ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 9

Вероятностные методы классификации: наивный байесовский классификатор

Неделя 9 · Tan, Steinbach, Kumar: 4.2, 5.3

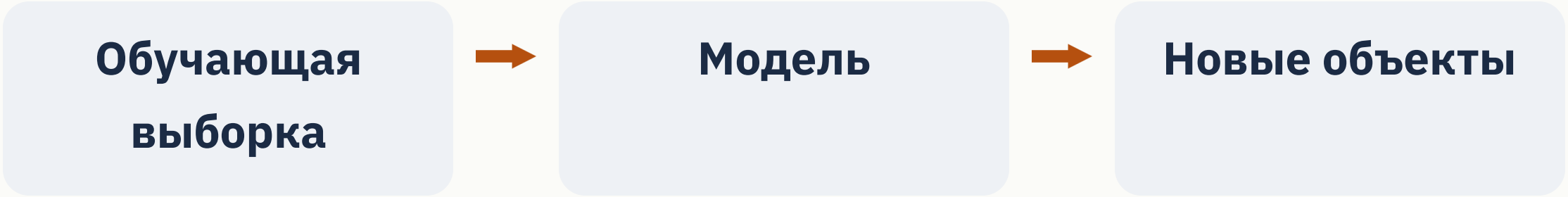
Цель и план лекции

Цель: понять вероятностные основы классификации и применять наивный Байес

1. Задача классификации
2. Условная вероятность и теорема Байеса
3. Наивный Байес: идея и оценка вероятностей
4. Два примера ручного расчёта
5. Сглаживание и реализация в scikit-learn

Задача классификации

Найти функцию, которая по признакам x выдаёт **метку класса** y



Главное — **обобщающая способность**: верно классифицировать объекты, которых модель не видела

Зависимые и независимые события

Независимы: $P(A \cap B) = P(A) \cdot P(B)$

- **Независимые** — две монеты, два кубика
- **Зависимые** — туз из колоды без возврата: $\frac{4}{36}$, затем $\frac{3}{35}$
- **Несовместные** (орёл и решка в одном броске) — зависимы!

Условная и полная вероятность

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

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

Два банка: 30 % и 40 % фальшивых монет, банк выбирается случайно → $P = 0{,}5 \cdot 0{,}3 + 0{,}5 \cdot 0{,}4 = \mathbf{0{,}35}$

Теорема Байеса

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

Болезнь у 1 % людей; тест находит 99 % больных, но даёт 5 % ложных тревог

$$P(\text{болен} \mid +) = \frac{0{,}0099}{0{,}0594} \approx \mathbf{0{,}17}$$

Даже точный тест ошибается при редкой болезни — важна априорная вероятность

Нормальное распределение

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Задаётся средним μ и стандартным отклонением σ
- Правило 68–95–99,7: доля значений в пределах 1σ, 2σ, 3σ
- Центральная предельная теорема: среднее многих величин ≈ нормально
- Используется для непрерывных признаков в наивном Байесе

Теорема Байеса для классификации

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}$$

- $P(Y)$ — доля класса в обучающей выборке
- $P(X \mid Y)$ — насколько типичны признаки для класса
- $P(X)$ одинакова для всех классов — сравниваем числители
- Трудность: оценить $P(X \mid Y)$ для всех комбинаций признаков

«Наивное» предположение

$$P(X \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y) \cdot \ \ldots \ \cdot P(X_d \mid Y)$$

Признаки **условно независимы** при известном классе. Выбираем класс с наибольшим $P(Y) \cdot \prod_{i=1}^{d} P(X_i \mid Y)$

Пример учебника: длина рук и навыки чтения связаны, но при фиксированном возрасте — независимы

Оценка вероятностей

Категориальные

Доля объектов класса с этим значением:

$$P(\text{Жильё} = \text{Да} \mid \text{Нет}) = \frac{3}{7}$$

Непрерывные

Нормальное распределение: среднее и дисперсия признака в классе → плотность

Плотность — не вероятность, но множитель ε одинаков для всех классов и сокращается

Пример 1: угонят ли автомобиль?

