Машинное обучение ФБМФ

Лекция 4

Решающие деревья

Свойства линейных моделей

- Легко обучаются с помощью градиентного спуска
- Востанавливают только простые зависимости

  мало степеней свободы:

  обычно число параметров $\approx$ количество признаков
- Не всегда отражают то, как люди принимают решения.

  Иногда не логично расставлять коэффициенты перед признаками.

Как люди принимают решения?

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Общий случай

Решающее дерево:

- Бинарное дерево.
- В каждой вершине записано некоторое условие.
- В зависимости от условия идем в правую или левую вершину.
- В листьях дерева — предсказания.

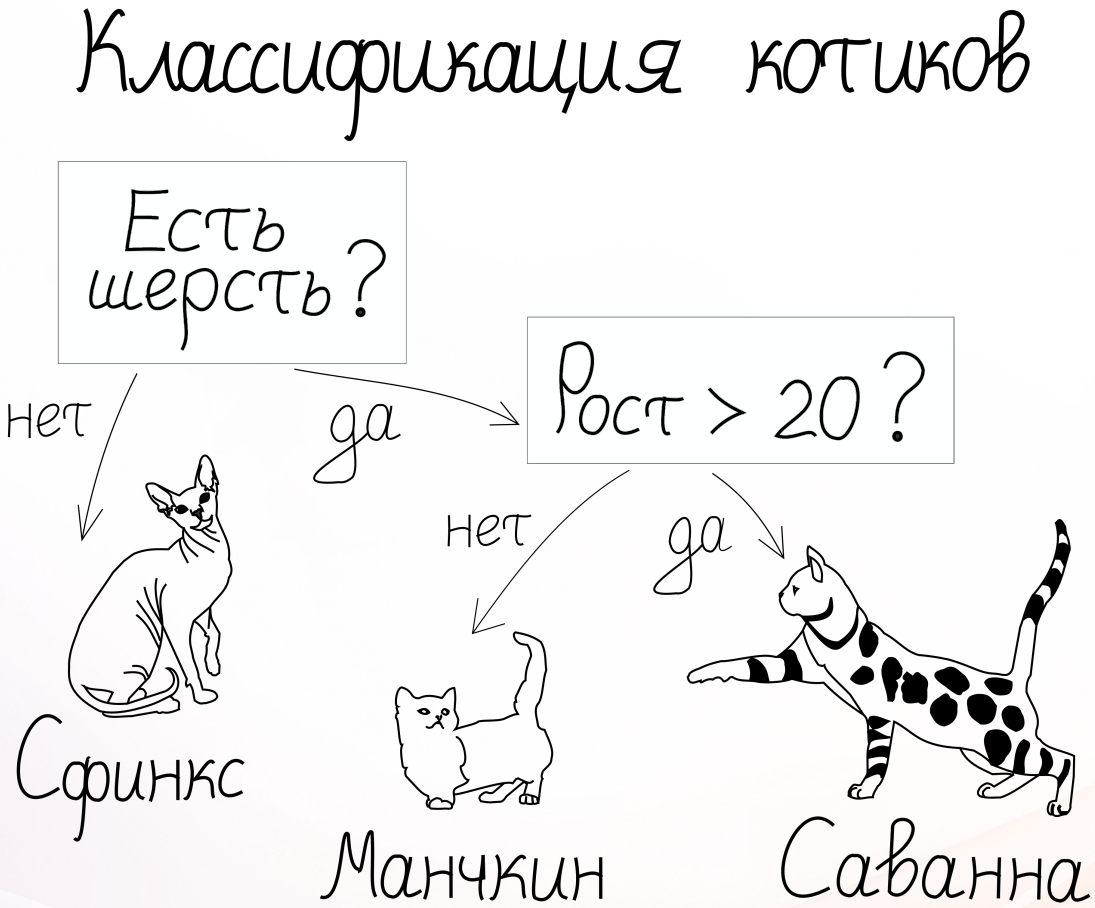
Общий случай

Решающее дерево:

