

Лекция 5. Математические основы анализа данных

Тема: Теория вероятностей, статистика, основы линейной алгебры

1. Введение

Любая наука, связанная с данными, имеет прочный математический фундамент.

Интеллектуальный анализ данных (Data Mining), машинное обучение (Machine Learning) и искусственный интеллект (AI) невозможно понять без знания основных разделов математики: **теории вероятностей, статистики и линейной алгебры.**

Эти дисциплины лежат в основе всех алгоритмов анализа и прогнозирования. Математика помогает:

- описывать неопределённость и случайность в данных;
- делать обоснованные выводы на основе выборки;
- работать с многомерными признаками и пространственными структурами.

Таким образом, математика — это язык, на котором "разговаривают" данные и алгоритмы.

2. Теория вероятностей: основы и роль в анализе данных

Теория вероятностей изучает закономерности случайных явлений и позволяет количественно оценивать неопределённость.

Она лежит в основе моделей прогнозирования, классификации и оценки рисков.

2.1. Основные понятия

- **Случайное событие** — результат эксперимента, который может произойти или не произойти.
- **Вероятность события** — числовая мера возможности его наступления (от 0 до 1).
- **Пространство элементарных исходов** — множество всех возможных исходов эксперимента.
- **Случайная величина** — функция, сопоставляющая каждому исходу числовое значение.

Пример:

Если бросить кубик, вероятность выпадения числа 3 равна $\frac{1}{6}$.

2.2. Основные свойства вероятности

1. $0 \leq P(A) \leq 1$
 2. $P(\Omega) = 1$ — вероятность достоверного события равна 1.
 3. Если события несовместимы:
 $P(A \cup B) = P(A) + P(B)$
 4. Для любых событий:
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
-

2.3. Условная вероятность и независимость

- **Условная вероятность:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

— вероятность наступления события A при условии, что произошло событие B.

- **Независимость:**

События A и B независимы, если $P(A \cap B) = P(A)P(B)$

2.4. Формула Байеса

Формула Байеса используется в вероятностных классификаторах, в том числе в **наивном байесовском алгоритме**:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Она позволяет пересчитать вероятность гипотезы (A) после получения новых данных (B).

Пример:

Если вероятность болезни мала, но тест даёт ложноположительные

результаты, формула Байеса помогает корректно оценить реальный риск заболевания.

3. Статистика: оценка, выборка и проверка гипотез

Если теория вероятностей изучает случайные явления **в целом**, то **статистика** помогает делать выводы **на основе конкретных данных (выборок)**.

3.1. Выборка и генеральная совокупность

- **Генеральная совокупность** — вся совокупность объектов исследования.
- **Выборка** — часть генеральной совокупности, используемая для анализа.

Задача статистики — по выборке **оценить свойства генеральной совокупности**.

3.2. Описательная статистика

Используется для суммирования и визуального представления данных.

Основные показатели:

- Среднее значение $\bar{x} = \frac{1}{n} \sum x_i$
- Медиана — центральное значение в упорядоченной выборке
- Мода — наиболее часто встречающееся значение
- Дисперсия $D = \frac{1}{n} \sum (x_i - \bar{x})^2$
- Стандартное отклонение $\sigma = \sqrt{D}$
- Ковариация и корреляция — мера зависимости между переменными

Пример:

Если между доходом и расходами корреляция +0.85, то чем выше доход, тем больше расходы.

3.3. Инференциальная статистика

Инференциальная (выводная) статистика занимается **оценкой параметров** и **проверкой гипотез**.

Оценка параметров:

- **Точечная оценка** — одно число (например, среднее выборки).
- **Доверительный интервал** — диапазон, в котором с заданной вероятностью находится истинное значение параметра.

Проверка гипотез:

1. Формулируется **нулевая гипотеза H_0** (например, «средние равны»).
2. Вычисляется статистика критерия (например, t-критерий Стьюдента).
3. Рассчитывается p-значение (p-value).
4. Если $p < 0.05$, гипотеза отвергается.

