

# Решающие деревья

Виктор Владимирович Китов

МГУ им.Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

## Определение решающего дерева

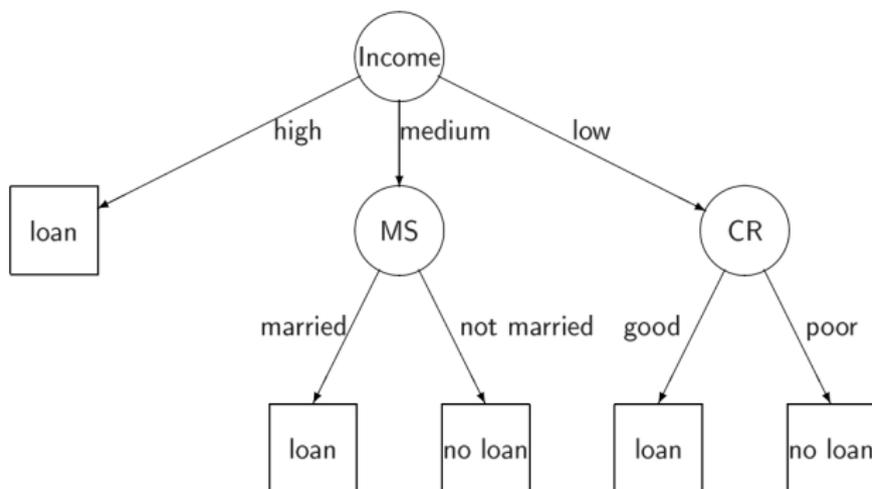
- Прогнозирование осуществляется деревом  $T$ :
  - направленный граф
  - без циклов
  - с одной корневой вершиной

## Определение решающего дерева

- каждому узлу  $t$  соответствуют функции от признаков  $Q_t(x^1, x^2, \dots, x^D)$ 
  - чаще всего проверяется значение отдельного признака  $Q_t(x^1, x^2, \dots, x^D) = x^{i(t)}$
- исходящим ребрам  $r_1(t), \dots, r_{K(t)}(t)$  соответствуют области значений  $S_1(t), \dots, S_{K(t)}(t)$  для  $Q_t(x^1, \dots, x^D)$ , такие что:
  - $S_1(t), \dots, S_{K(t)}(t)$  покрывают все множество значений  $Q_t$  и  $S_i \cap S_j = \emptyset \forall i \neq j, i, j \in \{r_1(t), \dots, r_{K(t)}(t)\}$ .
  - чаще всего  $K(t) = 2$ ,  $S_1 = \{x^{i(t)} \leq \text{threshold}(t)\}$ ,  $S_2 = \{x^{i(t)} > \text{threshold}(t)\}$
  - варианты:  $S_i = \{l_i < x \leq h_i\}$  либо  $S = \{v_k\}$ , где  $\{v_1, v_2, \dots\}$ -множество уникальных значений  $x^{i(t)}$ .

## Определение решающего дерева

- множество вершин делится на:
  - внутренние вершины  $int(T)$ , каждая из которых имеет  $\geq 2$  дочерних вершины
  - внешние вершины  $terminal(T)$ , которые не содержат потомков, но которым сопоставлены прогнозные значения.



## Процесс прогнозирования

- Каждому листу сопоставлен прогноз некоторой константой:
  - классификация: номер класса
  - регрессия: значение.
- Процесс прогнозирования для дерева  $T$ :
  - $t = \text{root}(T)$
  - пока  $t$  не является листом:
    - рассчитать  $Q_t(x)$
    - определить номер множества  $j$ , среди  $S_1(t), \dots, S_{K(t)}(t)$ , куда попадет  $Q_t(x)$ :  $Q_t(x) \in S_j(t)$
    - перейти по ребру  $r_j(t)$  к потомку  $\tilde{t}_j$  :  $t = \tilde{t}_j$
  - вернуть прогноз, сопоставленный листу.

