ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 10

Логистическая регрессия и деревья решений

Неделя 10 · Tan, Steinbach, Kumar: 4.1–4.3, 4.4.5

Цель и план лекции

Цель: понять, как работают логистическая регрессия и деревья решений, научиться строить, интерпретировать и сравнивать их

1. От регрессии к классификации: сигмоида
2. Шансы, логарифм шансов, log-loss
3. Дерево решений и алгоритм Ханта
4. Меры неоднородности и выбор разбиения
5. Переобучение, обрезка, scikit-learn

Почему не линейная регрессия?

Классы можно закодировать как 0 и 1, но:

- прямая даёт значения < 0 и > 1 — это не вероятности
- далёкие объекты «тянут» прямую и сдвигают границу
- квадратичная ошибка плохо подходит для ответа 0/1

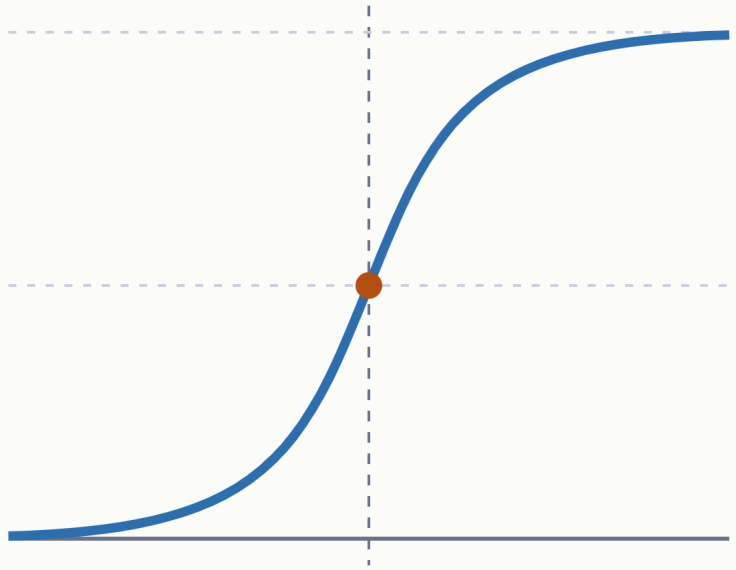
Логистическая регрессия сохраняет линейную модель, но превращает её выход в вероятность от 0 до 1

Сигмоидная функция

$$z = a_1 x_1 + \ \ldots \ + a_k x_k + b$$

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-z}}$$

- $z = 0 \to P = 0{,}5$
- класс 1, если $z \geq 0$
- граница $z = 0$ — прямая: **линейный классификатор**



Шансы и логарифм шансов

- **Шансы** (odds) $= \frac{p}{1-p}$; при $p = 0{,}8$ шансы 4 к 1
- **Логит:** $\ln\left(\frac{p}{1-p}\right) = z = a_1 x_1 + \ \ldots \ + a_k x_k + b$
- Логистическая регрессия — линейная модель для логарифма шансов

Рост признака x_j на 1 умножает шансы в e^{a_j} раз — это **отношение шансов** (odds ratio)

Пример: сдаст ли студент экзамен

Модель: $z = -4 + 1{,}5 \cdot \text{hours}$

Часы	z	$P(\text{сдал})$	Шансы
1	−2,5	0,076	0,08
2	−1,0	0,269	0,37
3	0,5	0,622	1,65
4	2,0	0,881	7,39

Граница: $\text{hours} = \frac{4}{1{,}5} \approx 2{,}67$ ч · каждый час умножает шансы в $e^{1{,}5} \approx 4{,}48$ раза

Обучение: функция потерь log-loss

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln p_i + (1 - y_i)\ln(1 - p_i)\right]$$

