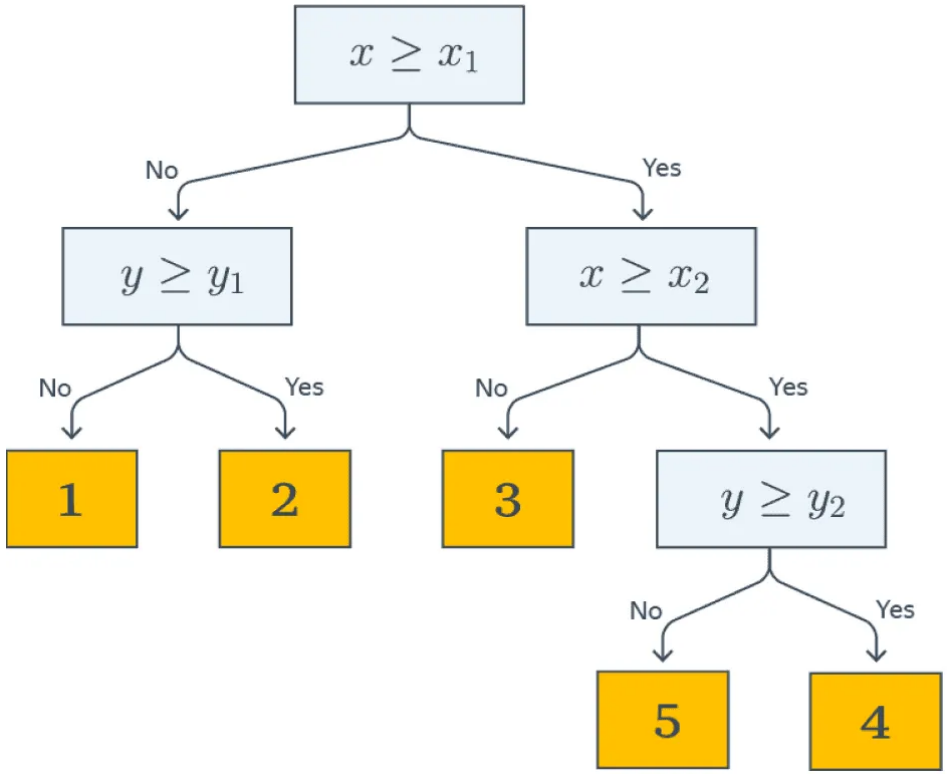
Решающие деревья

Решающие деревья

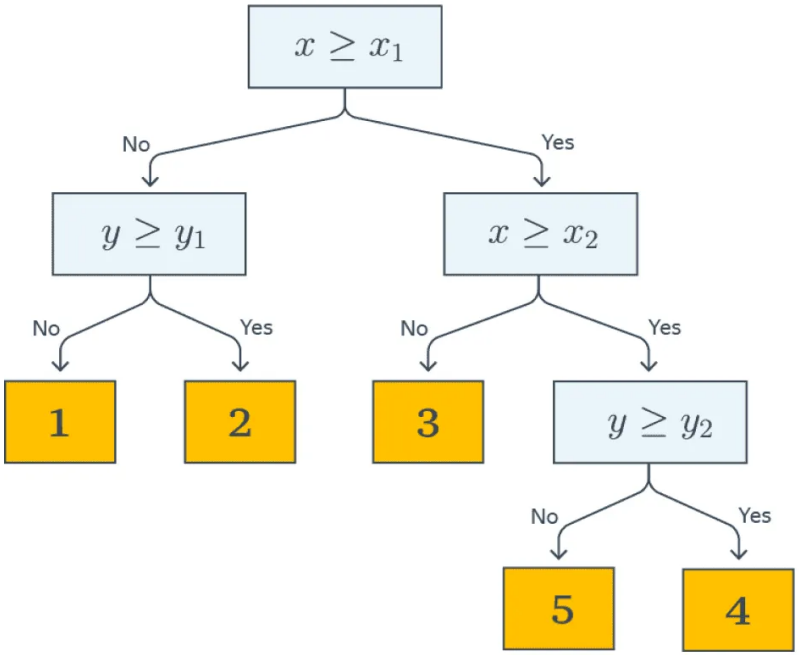
Предсказывают значение целевой переменной с помощью применения последовательности простых решающих правил[1].

