являются неотъемлемой частью современного анализа данных. Они позволяют:

- упростить сложные модели;
- улучшить визуальное восприятие;
- выделить скрытые закономерности;
- ускорить вычисления и повысить качество анализа.

Выбор метода зависит от задач — если нужно сохранить глобальную структуру и интерпретируемость, подойдёт **PCA**; если важно разделить классы — **LDA**; для визуализации сложных данных — **t-SNE** или **UMAP**.

Список литературы

1. Хэн, Дж., Камбер, М., Пей, Дж. *Интеллектуальный анализ данных: концепции и методы.* — М.: Вильямс, 2019.
2. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.* — O’Reilly Media, 2022.
3. Van der Maaten, L., Hinton, G. *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008.
4. McInnes, L., Healy, J., Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv preprint, 2018.
5. Jolliffe, I. T. *Principal Component Analysis.* — Springer, 2016.