$y = 1$, $p = 0{,}9$

штраф $-\ln 0{,}9 \approx 0{,}11$

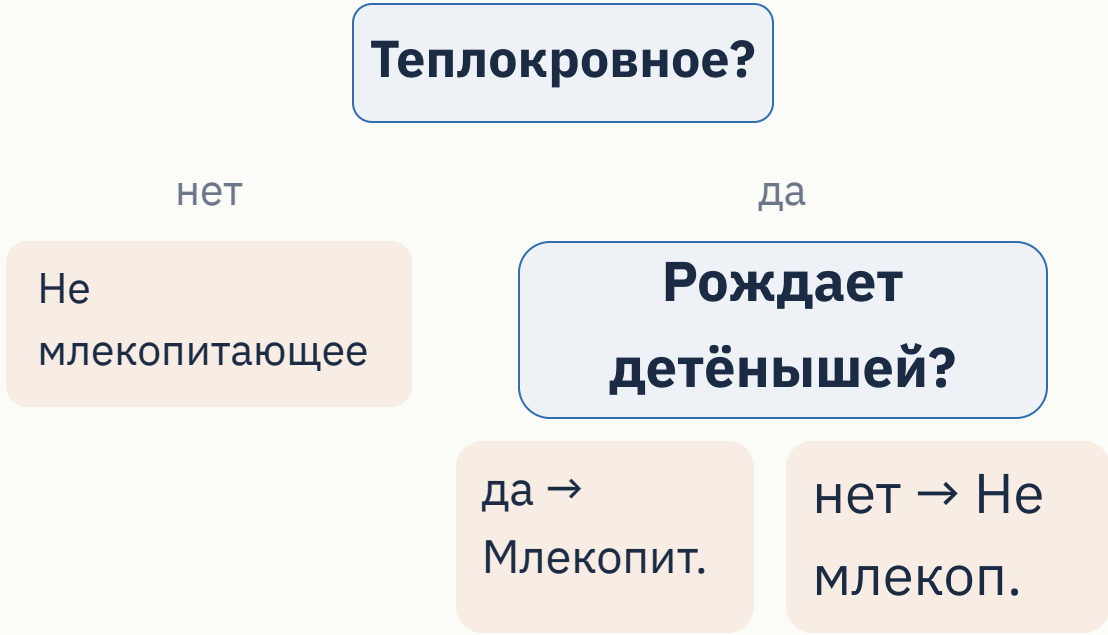
$y = 1$, $p = 0{,}1$

штраф $-\ln 0{,}1 \approx 2{,}30$

- Метод максимального правдоподобия, численная оптимизация (градиентный спуск)
- scikit-learn: L2-регуляризация по умолчанию, параметр C; стандартизируйте признаки

Дерево решений: устройство

- **Корень** — нет входящих рёбер
- **Внутренние узлы** — условия проверки признаков
- **Листья** — метки классов
- Классификация: от корня по ветвям до листа



Фламинго: теплокровное, не рождает детёнышей → «не млекопитающее»

Алгоритм Ханта

Жадное построение сверху вниз (основа ID3, C4.5, CART). D_t — объекты в узле t:

Шаг 1

Все объекты D_t одного класса → t — лист с этой меткой

Шаг 2

Классы разные → выбрать условие, разбить, повторить для каждого потомка

Два вопроса: **как разбивать** и **когда остановиться**

Условия разбиения по типам признаков

Признак	Разбиение
Бинарный	Два исхода: теплокровное / холоднокровное
Номинальный	По ветви на значение или две группы: {Спортивный, Люкс} / {Семейный}
Порядковый	Группы без нарушения порядка: {Малый, Средний} / {Большой}
Непрерывный	$x \leq v$ / $x > v$; порог — середины между соседними значениями

CART использует только бинарные разбиения

Меры неоднородности

$$\text{Entropy} = -\sum_i p(i \mid t)\log_2 p(i \mid t)$$

$$\text{Gini} = 1 - \sum_i p(i \mid t)^2$$

$$\text{Error} = 1 - \max_i p(i \mid t)$$

Узел	Класс 0 / 1	Gini	Энтропия	Ошибка
N1	0 / 6	0	0	0
N2	1 / 5	0,278	0,650	0,167
N3	3 / 3	0,5	1	0,5

Прирост $\Delta = I(\text{родитель}) - \sum_j \frac{N(v_j)}{N} \cdot I(v_j)$; с энтропией — прирост информации

Пример: первое разбиение для задачи о кредите

Корень: 3 «Да», 7 «Нет» → $\text{Gini} = 1 - 0{,}3^2 - 0{,}7^2 = \mathbf{0{,}42}$

