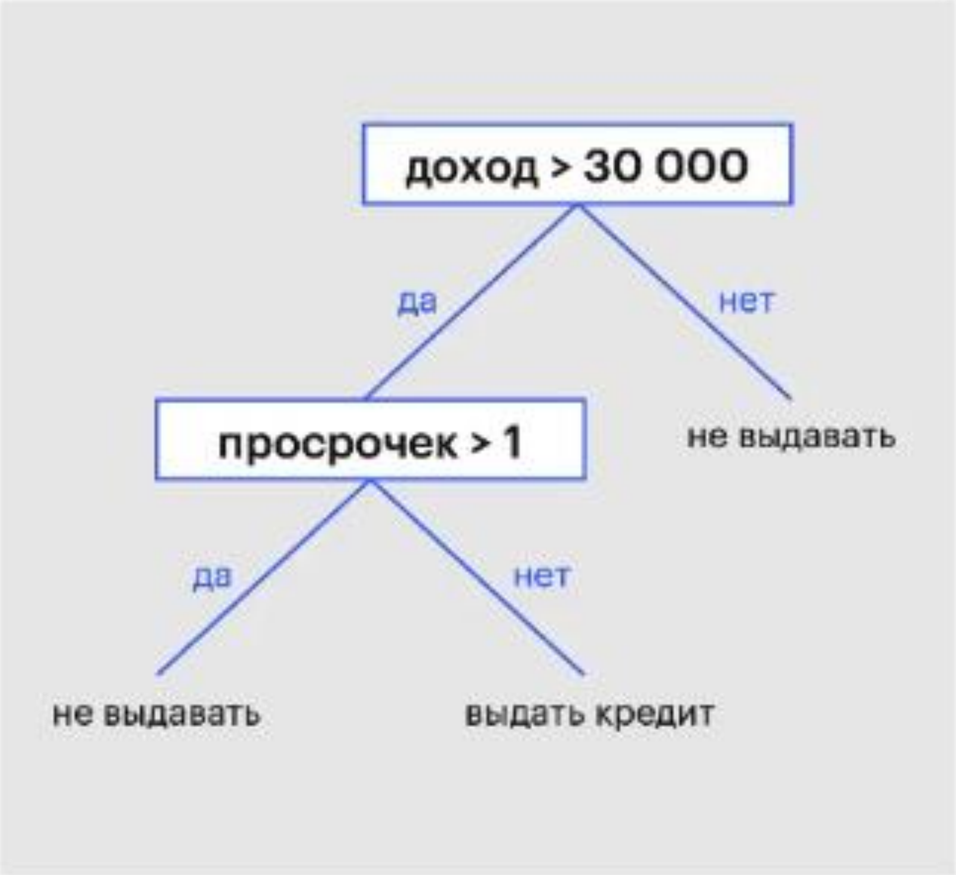
Системы искусственного интеллекта

Лекция 6
Решающие деревья
(Дерево решений, Decision Tree)

Запорожцев Иван Федорович
zaporozhtsev.if.work@gmail.com

Решающее дерево (Decision Tree)



- **Лист или терминальная вершина** (leaf / terminal node) → метка (вероятности меток)
- **Внутренняя вершина** (internal node) → ветвление, предикат (признак, порог)
- **Дуга** → значение предиката

CART = Classification and Regression Trees

Бинарные деревья (binary trees) – каждая вершина имеет двух потомков

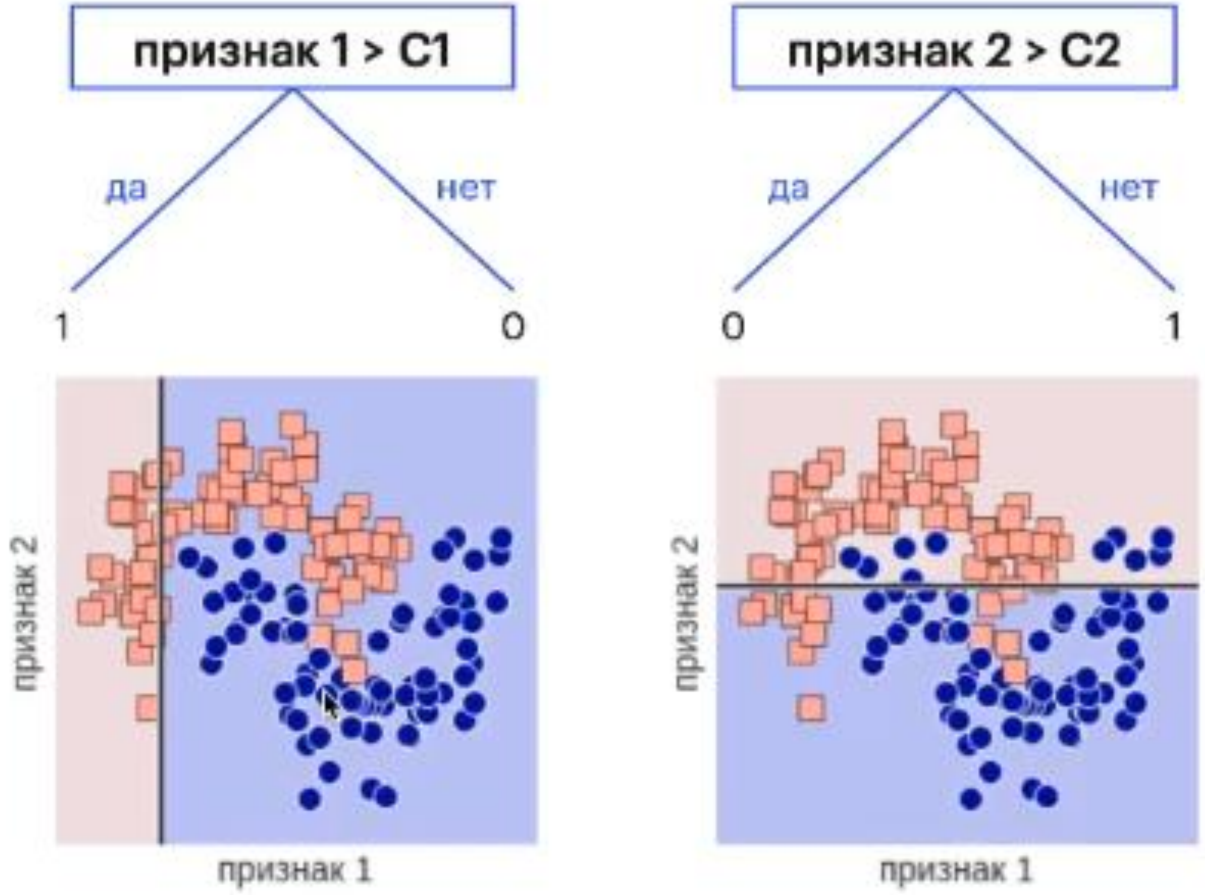
Дальше, в основном, рассматриваем их

Разбиение на области

Расщепление по переменной (splitting)