## Спецификация решающего дерева

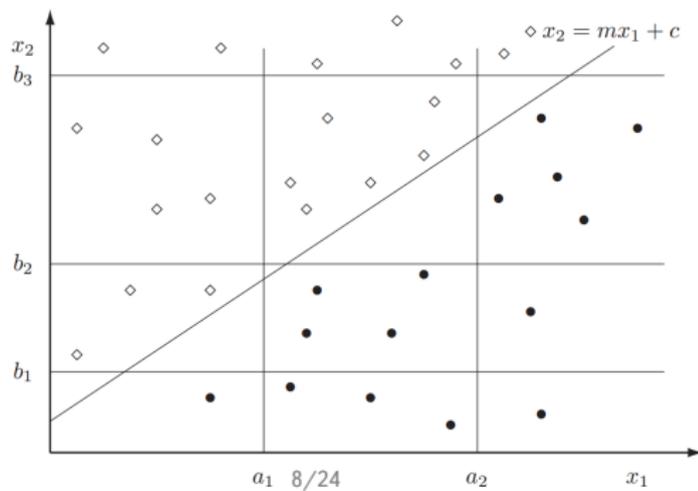
- Определить решающее правило:  $Q_t(x)$ ,  $K(t)$  и  $S_1(t), \dots, S_{K(t)}(t)$
- Определить, критерий останова (когда делаем вершину листом дерева)
- Определить сопоставление прогнозов каждому листу.

# Наиболее распространенное решающее правило

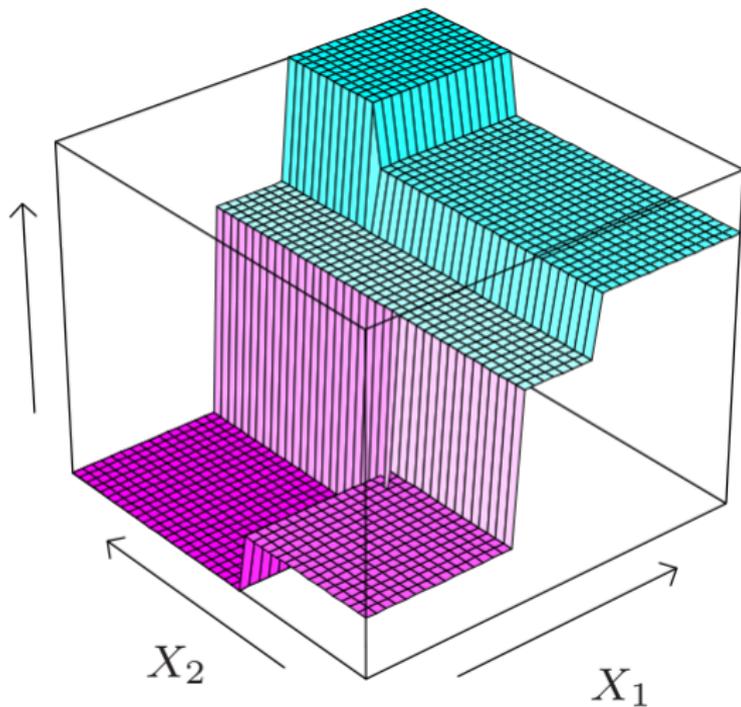
- Чаще всего:
  - в качестве  $Q_t$  берут  $Q_t(x) = x^{i(t)}$
  - $K(t) = 2 \forall t \in \text{int}(T)$ , где  $\text{int}(T)$  - множество внутренних вершин дерева.
  - $S_1(t) = \{x^{i(t)} \leq \text{threshold}(t)\}$ ,  $S_2(t) = \{x^{i(t)} > \text{threshold}(t)\}$ 
    - $\text{threshold}(t) \in \{x_1^{i(t)}, x_2^{i(t)}, \dots, x_N^{i(t)}\}$
    - применимо только для вещественных, порядковых и бинарных признаков
    - в случае дискретных признаков - их необходимо преобразовать в бинарные через one-hot-encoding.