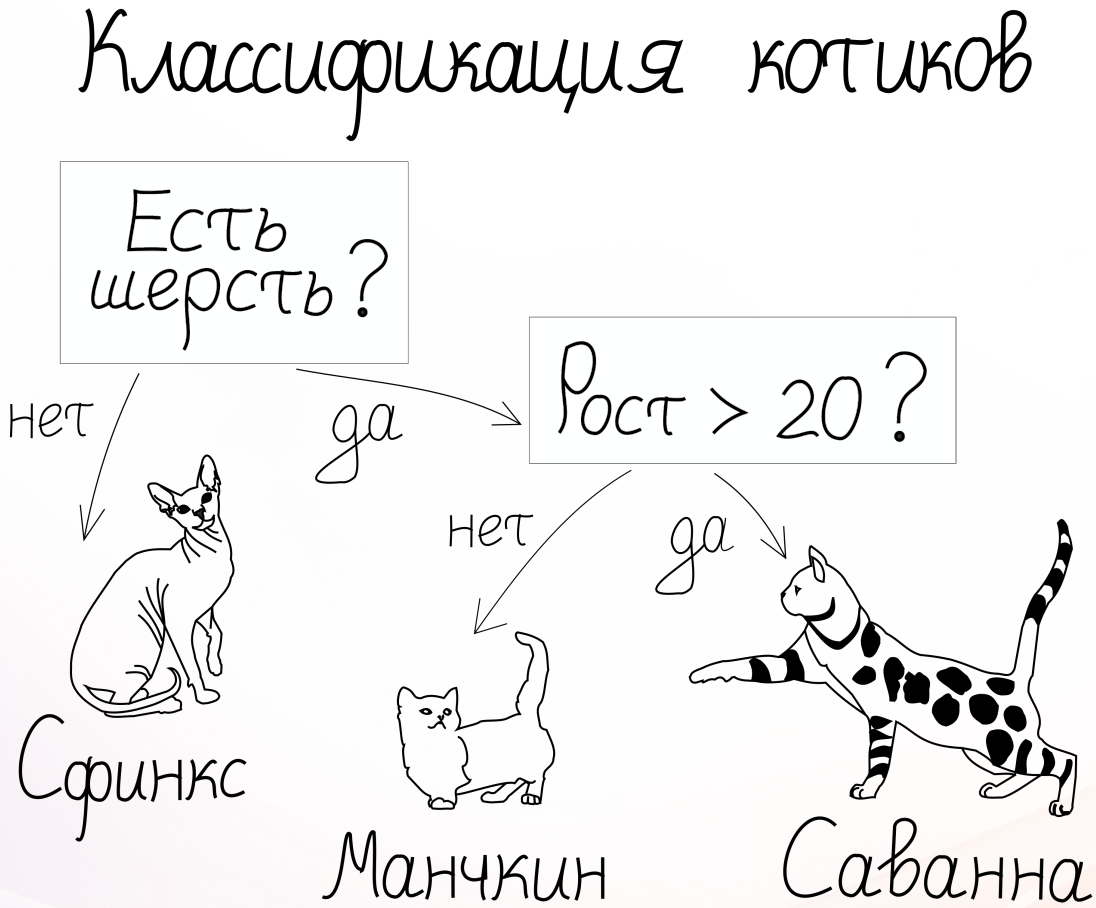
- Бинарное дерево.
- В каждой вершине записано некоторое условие.
- В зависимости от условия идем в правую или левую вершину.
- В листьях дерева — предсказания.

Замечание: Существуют и не бинарные решающие деревья, однако в основном используются именно бинарные