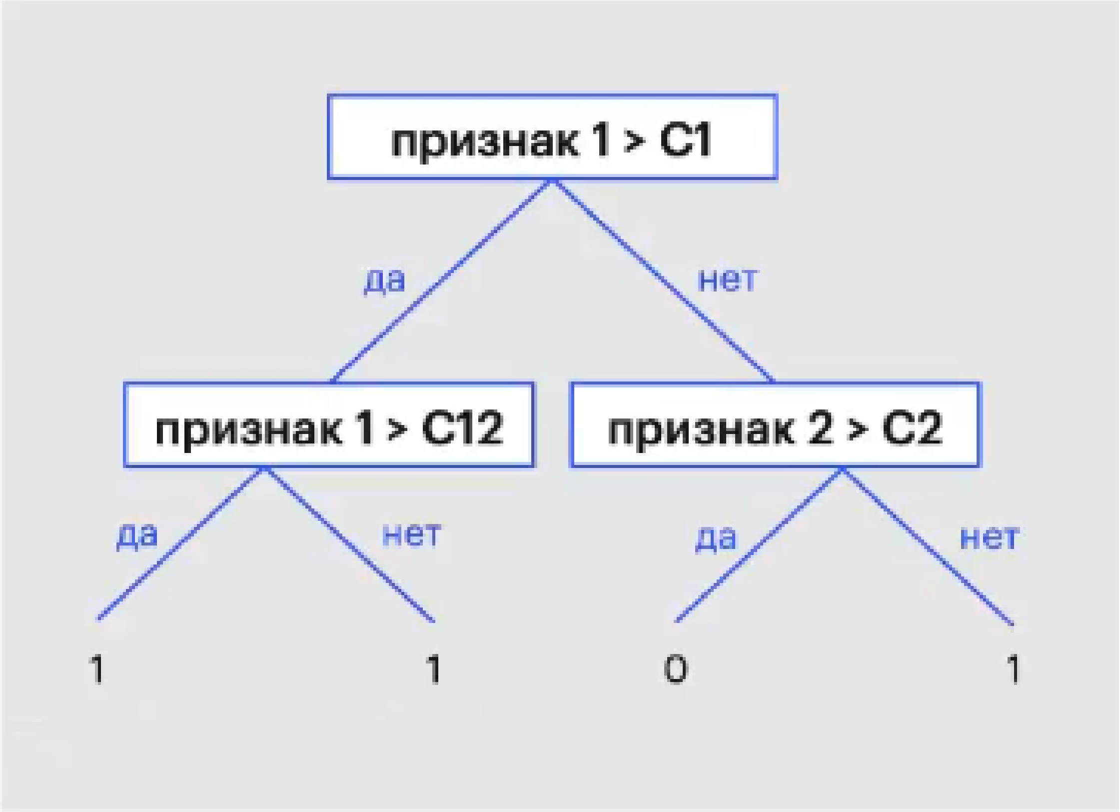
↓

разбиение (stratifying / segmenting) **на области / регионы**



– это, кстати, «решающие пни» (decision stumps)

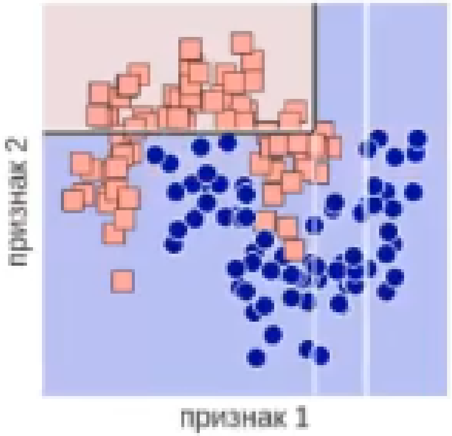
Решающее дерево как функция – кусочно-постоянная



$$a(x) = \sum_j a_{R_j} I[x \in R_j]$$

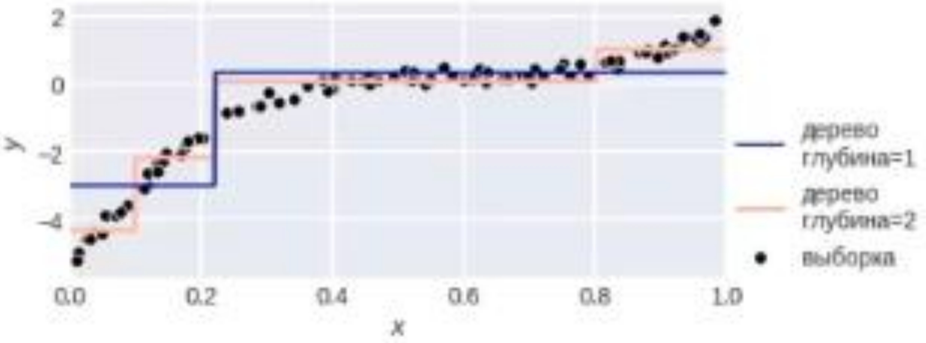
$$\bigcup_j R_j = \aleph$$

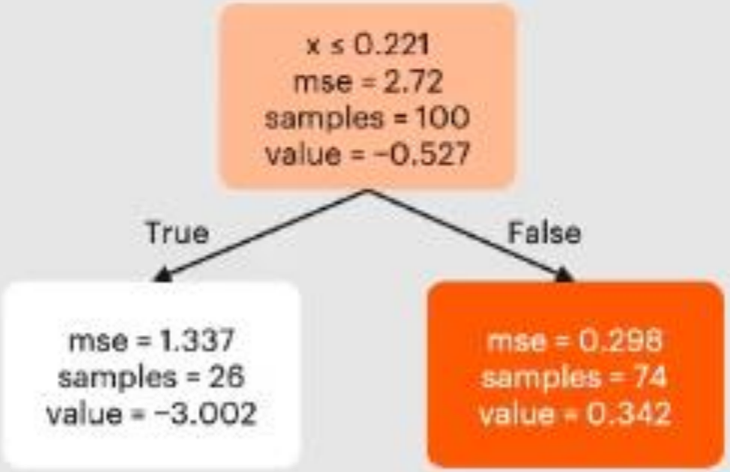
$$R_i \cap R_j = \varnothing \ \forall i \neq j$$