Пример:

Проверка, влияет ли новый метод обучения на среднюю успеваемость студентов.

3.4. Распределения вероятностей

В анализе данных часто встречаются различные **распределения случайных величин**:

- **Нормальное распределение (Гауссово)** — симметричное, с колоколоподобной формой.
- **Бернулли / Биномиальное распределение** — для двоичных событий (да/нет).
- **Пуассоновское распределение** — для подсчёта редких событий.
- **Экспоненциальное распределение** — для анализа времени между событиями.

Многие алгоритмы (например, линейная регрессия, статистические тесты) предполагают нормальное распределение данных.

4. Линейная алгебра: работа с многомерными данными

Линейная алгебра является основой машинного обучения и анализа данных, так как все данные в моделях представляются в виде **векторов и матриц**.

4.1. Векторы и матрицы

- **Вектор** — упорядоченный набор чисел, описывающий объект (например, признаки пользователя).
- **Матрица** — таблица чисел, где каждая строка — объект, а каждый столбец — признак.

$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$

4.2. Операции линейной алгебры

- **Скалярное произведение:** $a \cdot b = \sum a_i b_i$ — измеряет схожесть векторов.
 - **Матрица признаков:** используется для хранения данных.
 - **Транспонирование:** A^T — замена строк и столбцов.
 - **Обратная матрица:** A^{-1} — используется для решения систем уравнений.
 - **Детерминант:** характеристика матрицы, важная для анализа линейных зависимостей.
-

4.3. Собственные значения и собственные векторы

Эти понятия лежат в основе **анализа главных компонент (PCA)** — метода снижения размерности данных.

Если:

$$Av = \lambda v$$

то v — собственный вектор, а λ — собственное значение.

Интерпретация:

Собственные векторы показывают направления наибольшей изменчивости данных, а собственные значения — величину этой изменчивости.

4.4. Геометрическая интерпретация

Линейная алгебра помогает понять:

- как данные проецируются на новое пространство признаков;
- как модели (например, регрессия) подбирают гиперплоскость, минимизирующую ошибку;
- как нейронные сети преобразуют данные через матричные операции.

Пример:

В линейной регрессии веса модели w вычисляются как:

$$w = (X^T X)^{-1} X^T y$$

где X — матрица признаков, а y — вектор целевых значений.

5. Взаимосвязь трёх дисциплин

Область	Роль в анализе данных
Теория вероятностей	Моделирует случайность и неопределённость
Статистика	Делает выводы на основе выборок
Линейная алгебра	Представляет и преобразует многомерные данные

Совместное использование этих дисциплин делает возможным:

- обучение моделей (через оптимизацию и вероятности);
- оценку качества предсказаний;
- интерпретацию результатов.

6. Заключение

Математика — это фундамент анализа данных.

Без понимания вероятности, статистики и линейной алгебры невозможно объяснить, **почему** и **как** работает тот или иной алгоритм.

Эти дисциплины позволяют:

- правильно интерпретировать данные;
- избегать ложных выводов;
- создавать модели, устойчивые к шуму и неопределённости.

Как говорил Галилео Галилей:

«Книга природы написана языком математики».

В современном мире данных эта фраза приобретает новое значение — **данные стали новой природой**, а математика остаётся ключом к её пониманию.

Список литературы

1. Гмурман, В. Е. *Теория вероятностей и математическая статистика*. — М.: Высшая школа, 2018.
2. Ланг, С. *Линейная алгебра*. — М.: Мир, 2019.
3. DeGroot, M. H., Schervish, M. J. *Probability and Statistics*. — Pearson, 2014.
4. Strang, G. *Linear Algebra and Its Applications*. — Cengage, 2016.
5. Freedman, D., Pisani, R., Purves, R. *Statistics*. — W. W. Norton & Company, 2007.
6. Bishop, C. M. *Pattern Recognition and Machine Learning*. — Springer, 2016.