Пример

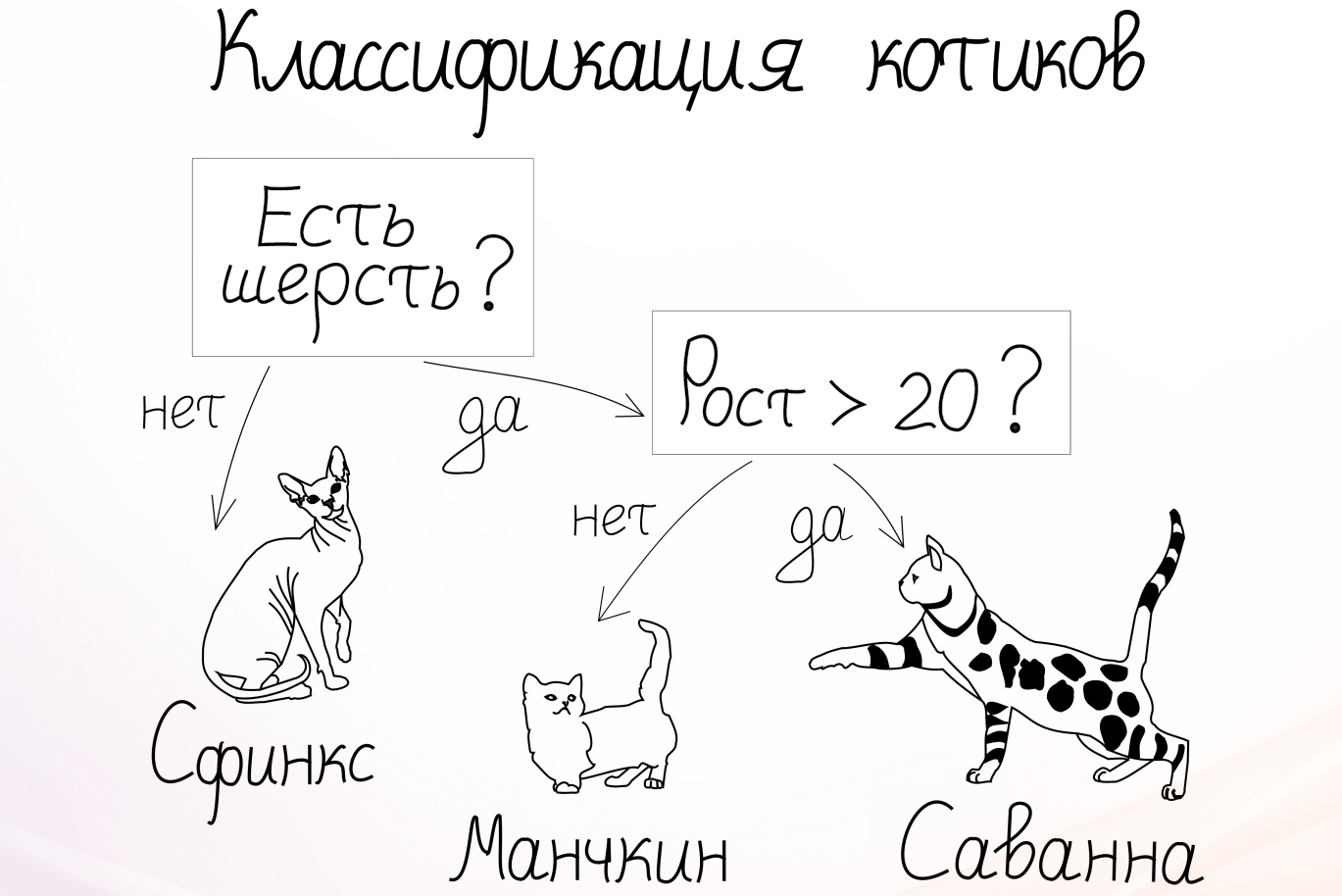

Классификация котиков
Есть шерсть?
нет
да
Рост > 20?
нет
да
Сфинкс
Манчкин
Саванна

Пример



котик	порода	рост	шерсть
?	?	50 см	да

Пример



котик	порода	рост	шерсть
	Саванна	50 см	да

Какие условия в вершинах?

Пусть $x^{(j)}$ - j-ый признак объекта.

t - некоторый порог.

Самый популярный подход - делать разбиение в вершине по правилу вида $I\{x^{(j)} < t\}$.

Какие условия в вершинах?

Пусть $x^{(j)}$ - j-ый признак объекта.

t - некоторый порог.

Самый популярный подход - делать разбиение в вершине по правилу вида $I\{x^{(j)} < t\}$.

Оптимальные значения t и j подбираются по некоторому критерию (об этом позже).

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Строим бинарное дерево по следующему принципу:

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Строим бинарное дерево по следующему принципу:

Начало: один лист с меткой $\overline{Y}$, к нему относятся все объекты.

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Строим бинарное дерево по следующему принципу:

Начало: один лист с меткой $\overline{Y}$, к нему относятся все объекты.

Деление листа: Пусть $X_{leaf} \subset X$ — множество объектов в листе.

Принцип деления: наилучшее приближение двумя константами в дочерних листах:

$$\sum_{x_i \in X_l} (Y_i - y_l)^2 + \sum_{x_i \in X_r} (Y_i - y_r)^2 \longrightarrow \min_{X_l, X_r \,:\, X_l \sqcup X_r = X_{leaf}}$$

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Строим бинарное дерево по следующему принципу:

Начало: один лист с меткой $\overline{Y}$, к нему относятся все объекты.

Деление листа: Пусть $X_{leaf} \subset X$ — множество объектов в листе.

Принцип деления: наилучшее приближение двумя константами в дочерних листах:

$$\sum_{x_i \in X_l} (Y_i - y_l)^2 + \sum_{x_i \in X_r} (Y_i - y_r)^2 \longrightarrow \min_{X_l, X_r \,:\, X_l \sqcup X_r = X_{leaf}}$$

Какие y_l и y_r выбрать для наилучшего приближения по MSE?

Регрессионное дерево

Пусть имеем выборку $(x_1, Y_1), ..., (x_n, Y_n)$.

Строим бинарное дерево по следующему принципу:

Начало: один лист с меткой $\overline{Y}$, к нему относятся все объекты.