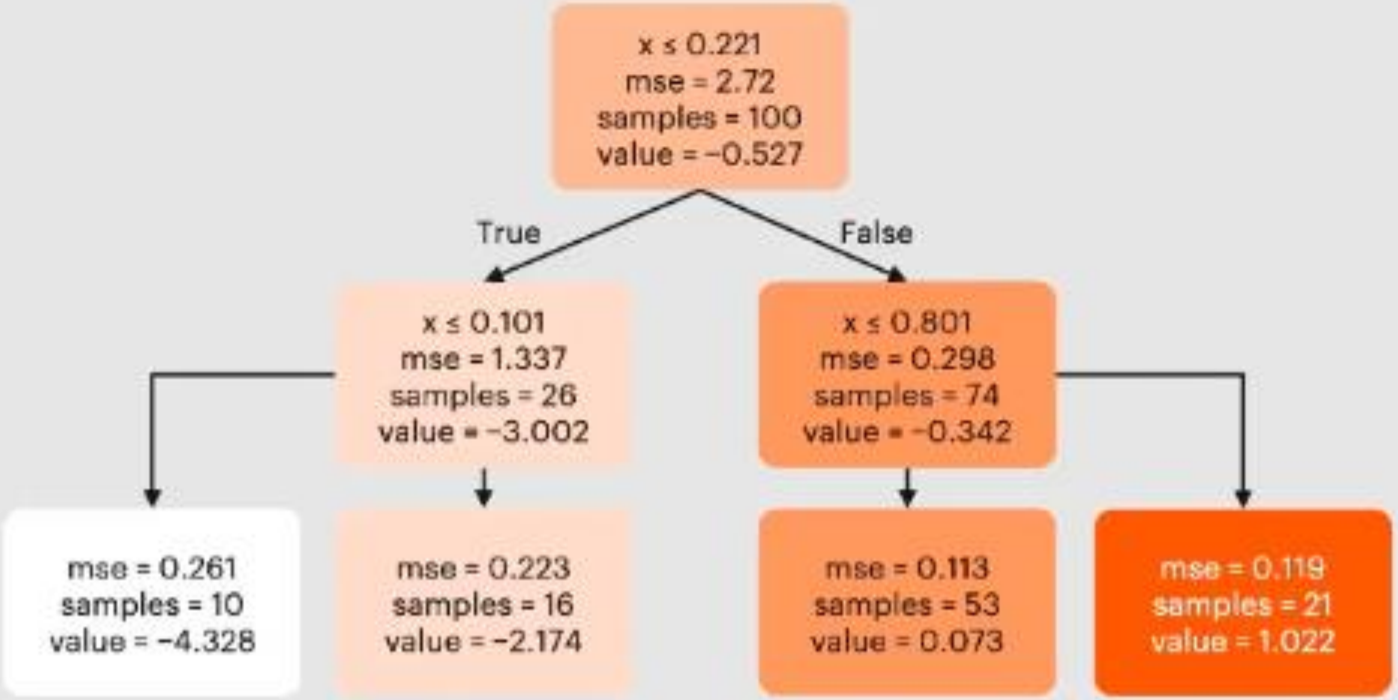
В соседних регионах метки могут быть одинаковые

Решающее дерево в задаче регрессии





$$y = -3.002 I[x \leq 0.221] + 0.342 I[x > 0.221]$$



Как строить дерево?

В идеале в задаче регрессии с MSE нужно решить такую задачу минимизации:

$$\sum_i \sum_j I[x_i \in R_j](y_i - a_{R_j})^2 \to \min$$

– Residual Sum of Squares (RSS)

минимизация по всем разбиениям на области $\{R_j\}$

и по всем выборам a_{R_j}

Трудоёмко → строим дерево «по уровням», последовательно минимизируя RSS
(top-down greedy approach)

Построение дерева

Сверху-вниз

Стартуя от дерева, состоящего из одной вершины, можно проводить расщепления выбирая признак и порог так, чтобы минимизировать RSS

Сейчас уточним – что будем оптимизировать

Расщепления производятся пока не выполнятся некоторые критерии останова (ограничения на глубину дерева, число объектов обучающей выборки в листьях, на изменение RSS, см. дальше)

Построение дерева

Приписывание меток областям

Заметим, что если зафиксировать область, то оптимальное значение метки для области находится просто
(в смысле минимизации эмпирического риска)

В задаче регрессии (MSE)

$$a_{R_j} = \frac{1}{|\{x_i : x_i \in R_j\}|} \sum_{x_i \in R_j} y_i$$

В задаче классификации (точность)

$$a_{R_j} = \mathrm{mode}(\{y_i : x_i \in R_j\})$$

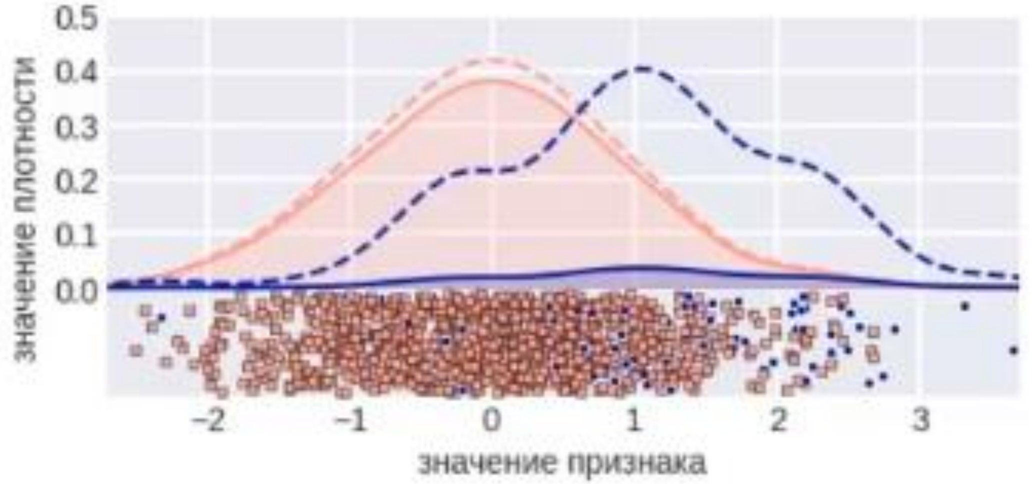
Поэтому именно так они и выбираются!