# Анализ решающего правила

- Преимущества:
  - простота
  - интерпретируемость
- Недостатки
  - много узлов может потребоваться, если разделяющая кривая не параллельна осям координат:

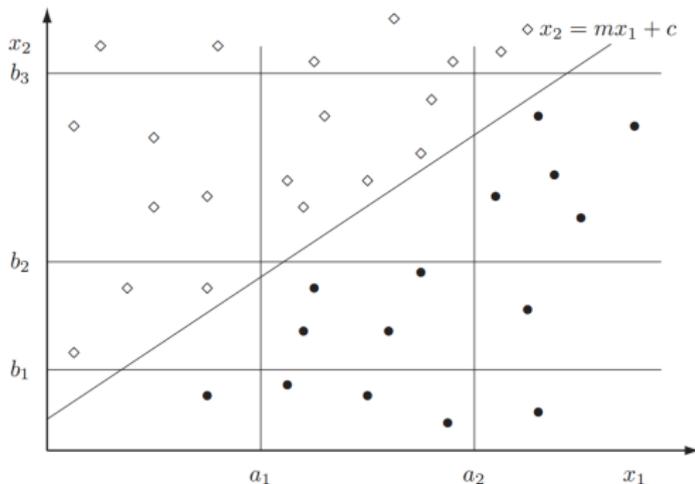


# Кусочно-постоянное решение



## Более общие решающие правила

- Вместо отдельного признака,  $Q(x)$  может иметь более общий вид:  $Q_t(x) = \langle a_t, x \rangle$  или нелинейная функция.
  - тоже кусочно-постоянное решение, но с др. границами
  - менее интерпретируемое решение
  - зато может оказаться значительно меньше вершин!



## Критерий останова

- Противоречие: смещение/дисперсия
  - наиболее разросшиеся деревья - переобучение
  - слишком простые - недообучение
- Подходы к остановке
  - основанные на правиле: сравниваем критерий с порогом
  - основанные на обрезке дерева (pruning)

## Варианты правилых подходов

- Правильный подход: сравниваем критерий с порогом.
- Варианты критерия:
  - глубина дерева
  - количество наблюдений в вершине
  - минимальное количество наблюдений в одной из дочерних вершин
  - делимость классов
  - изменение делимости классов после дробления

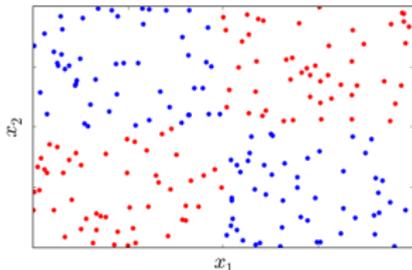
## Обсуждение правилых подходов

Преимущества:

- простота
- эффективность

Недостатки:

- требует спецификации порога
- субоптимальна: последующие дробления могут оказаться лучше, но процесс построения дерева до них не дойдет
  - пример субоптимальности правил (основанных на делимости классов):



## Функция смешанности классов

- Вероятности классов в узле:

$$p(\omega_j | x \in u(t)) = p(\omega_j | t) \approx \frac{N_j(t)}{N(t)}$$

- Функция смешанности (impurity function):

$$I(t) = \phi(p(\omega_1 | t), \dots, p(\omega_C | t))$$

- $\phi(q_1, q_2, \dots, q_C)$  определена для  $q_j \geq 0$  и  $\sum_j q_j = 1$ .
- $\phi$  достигает максимума при  $q_j = 1/C$  для всех  $j$ .
- $\phi$  достигает минимума, когда  $\exists j : q_j = 1, q_i = 0$  для всех  $i \neq j$ .
- $\phi$  симметричная функция от  $q_1, q_2, \dots, q_C$ .

## Типичные функции смешанности классов

- Критерий Джини

- вероятность сделать ошибку, сопоставляя класс случайно с вероятностями  $[p(\omega_1|t), \dots, p(\omega_C|t)]$ :

$$I(t) = \sum_i p(\omega_i|t)(1 - p(\omega_i|t)) = 1 - \sum_i [p(\omega_i|t)]^2$$