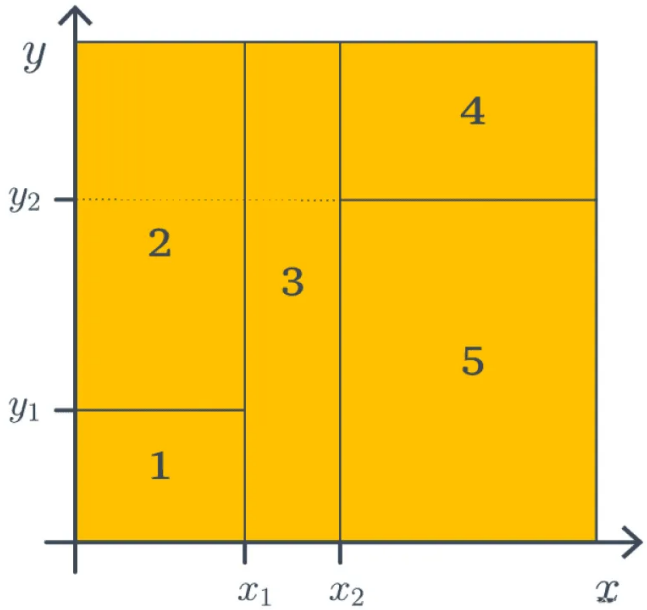


[1] https://education.yandex.ru/handbook/ml/article/reshayushchiye-derevya

Решающие деревья

Разбивают пространство на области.





Это задача регрессии или классификации?

Условия в листах

- **Одномерные предикаты**

$$B_{j,t}(x_i) = [x_{i,j} < t]$$

- Линейные предикаты

$$B_{j,t}(x_i, w) = [w^T x_i < t]$$

- Метрические

$$B_{j,t}(x_i, w) = [\rho(x_i, , w) < t]$$

Почти всегда используются одномерные предикаты.

Ответ листа

Пусть в лист попало множество объектов U из обучающей выборки. Какое значение целевой переменной для них назначать модели:

- **число** (метка самого частого класса, среднее, медиана);
- вектор (оценка дискретного распределения вероятностей классов);
- модель от данных в листе.

Почти всегда используется константа — одинаковое предсказание для всех объектов, попавших в лист.

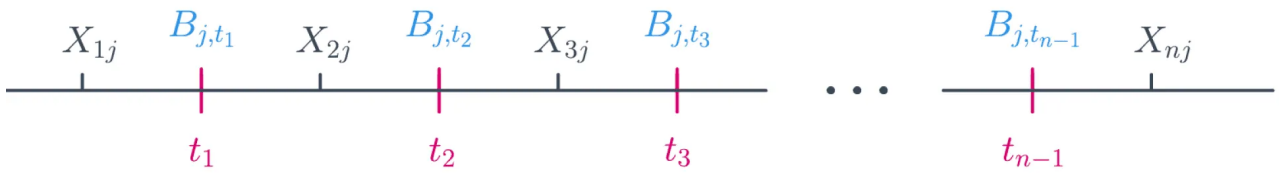
Построение дерева

Пусть дана обучающая выборка (X, y), где $X \in \mathbb{R}^{N \times D}$ и задана функция потерь $L(f, X, y)$. Разбиение

$$B_{j,t}(x_i) = [x_{i,j} < t].$$

Оптимальное из $(N-1)D$ вариантов разбиение

$$(j^*, t^*) = \arg\min_{j,t} L(B_{j,t}, X, y).$$



Поиск оптимального с точки зрения качества на обучающей выборке дерева минимальной глубины — NP-полная задача.

Критерий информативности

Дана функция потерь $L(y, \hat{y})$ и лист, в который попало множество объектов U. Тогда информативность (impurity)

$$H(U) = \min_{\hat{y}} \frac{1}{|U|} \sum_{(x_i, y_i) \in U} L(y_i, \hat{y}_i)$$

Поделим вершину на две L и R. Тогда

$$H(L \cup R) = \min_{\hat{y}^L} \frac{1}{|L|} \sum_{(x_i, y_i) \in L} L(y_i, \hat{y}_i^L) + \min_{\hat{y}^R} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, \hat{y}_i^R)$$

$$H(L \cup R) = \frac{|L|}{|U|} H(L) + \frac{|R|}{|U|} H(R)$$

Критерий ветвления

Насколько улучшится некоторая финальная метрика качества дерева в случае, если получившиеся два листа будут терминальными, по сравнению с ситуацией, когда сама исходная вершина — это лист:

$$B_{j,t}(U) = |U|H(U) - (|L|H(L) + |R|H(R))$$

Максимизируем для нахождения наилучшего сплита.

Рекурсивное построение дерева

1. Вычисляем критерий информативности текущего множества точек в листе
2. Для всех факторов и всех возможных разбиений вычисляем изменения критерия информативности
3. Выбираем лучшее
4. Проверяем критерий остановки, если он есть
5. Делаем прунинг, если он предусмотрен

Критерии информативности. Регрессия

1. MSE:

$$H(U) = \frac{1}{|U|} \sum_{i \in U} \left(y_i - \frac{1}{|U|} \sum_{j \in U} y_j \right)^2$$

Оценка значения в каждом листе — среднее, а лучший сплит минимизирует сумму дисперсий в листьях.