Возможно другие значения для специальных функционалов качества

Построение дерева

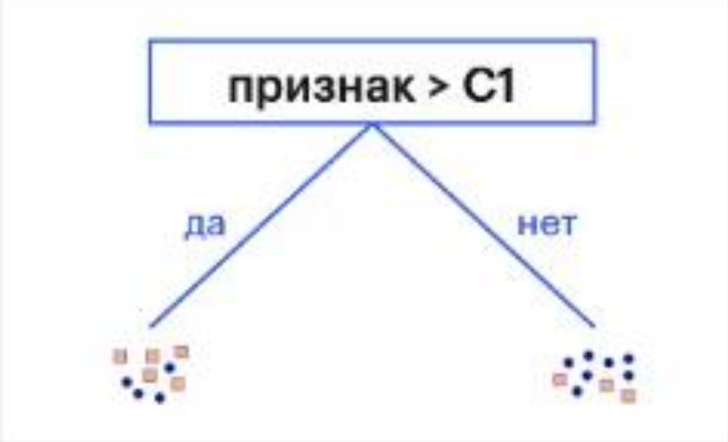
Как делать расщепления

Как выбрать порог для расщепления при построении дерева?



«Хорошо»:

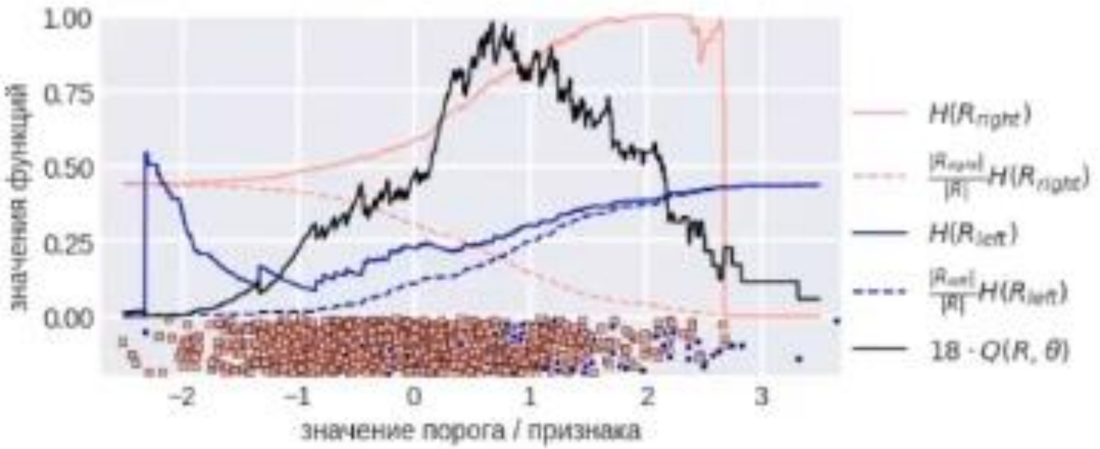
«Плохо»:



Критерии расщепления в задачах классификации

Идея: **ввести меру неоднородности / зашумлённости (impurity) множества** $H(R)$
~ насколько в области «почти все объекты одного класса»
при расщеплении области R на R_{left} и R_{right} оптимизировать

$$Q(R,\theta) = H(R) - \frac{|R_{\text{left}}|}{|R|} H(R_{\text{left}}) - \frac{|R_{\text{right}}|}{|R|} H(R_{\text{right}})$$



Меры impurity

Неоднородности / зашумлённости

В задачах классификации

Пусть есть область R

в ней доли объектов всех классов: $p_1, \ldots, p_l$

Missclassification criteria → $H(R) = 1 - p_{\max}$

Энтропийный → $H(R) = -\sum_j p_j \log_2 p_j$

Джини → $H(R) = \sum_j p_j (1 - p_j) = 1 - \sum_j p_j^2$

Мера неоднородности (impurity) минимальна (=0) только если все объекты принадлежат одному классу

Критерии расщепления

Частный случай двух классов

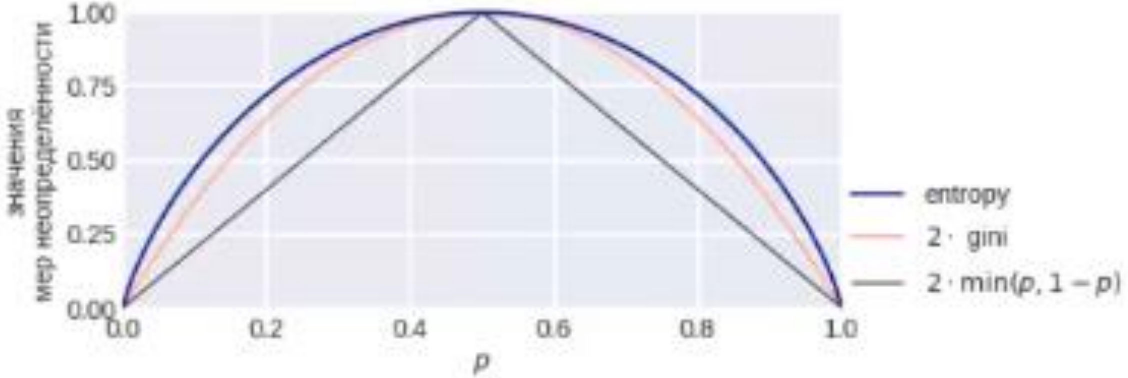
Пусть есть область R

в ней доли объектов всех классов: $p_1 = p, p_2 = 1 - p$

Missclassification criteria → $H(R) = \min[p, 1 - p]$

Энтропийный → $H(R) = -p\log_2 p - (1-p)\log_2(1-p)$

Джини → $H(R) = 2p(1-p) = 1 - p^2 - (1-p)^2$



Критерии расщепления

4 **AUC_ROC** (не будем обосновывать)

$$Q(R,\theta) = \left| \frac{|R_{\text{right}} \cap K_0|}{|K_0|} - \frac{|R_{\text{right}} \cap K_1|}{|K_1|} \right| =$$

$$= \left| \frac{|R_{\text{left}} \cap K_0|}{|K_0|} - \frac{|R_{\text{left}} \cap K_1|}{|K_1|} \right|$$