№	Цвет	Тип	Производство	Угнан?
1	Красный	Спортивный	Отечественный	Да
2	Красный	Спортивный	Отечественный	Нет
3	Красный	Спортивный	Отечественный	Да
4	Жёлтый	Спортивный	Отечественный	Нет
5	Жёлтый	Спортивный	Импортный	Да
6	Жёлтый	Внедорожник	Импортный	Нет
7	Жёлтый	Внедорожник	Импортный	Да
8	Жёлтый	Внедорожник	Отечественный	Нет
9	Красный	Внедорожник	Импортный	Нет
10	Красный	Спортивный	Импортный	Да

Красный отечественный внедорожник

Признак	$P(\cdot \mid \text{Да})$	$P(\cdot \mid \text{Нет})$
Красный	$\frac{3}{5} = 0{,}6$	$\frac{2}{5} = 0{,}4$
Внедорожник	$\frac{1}{5} = 0{,}2$	$\frac{3}{5} = 0{,}6$
Отечественный	$\frac{2}{5} = 0{,}4$	$\frac{3}{5} = 0{,}6$

«Да»: $0{,}5 \cdot 0{,}6 \cdot 0{,}2 \cdot 0{,}4 = \mathbf{0{,}024}$
«Нет»: $0{,}5 \cdot 0{,}4 \cdot 0{,}6 \cdot 0{,}6 = \mathbf{0{,}072}$

Ответ: «Нет», $P(\text{Нет} \mid X) = \frac{0{,}072}{0{,}096} = 0{,}75$

Пример 2: кредит (учебник Tan)

Заёмщик: не владеет жильём, женат, доход 120K. $P(\text{Нет}) = 0{,}7$, $P(\text{Да}) = 0{,}3$

- Доход в классе «Нет»: среднее 110, дисперсия 2975 → плотность при 120 ≈ 0,0072
- $P(X \mid \text{Нет}) = \frac{4}{7} \cdot \frac{4}{7} \cdot 0{,}0072 \approx 0{,}0024$
- $P(X \mid \text{Да}) = 1 \cdot 0 \cdot 1{,}2 \cdot 10^{-9} = \mathbf{0}$ — женатых должников нет

Ответ: «Нет» ($0{,}7 \cdot 0{,}0024 \approx 0{,}0016$ против 0)

Проблема нулевых вероятностей

Один нулевой множитель обнуляет весь класс

$$\text{m-оценка: } P(x_i \mid y_j) = \frac{n_c + m \cdot p}{n + m}$$

- $m = 3$, $p = \frac{1}{3}$: $P(\text{Женат} \mid \text{Да}) = \frac{0 + 1}{6} = \frac{1}{6}$
- **Сглаживание Лапласа** — +1 к каждому счётчику
- В scikit-learn — параметр alpha (по умолчанию 1)

Свойства наивного Байеса

- Устойчив к шуму и нерелевантным признакам
- Естественно обрабатывает пропуски
- Чувствителен к сильно коррелированным признакам
- Очень быстрый, работает на малых выборках
- Применение: спам, тональность отзывов, рубрики новостей, диагностика

Наивный Байес в scikit-learn

Класс	Признаки
GaussianNB	Непрерывные (нормальное распределение)
CategoricalNB	Категориальные
MultinomialNB	Счётчики, частоты слов
BernoulliNB	Бинарные

```
gnb = GaussianNB().fit(X_train, y_train)  # Iris
accuracy_score(y_test, gnb.predict(X_test)) # ≈ 0.91
```

Выводы

- Теорема Байеса пересчитывает вероятность класса по признакам
- Наивный Байес предполагает условную независимость признаков
- Категории — доли, числа — нормальное распределение
- Нули устраняет сглаживание

Далее: логистическая регрессия и деревья решений

Контрольные вопросы

1. Решите задачу о двух банках при 20 % и 50 % фальшивых монет.
2. Объясните априорную, условную и апостериорную вероятности в классификации.
3. Что такое условная независимость и в чём «наивность» метода?
4. Угонят ли жёлтый импортный спортивный автомобиль?
5. Вычислите m-оценку при $n = 5$, $n_c = 0$, $m = 2$, $p = 0{,}5$.
6. Обучите GaussianNB на наборе Wine и выведите точность и вероятности классов.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — 4.2; 5.3. Bayesian Classifiers.
2. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Главы 6, 13.
3. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. — 2017. — Глава 2.
4. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapter 5.
5. Документация scikit-learn: Naive Bayes.