- Энтропия

- мера неопределенности дискретной случайной величины

$$I(t) = - \sum_i p(\omega_i|t) \ln p(\omega_i|t)$$

- Ошибка классификации

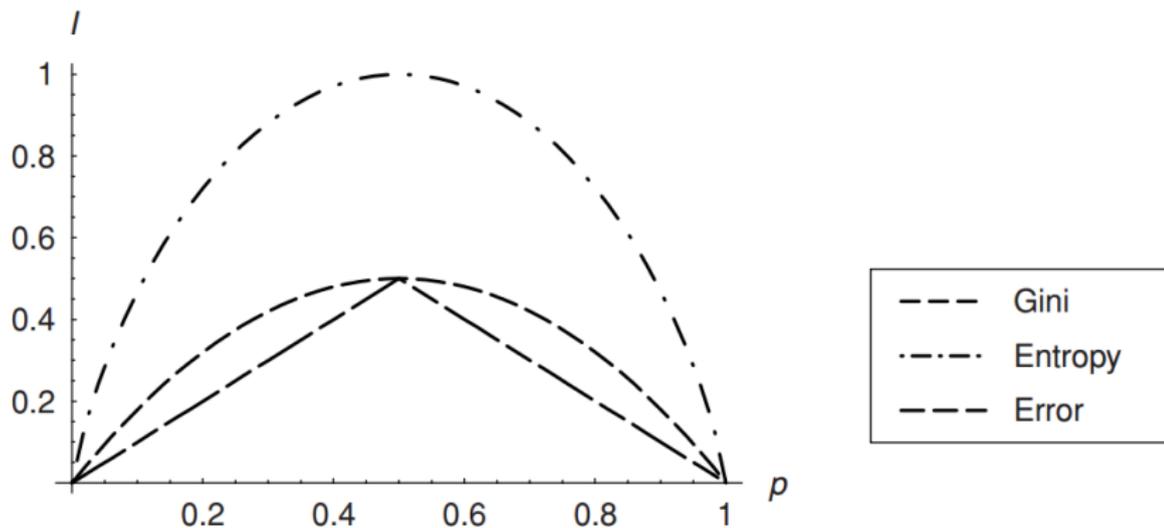
- показывает частоту ошибок при классификации самым частным классом

$$I(t) = 1 - \max_i p(\omega_i|t)$$

## Типичные функции смешанности классов

Функции смешанности в случае 2х классов:

$$p = p(\omega_1|t), \quad 1 - p = p(\omega_2|t).$$



## Критерий ветвления

- Выбрать правило ветвления, максимизирующее снижение смешанности классов:

$$\Delta I(t) = I(t) - \sum_{i=1}^S I(t_i) \frac{N_i(t)}{N(t)}$$

где  $S$  - число ветвей,  $t_1, \dots, t_S$  дочерние вершины для  $t$ ,  $N(t)$  - число наблюдений в  $t$ , а  $N_i(t)$  - число наблюдений класса  $\omega_i$  в  $t$ .

- Если  $I(t)$  - энтропия, то  $\Delta I(t)$  называется *information gain*.

## Назначение классов листовым вершинам

- Регрессия:

$$\hat{y} = \arg \min_{\mu} \sum_{i: x_i \in u(t)} (y_i - \mu)^2 = \frac{1}{N_t} \sum_{i: x_i \in u(t)} y_i,$$

где  $N = |\{x_i : x_i \in u(t)\}|$

- могут использоваться и др. функции потерь.

- Классификация:

- Пусть  $\lambda(\omega_i, \omega_j)$  - цена сопоставления объекта класса  $\omega_i$  классу  $\omega_j$
- Сопоставление класса с минимальной совокупной ценой

$$c = \arg \min_{\omega} \sum_{i: x_i \in u(t)} \lambda(c_i, \omega)$$

- Для  $\lambda(\omega_i, \omega_j) = \mathbb{I}[\omega_i \neq \omega_j]$  будет сопоставляться наиболее частый класс в подвыборке листа.

## CART

- Определим  $R(t) = \frac{M(t)}{N}$ , для листовой вершины  $t \in \tilde{T}$ , где  $\tilde{T}$ -множество листовых вершин для  $T$ .  $M(t)$  - количество ошибок на валидационном множестве, а  $N$ -количество элементов на валидационном множестве.

$$R(T) = \sum_{t \in \tilde{T}} R(t)$$

$$R_\alpha(T) = \sum_{t \in \tilde{T}} R_\alpha(t) = R(T) + \alpha |\tilde{T}|$$

- Условие, когда дерево  $T_t$  с корнем  $t$  имеет такую же цену, как сам корень  $t$ :