2. MAE:

$$H(U) = \frac{1}{|U|} \sum_{i \in U} (y_i - med(Y))$$

Оценка значения в каждом листе — медиана.

Критерии информативности. Классификация

Пусть $p_1, \ldots, p_k$ — доли объектов классов $1, \ldots, K$ в U.

1. Доля ошибок. Функция потерь $L(y, \hat{y}) = \mathbb{I}[y_i \neq \hat{y}]$.

$$H(U) = 1 - p_{max}$$

2. Энтропийный критерий

$$H(U) = -\sum_{k=1}^{K} p_k \ln p_k$$

3. Критерий Джини. Функция потерь — метрика Бриера (MSE от вероятностей)

$$H(U) = -\sum_{k=1}^{K} p_k(1 - p_k)$$

Критерии останова

1. Максимальная глубина дерева
2. Минимальное число примеров в листе/узле
3. Максимальное количество листьев в дереве
4. Минимальный размер изменения критерия информативности

Pre-pruning (early-stoping): Ограничиваем рост дерева до того как оно построено. Если в какой-то момент информативность признаков в разбиении меньше порога – не разбиваем вершину

Post-prunning: упрощаем дерево после того как дерево построено

Признаки

- Бинарные — готовый предикат
- Числовые
- Категориальные — хотим научиться упорядочивать значения, чтобы работать с ними так же, как с обычными числами (делить на больше/меньше)

Способны обрабатывать пропуски в данных.

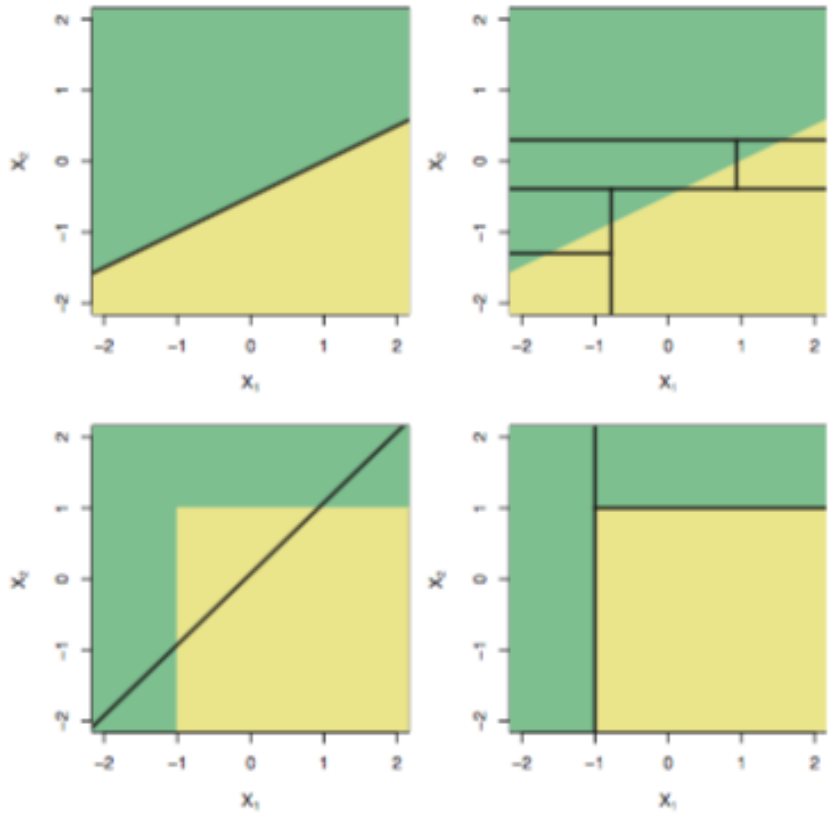
Свойства

- Кусочно-постоянная аппроксимация целевой зависимости
- Линеиная комбинация, произведение, степень деревьев — дерево
- Дерево деревьев — дерево
- Любое дерево можно представить как бинарное
- Изменчивость при изменениях обучающей выборки
- Способно идеально приблизить обучающую выборку
- За пределами обучающей выборки делает константные предсказания.

Плюсы и минусы

+ Простота, интерпретируемость модели
+ Встроенный отбор признаков
+ Работает с дискретными и непрерывными признаками, инвариантен к монотонным преобразованиям
+ Работают для задач с несколькими выходами
– Склонность в переобучению
– Сложность модели в случае разделяющей полосы, не параллельной осям координат
– Необходимость переобучения всего дерева при добавлении новых объектов
– Могут плохо работать при несбалансированной выборке

Линейные модели vs деревья



Деревья используют взаимное положение объектов вдоль каждой из осей-признаков, не используют расстояния.