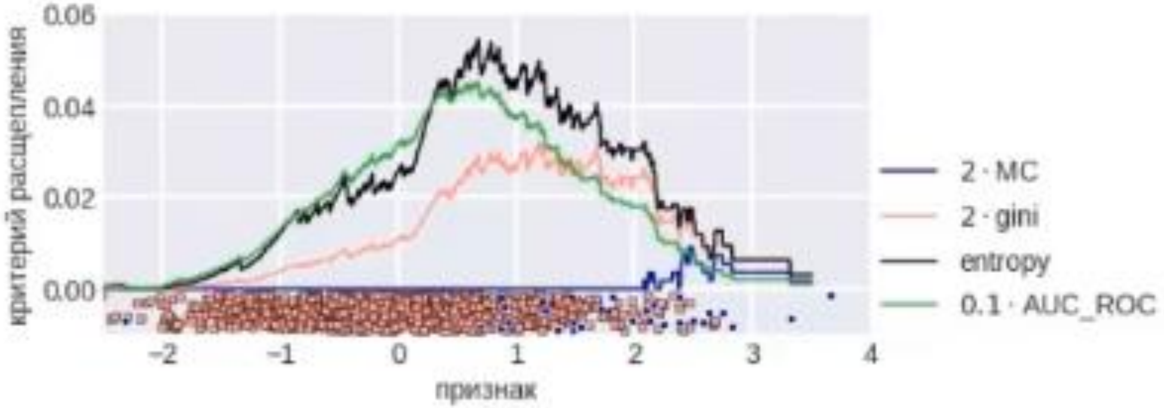
5 **Twoing**

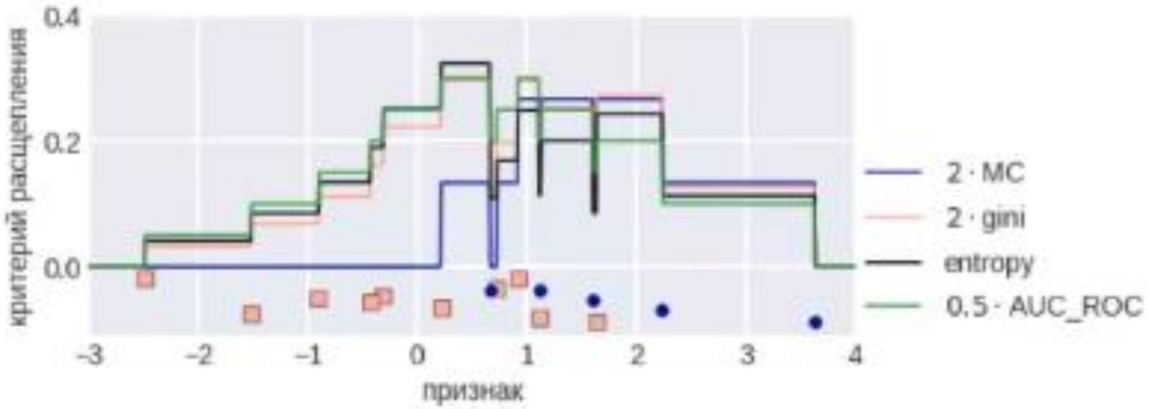
$$Q(R,\theta) = \frac{1}{4} \frac{|R_{\text{left}}|}{|R|} \frac{|R_{\text{right}}|}{|R|} \left(\sum_{j=1}^{l} \left| p_j(R_{\text{left}}) - p_j(R_{\text{right}}) \right| \right)^2$$

Для двух классов ~ Джини

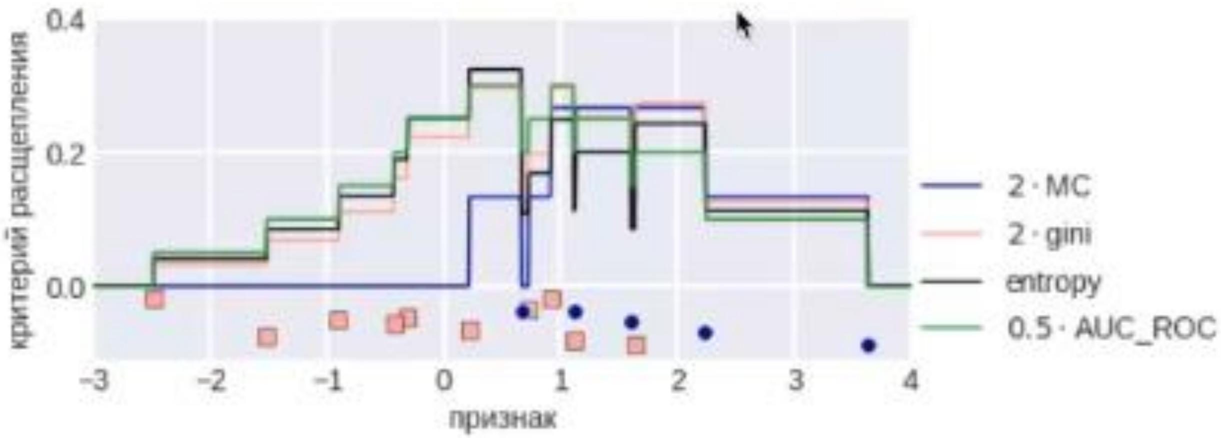
Критерии расщепления

Частный случай двух классов





Энтропия – мера неопределённости распределения



При пороге $\theta = 1$

$$\frac{|R_{\text{left}}|}{|R|} = \frac{9}{15} \qquad \frac{|R_{\text{right}}|}{|R|} = \frac{6}{15}$$

$$H(R) = -(5/15)\log_2(5/15) - (10/15)\log_2(10/15) \approx 0.918$$

$$H(R_{\text{right}}) = -(4/6)\log_2(4/6) - (2/6)\log_2(2/6) \approx 0.918$$

$$H(R_{\text{left}}) = -(1/9)\log_2(1/9) - (8/9)\log_2(8/9) \approx 0.503$$

$$Q(R, \theta) \approx 0.918 - \frac{9}{15}0.503 - \frac{6}{15}0.918 \approx 0.249$$