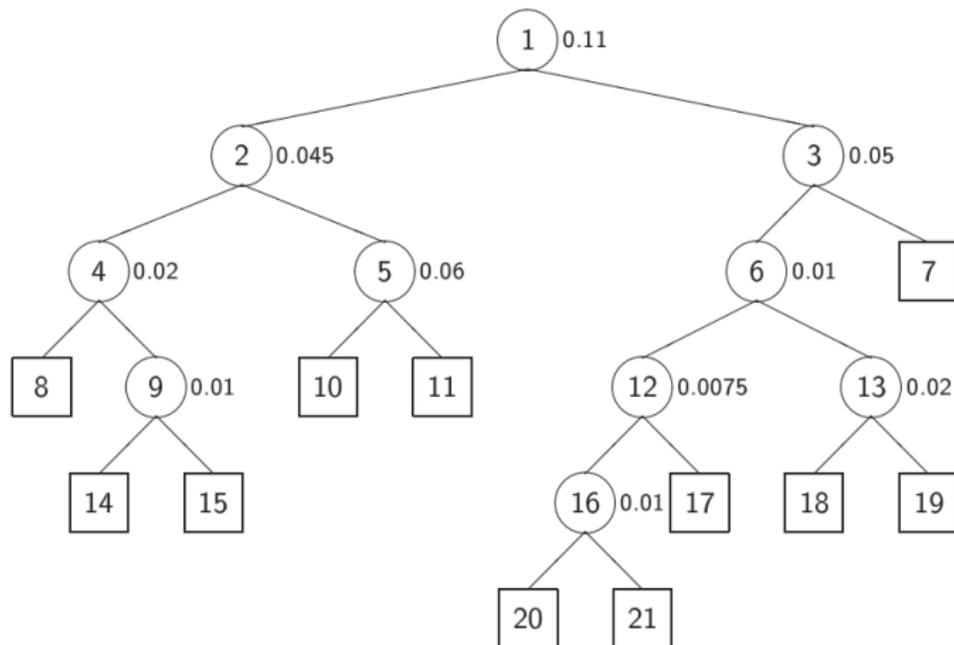
$$\alpha_t = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

## Алгоритм обрезки

- строим дерево, пока классы не становятся полностью разделимыми, получаем дерево  $T$ .
- строим иерархию вложенных поддеревьев  $T_0, T_1, \dots, T_{|T|}$  из  $|T|, |T| - 1, \dots, 1$  вершин, повторяя процедуру:
  - заменяем дерево с наименьшим  $\alpha_t$  его корнем  $t$
  - пересчитываем  $R(T_t), R_\alpha(T_t)$  и  $\alpha_t$  для всех предков  $t$ .
- Для каждого вложенного дерева  $T_0, T_1, \dots, T_{|T|}$  вычисляем ф-ции цены  $R(T_0), R(T_1), \dots, R(T_{|T|})$ .
- Выбираем то дерево  $T_i$ , дающее наименьшую цену

$$i = \arg \min_i R(T_i)$$

## Пример



## Пример

	$\alpha_k$	$ \tilde{T}^k $	$R(T^k)$
1	0	11	0.185
2	0.0075	9	0.2
3	0.01	6	0.22
4	0.02	5	0.25
5	0.045	3	0.34
6	0.05	2	0.39
7	0.11	1	0.5

## Дополнительно

Если нужный признак отсутствует:

- усреднение прогнозов по каждой ветви с весами  $\propto$  числу наблюдений идущих в ветви.
- surrogate splits
- заполнение пропущенных значений
  - среднее
  - прогноз по остальным признакам
- отдельная ветвь «значение пропущено».

Важность признаков, рассчитанная по дереву

## Анализ решающих деревьев в целом

- Преимущества решающих деревьев:
  - простота
  - интерпретируемость
  - встроенный отбор признаков
  - работает одновременно с дискретными и непрерывными признаками
  - прогноз инвариантен к монотонным преобразованиям признаков для  $Q_t(x) = x^{i(t)}$ 
    - не нужна нормализация
- Недостатки решающих деревьев:
  - если разделяющая кривая не параллельна осям координат, то может потребоваться много вершин
  - субоптимальность (пример XOR)
  - нет онлайнности - для новых наблюдений требуется полная перестройка всего дерева.