Разбиение	**Взвешенный Gini**	**Прирост Δ**
Владеет жильём	0,343	0,077
Семейное положение	0,300	0,120
Доход ≤ 97K	0,300	0,120

- Жильё: владельцы 0, не владельцы 0,490 → $0{,}7 \cdot 0{,}490 = 0{,}343$
- Доход ≤ 97K: слева $\frac{3}{3}$ (0,5), справа все «Нет» → $0{,}6 \cdot 0{,}5 = 0{,}300$

Отношение прироста

Энтропия и Gini предпочитают признаки с множеством значений. «ID клиента» даёт чистые листья, но ничего не предсказывает.

$$\text{Gain ratio} = \frac{\Delta_{\text{info}}}{\text{Split Info}}$$

$$\text{Split Info} = -\sum_i P(v_i) \log_2 P(v_i)$$

- k равных частей → Split Info $= \log_2 k$: больше k — сильнее штраф
- C4.5 — отношение прироста; CART — только бинарные разбиения

Свойства деревьев и переобучение

- Непараметрические, быстрые, легко читаются как правила «если — то»
- Граница из отрезков, параллельных осям («лесенка»)
- Фрагментация: в глубоких узлах мало объектов

Предварительная обрезка

Остановить рост: max_depth, min_samples_split, min_samples_leaf

Последующая обрезка

Построить полное дерево, заменить поддеревья листьями: ccp_alpha

Дерево для задачи о кредите в scikit-learn

```
X = pd.get_dummies(
  loan[['home','marital',
       'income']]).astype(int)
tree = DecisionTreeClassifier(
  criterion='gini',
  random_state=0)
tree.fit(X, loan['default'])
print(export_text(tree,
  feature_names=list(X.columns)))
```

```
|--- income ⩽ 97.50
|   |--- income ⩽ 80.00
|   |   |--- class: Нет
|   |--- income >  80.00
|   |   |--- class: Да
|--- income >  97.50
|   |--- class: Нет
```

Порог 97,5 совпадает с ручным расчётом; правило «должник — доход от 80K до 97,5K» на 10 объектах, скорее всего, случайно

Сравнение на наборе Breast Cancer

```
logreg = make_pipeline(StandardScaler(), LogisticRegression())
deep   = DecisionTreeClassifier(random_state=42)
pruned = DecisionTreeClassifier(max_depth=3, random_state=42)
```

Модель	Обучение	Тест
Логистическая регрессия	0,988	0,986
Дерево без ограничений (глубина 7)	1,000	0,923
Дерево глубины 3	0,977	0,944

Полное дерево переобучено; ограничение глубины сокращает разрыв. Лучше всех — логистическая регрессия

Сравнение классификаторов

Свойство	Наивный Байес	Логист. регрессия	Дерево решений
Граница	По распределениям	Линейная	Ступенчатая
Интерпретация	Вероятности	Отношения шансов	Правила «если — то»
Масштабирование	Не нужно	Желательно	Не нужно
Категории	Естественно	Кодирование	Кодирование
Переобучение	Низкое	Низкое / среднее	Высокое без обрезки
Нелинейность	Ограниченно	Нет	Да

Выводы

- Логистическая регрессия: сигмоида от линейной комбинации, log-loss, отношения шансов
- Дерево решений: серия вопросов, жадный алгоритм Ханта
- Разбиение выбирают по наибольшему уменьшению энтропии или Gini
- Деревья склонны к переобучению — ограничиваем или обрезаем

Далее: метрики качества моделей

Контрольные вопросы

1. Почему линейная регрессия плохо подходит для классификации? Запишите сигмоиду.
2. Шансы, логит, отношение шансов: P и шансы для $z = -2 + 0{,}5x$ при $x = 6$.
3. Устройство дерева, алгоритм Ханта, разбиения для разных типов признаков.
4. Энтропия, Gini и ошибка для узла: 2 объекта A и 6 объектов B.
5. Почему «ID клиента» бесполезен и как помогает отношение прироста?
6. Логистическая регрессия и дерево на наборе Wine: точность и 5 важных признаков.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — 4.1–4.3; 4.4.5.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. — 2017. — Глава 2.
3. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Главы 16, 17.
4. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapter 5.
5. Документация scikit-learn: Logistic regression; Decision Trees.