Напоминание

Теория информации

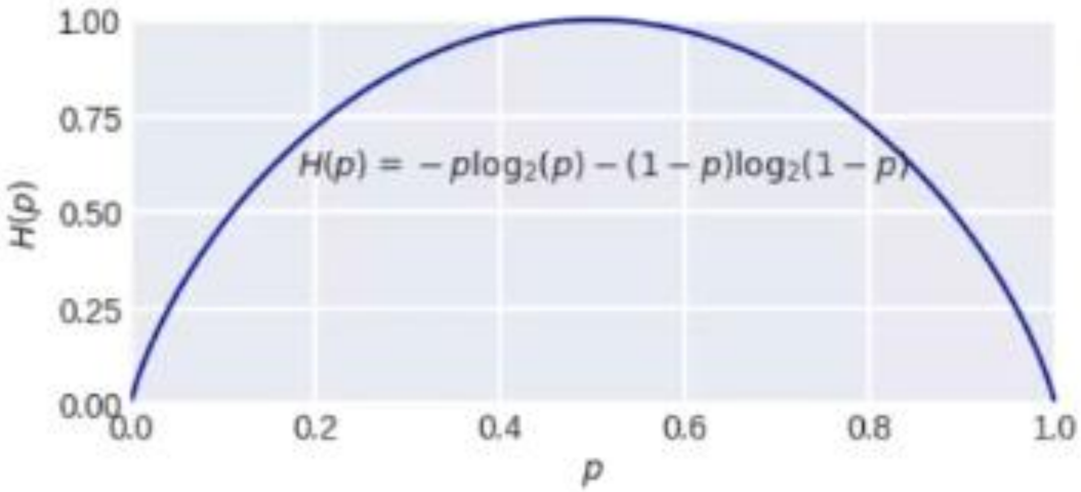
Информационная энтропия (Entropy) – мера неопределённости некоторой системы

$$x \sim (x_1, p_1), (x_2, p_2), \ldots$$

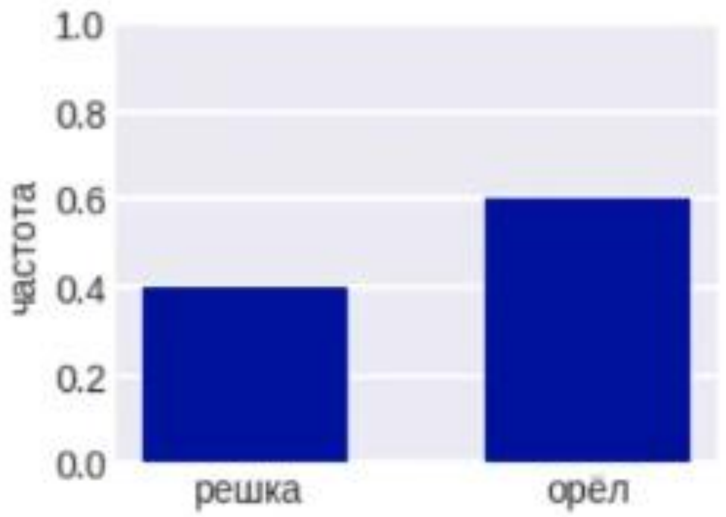
$$H(x) = -\sum_i p_i \log p_i$$

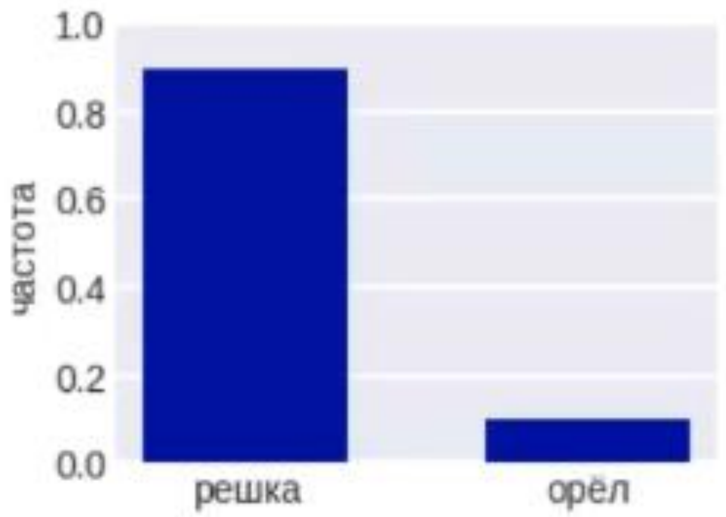
$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

Что зависит от основания логарифма?



Напоминание: теория информации





$$-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10}\approx 0.97$$

$$-\frac{9}{10}\log_2\frac{9}{10}-\frac{1}{10}\log_2\frac{1}{10}\approx 0.47$$

Результат подбрасывания честной монеты – 1 бит информации

Обоснование энтропийного критерия расщепления

	облачно	ясно
дождь	3	1
сухо	1	5

X = {дождь, сухо}
Y = {облачно, ясно}

Совместная энтропия

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) =$$

$$= -\frac{3}{10}\log_2\frac{3}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

Энтропия при условии знаем, что дождь

$$H(Y \mid X = x) = -\sum_{x\in X}\sum_{y\in Y} p(y \mid x)\log_2 p(y \mid x) =$$

$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

Обоснование энтропийного критерия расщепления

	облачно	ясно
дождь	3	1
сухо	1	5

X = {дождь, сухо}
Y = {облачно, ясно}

Ожидаемая условная энтропия

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x)$$

Энтропия при условии, что знаем X

Information Gain / Mutual Information

$$IG(Y \mid X) = H(Y) - H(Y \mid X)$$

Насколько знание X уменьшило неопределённость

Как раз это и считаем!

$$Q(R, \theta) = H(R) - \frac{|R_{\text{left}}|}{|R|} H(R_{\text{left}}) - \frac{|R_{\text{right}}|}{|R|} H(R_{\text{right}})$$

Критерии расщепления

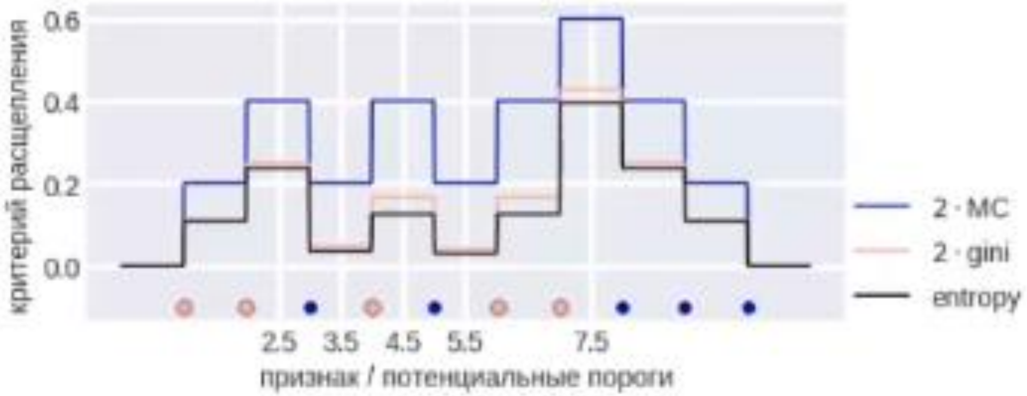
Тонкости

При выборе расщепления мы выбираем порог

- достаточно рассматривать только «средние точки»
- достаточно рассматривать только «границы регионов»

Но в чём тут подвох?

Для начала надо отсортировать все значения
Есть проблема константных признаков



Критерии расщепления в задачах регрессии

Аналогично... но тут «неоднородность» – дисперсия

$$H(R) = \mathrm{var}(\{y_i \mid y_i \in R\})$$

$$Q(R,\theta) = H(R) - \frac{|R_{\mathrm{left}}|}{|R|} H(R_{\mathrm{left}}) - \frac{|R_{\mathrm{right}}|}{|R|} H(R_{\mathrm{right}})$$

Делаем разбиение, которое максимизирует Q
Повторяем процедуру в листьях

Кстати, для бинарной случайной величины $y_i \in \{0,1\}$

Моменты распределения Бернулли [пр

$\mathbb{E}[X] = p$,

$\mathrm{D}[X] = p(1-p) = pq$, так как:

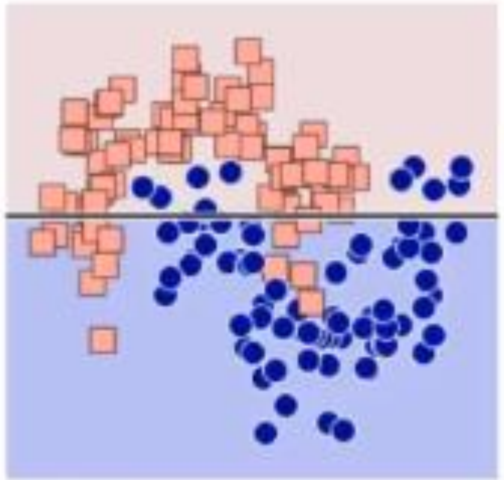
$\mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = p - p^2 = p \cdot (1-p) = pq$

GINI

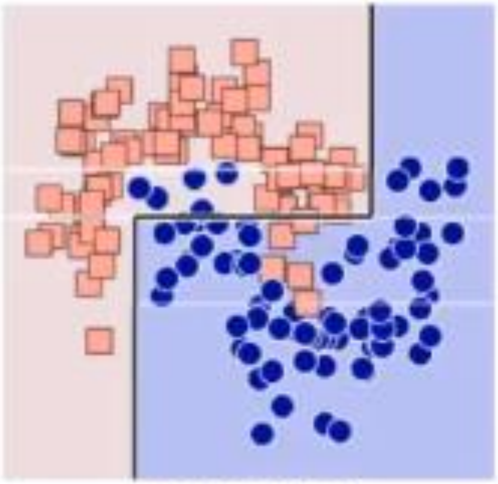
Как долго строить дерево

Критерии останова

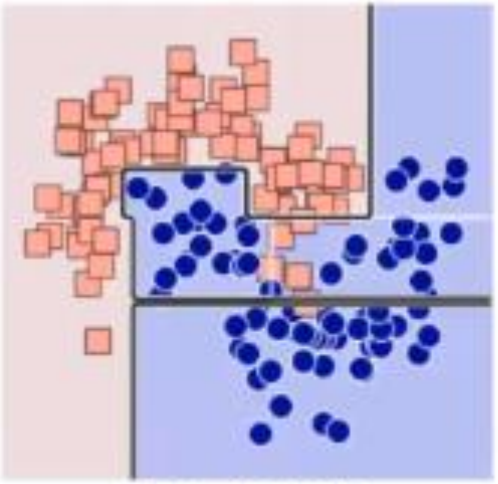
- ограничение на глубину / на число листьев
- ограничение на число объектов в листьях / на число, когда делаем деление
- «естественное ограничение» (все объекты одного класса) обобщение: почти все объекты одного класса
- изменение impurity



max_depth=1



max_depth=3



max_depth=5

Минутка кода: «Решающее дерево»

```
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier(criterion='gini', # критерий расщепления
                                                  # «gini» / «entropy»
     splitter='best',                # разбиение «best» / «random»
     max_depth=None,                 # допустимая глубина
     min_samples_split=2,            # минимальная выборка для разбиения
     min_samples_leaf=1,             # минимальная мощность листа
     min_weight_fraction_leaf=0.0,   # аналогично с весом
     max_features=None,              # #признаков для нахождения разбиения
     random_state=3,
     max_leaf_nodes=None,            # допустимое число листьев
     min_impurity_decrease=0.0,      # порог изменения «загрязнённости»
     min_impurity_split=None,        # порог «загрязнённости» для останова
     class_weight=None,              # веса классов («balanced» или словарь)
    ccp_alpha=0.0)                   # для подрезки

model.fit(X, y)
```

Особенности

Изменение impurity – порог на

$$\frac{|R|}{m}\left(H(R)-\frac{|R_{\text{left}}|}{|R|}H(R_{\text{left}})-\frac{|R_{\text{right}}|}{|R|}H(R_{\text{right}})\right)$$

$$Q(R,\theta)=H(R)-\frac{|R_{\text{left}}|}{|R|}H(R_{\text{left}})-\frac{|R_{\text{right}}|}{|R|}H(R_{\text{right}})$$

Проблема переобучения деревьев

Глубокие деревья склонны к переобучению, поскольку «затачиваются» на отдельные объекты

1. **Прекращают построение достаточно рано** (см. критерии останова, stopping early)

   Можно на отложенной выборке выбрать точку останова

2. **Подрезают деревья** (post-pruning)

3. **Используют в ансамблях** (например, в случайном лесе)

   Отдельная тема

Подрезка (post-pruning)

Сейчас используется редко – если задачу надо решить деревом
(или ансамблем из нескольких деревьев)

Раньше...

- использовали отложенный контроль (удаляли листья, если это повышает качество)
- MDL (Minimum Description Length)

$$\sum_{j} \sum_{x_i \in R_j} (y_i - a_{R_j})^2 + \alpha \, | \{R_j\} | \to \min$$

Оптимальное значение α находят с помощью скользящего контроля, потом с этим значением параметра дерево перестраивается по всей выборке

↓

α регулирует баланс между стремлением обучиться и получить небольшое дерево

Важности признаков

Вспомним формулу → $Q(R,\theta) = H(R) - \frac{|R_{\text{left}}|}{|R|}H(R_{\text{left}}) - \frac{|R_{\text{right}}|}{|R|}H(R_{\text{right}})$

– это уменьшение неоднородности при выборе такого расщепления!

Идея: чем больше признак уменьшает неоднородность, тем он важнее!

Важность признака = сумма уменьшений неоднородностей с помощью этого признака при построении дерева (иногда умножается на $|R|$ ~ sklearn)

Есть и другие способы оценки важности:

- коэффициенты в моделях
- OOB-оценки
- корреляции / функциональные зависимости и т.п.

Деревья: проблема пропусков

Missing Values

- Удалить
- Заменить (средним)
- Рассматривать как отдельную категорию
- Пронести в обе ветви дерева
- Выбрать наиболее подходящую ветвь дерева
- Суррогатные расщепления (Surrogate Splits)

Суррогатные расщепления

Surrogate Splits

Для объектов, у которых расщепляемая переменная неизвестна подбираем другой сплит, который имитирует изначальный

```
A B C Y=target
0 ? 0 0
1 5 1 0
2 ? 0 0
3 8 1 1
? 7 1 1
? 7 1 1
```

«A > 2.5», но не понятно, в какое поддерево класть последние два объекта

«B > 6» на объектах с известными значениями признаков совпадает с исходным,

При этом определено для тех объектов, у которых были неизвестны значения признака А.

www.learnbymarketing.com/methods/

Деревья: категориальные признаки

Формально при расщеплении должны рассмотреть все подмножества множества категорий

↓

Реальнo (в задаче бинарной классификации): упорядочиваем по вероятности класса 1, каждая категория → номер по порядку

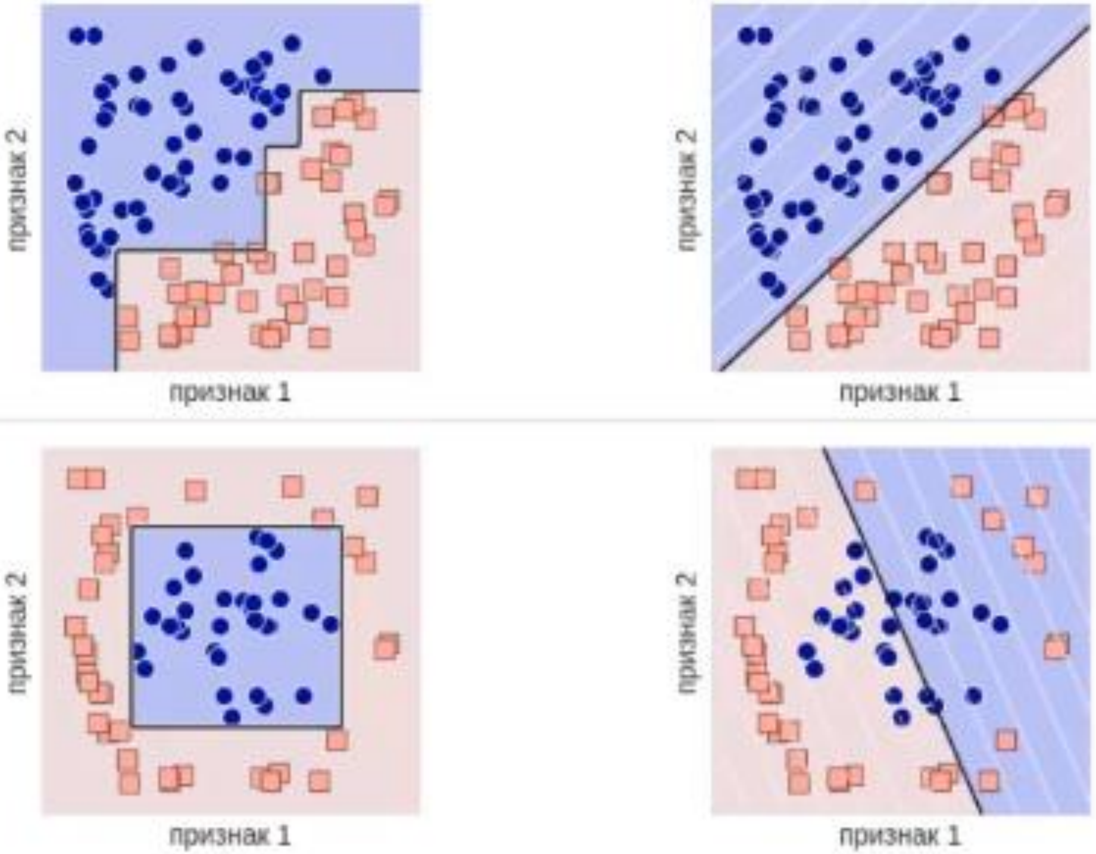
- находим для полученного числового признака оптимальное разбиение
- переобучение для мелких категорий

Деревья vs линейные модели

Дерево | Гиперплоскость

Линейная зависимость

Нелинейная зависимость



Итог: решающие деревья

Возможности

- способны обучиться на любой (непротиворечивой) выборке (при возможности построения неограниченного дерева)
- можно использовать при признаках разных типов (+ пропуски)
- можно сделать устойчивыми к выбросам
- универсальный метод – для всех типов задач машинного обучения
- встроенный отбор признаков
- нелинейный метод!

Качество

- не очень высокое качество решения задачи / переобучение
- хороши в ансамблях будет в ансамблировании

Итог: решающие деревья

Эффективность / стабильность

- достаточно быстро строятся
- нет ограничений на распределения признаков
- «неустойчивый алгоритм» (меняется при небольшом изменении выборки)
- плох для больших / изменяющихся данных

Понимание, интерпретация и анализ

- просто объяснить неспециалисту
- ближе к человеческой логики принятия решения
- можно изобразить (на слайде)
- нет красивой аналитической формулы для модели

Итог: решающие деревья

Особенности

- не использует геометрию (нет расстояний, неметрический)
- устойчив к масштабированию
- устойчив к дубликатам признаков, зависимостям в признаках и т.п.
- автоматическое решение проблемы пропусков
- неспособен к экстраполяции
- использует мало признаков!!!

Важно: эвристическое жадное обучение

(т.к. построение оптимального дерева очень сложная – NP-полная – задача)

Если категориальные признаки с большим число категорий – всё сваливается на них...

Спасибо за внимание!

Запорожцев Иван Федорович
zaporozhtsev.if.work@gmail.com