Деление листа: Пусть $X_{leaf} \subset X$ — множество объектов в листе.

Принцип деления: наилучшее приближение двумя константами в дочерних листах:

$$\sum_{x_i \in X_l} (Y_i - y_l)^2 + \sum_{x_i \in X_r} (Y_i - y_r)^2 \longrightarrow \min_{X_l, X_r \,:\, X_l \sqcup X_r = X_{leaf}}$$

Какие y_l и y_r выбрать для наилучшего приближения по MSE?

$$y_l = \frac{1}{|X_l|} \sum_{x_i \in X_l} y_i, \qquad y_r = \frac{1}{|X_r|} \sum_{x_i \in X_r} y_i \quad - \text{метки в новых листах}$$

Регрессионное дерево

Как разделить выборку X_m *на* X_l *и* X_r*?*

Регрессионное дерево

Как разделить выборку X_m на X_l и X_r?

Пусть $x^{(j)}$ — j-ый признак x.

Обычно правило выглядит так:

- $x^{(j)} < t \Rightarrow$ объект x попадает в левое поддерево X_l.
- $x^{(j)} \geq t \Rightarrow$ объект x попадает в правое поддерево X_r.

Регрессионное дерево

Как разделить выборку X_m на X_l и X_r?

Пусть $x^{(j)}$ — j-ый признак x.

Обычно правило выглядит так:

- $x^{(j)} < t \Rightarrow$ объект x попадает в левое поддерево X_l.
- $x^{(j)} \geq t \Rightarrow$ объект x попадает в правое поддерево X_r.

Для нахождения оптимальных j и t перебираются

все возможные их значения.

Регрессионное дерево

Оценка отклика для объекта — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Регрессионное дерево

Оценка отклика для объекта — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

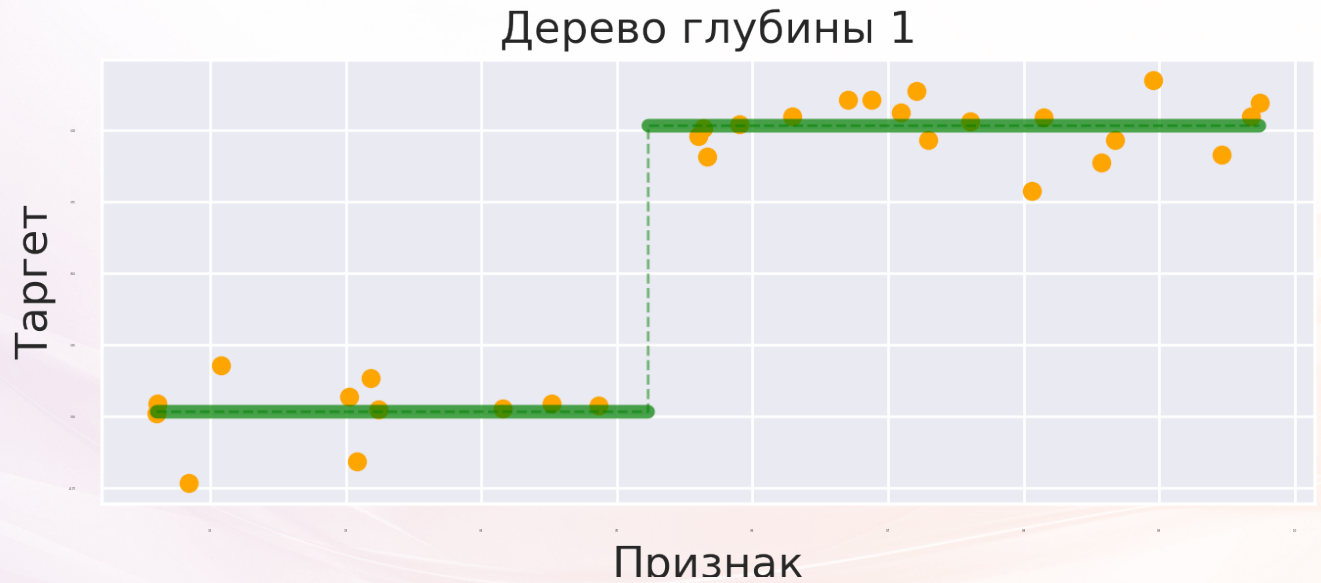
Пример 1:

Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 1:



Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

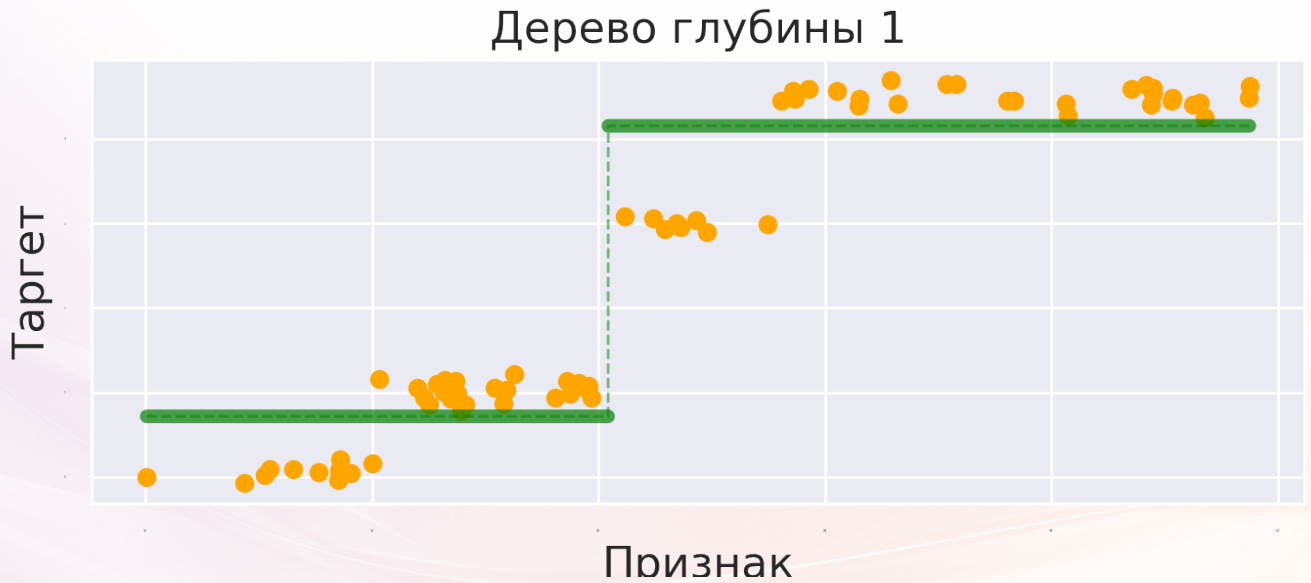
Пример 2:

Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 2:

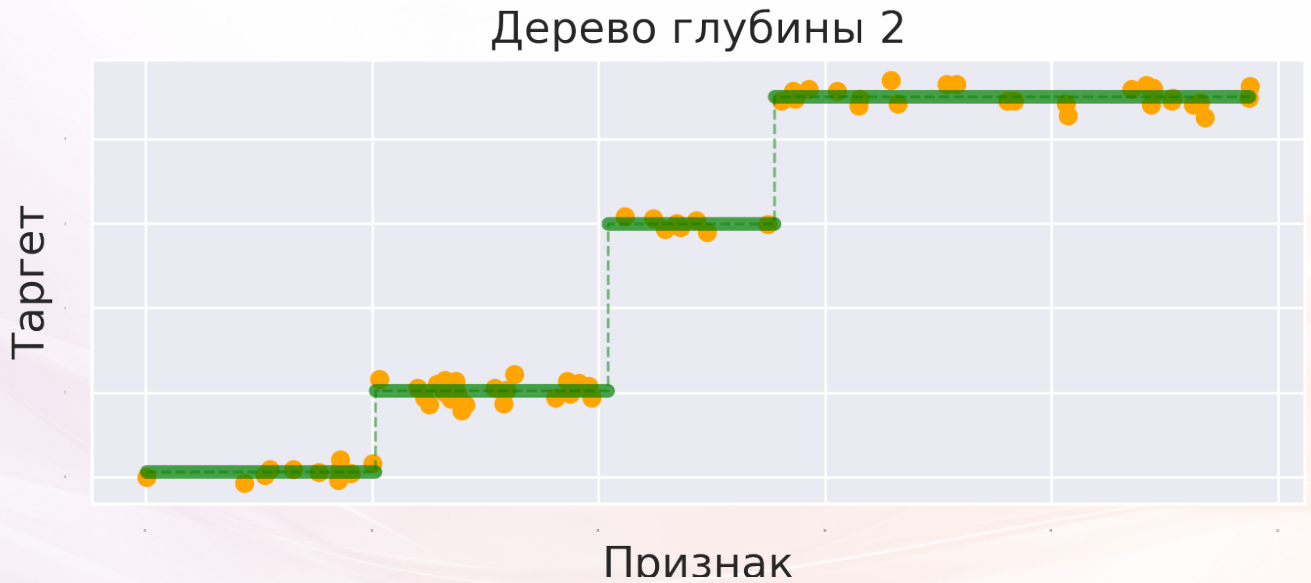


Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 2:

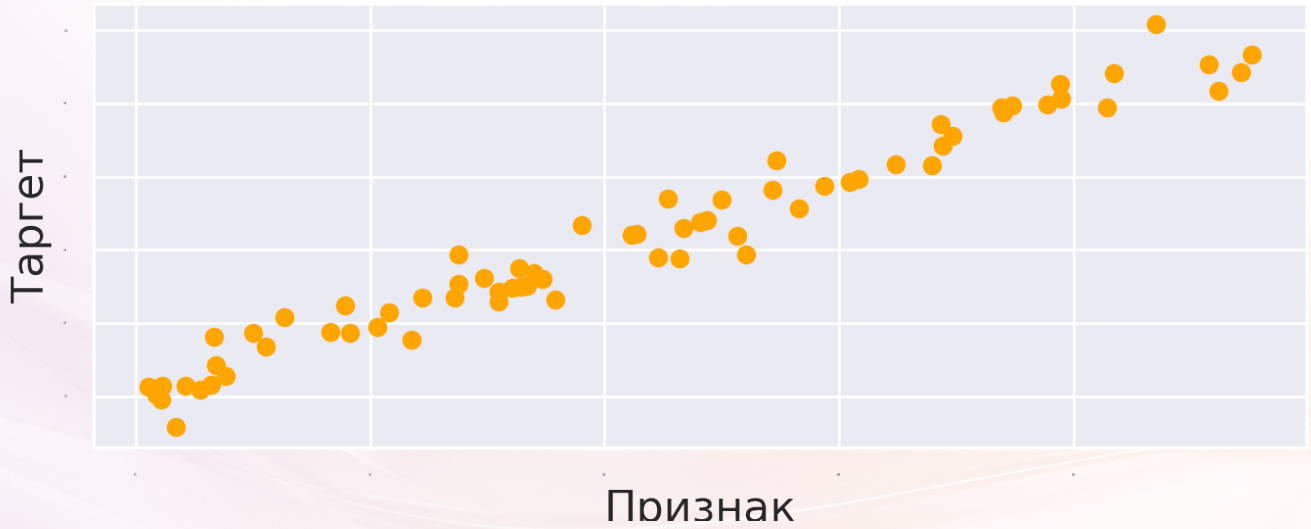


Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 3:

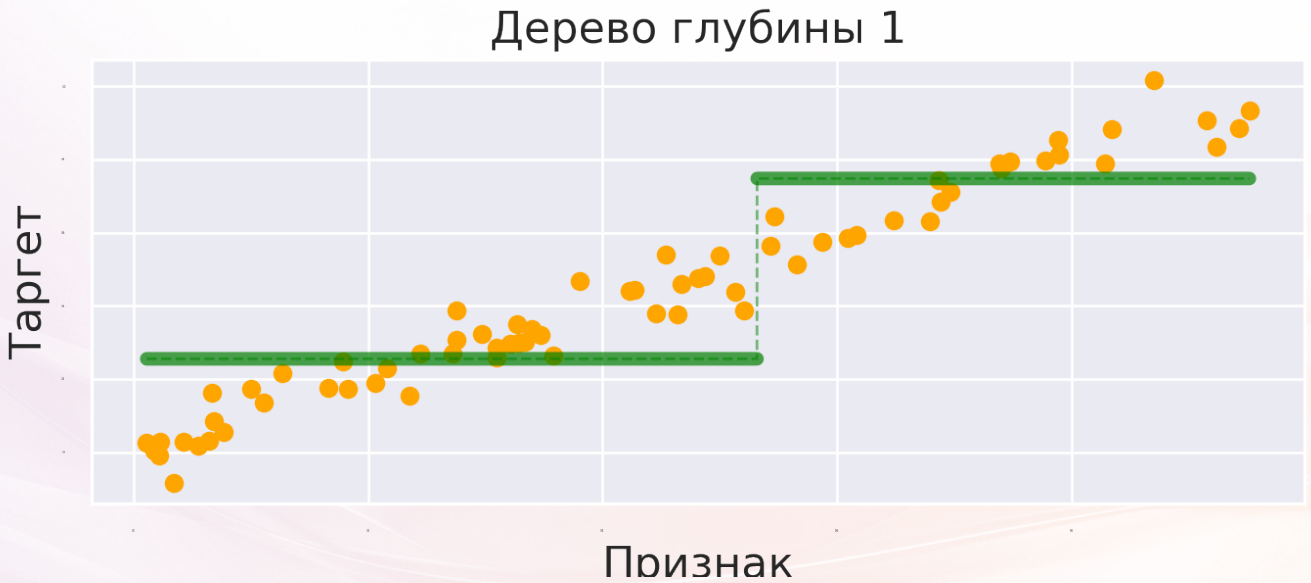


Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 3:

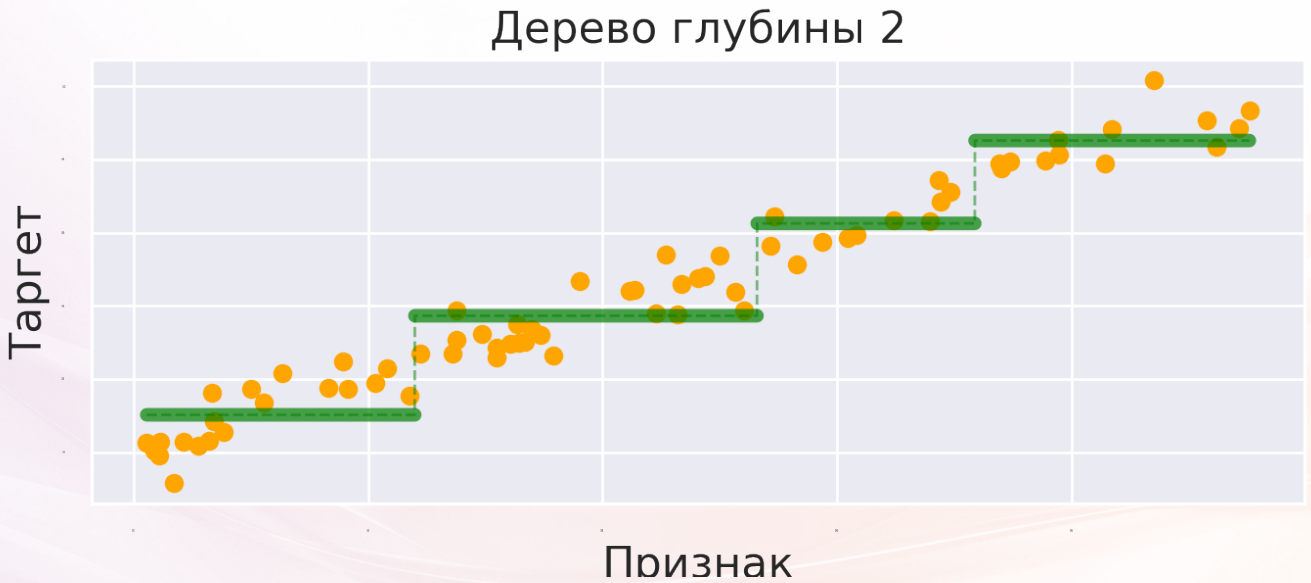


Регрессионное дерево

Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

Пример 3:

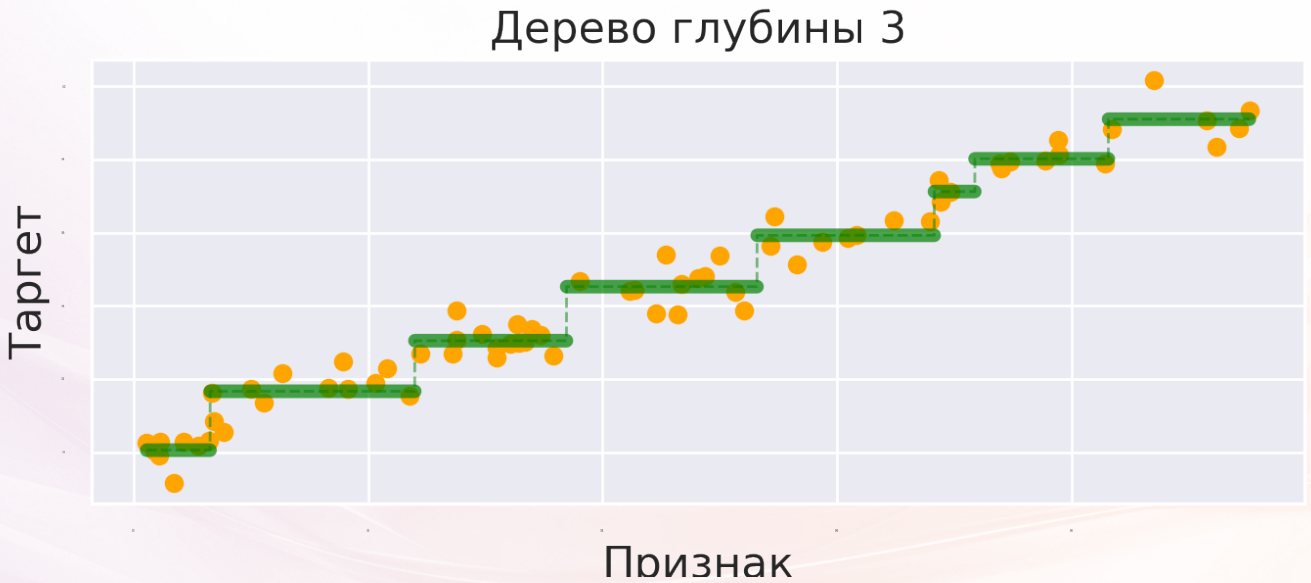


Регрессионное дерево

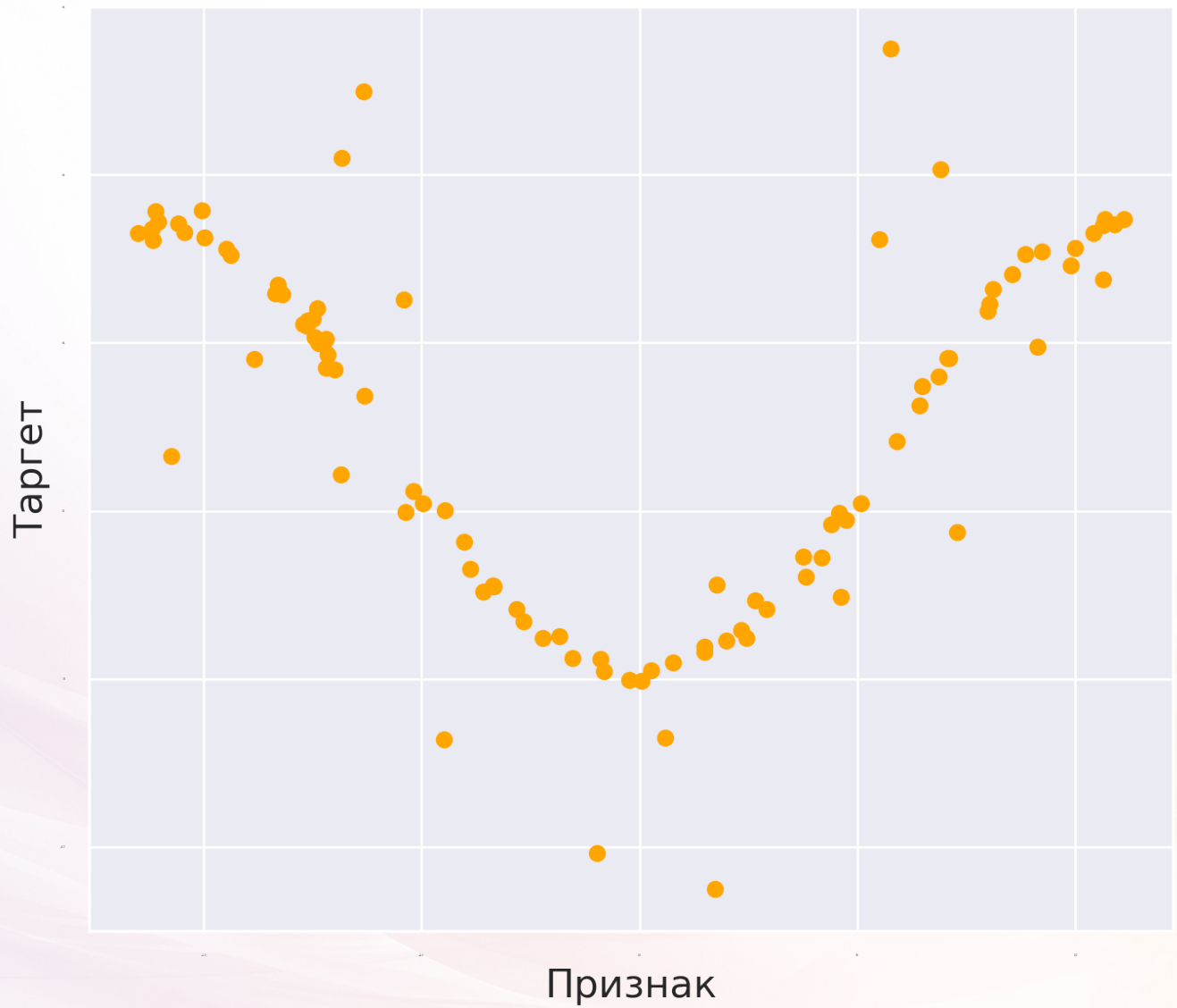
Оценка отклика — метка листа, в который попадет объект.

Т.е. строится кусочно-постоянная функция.

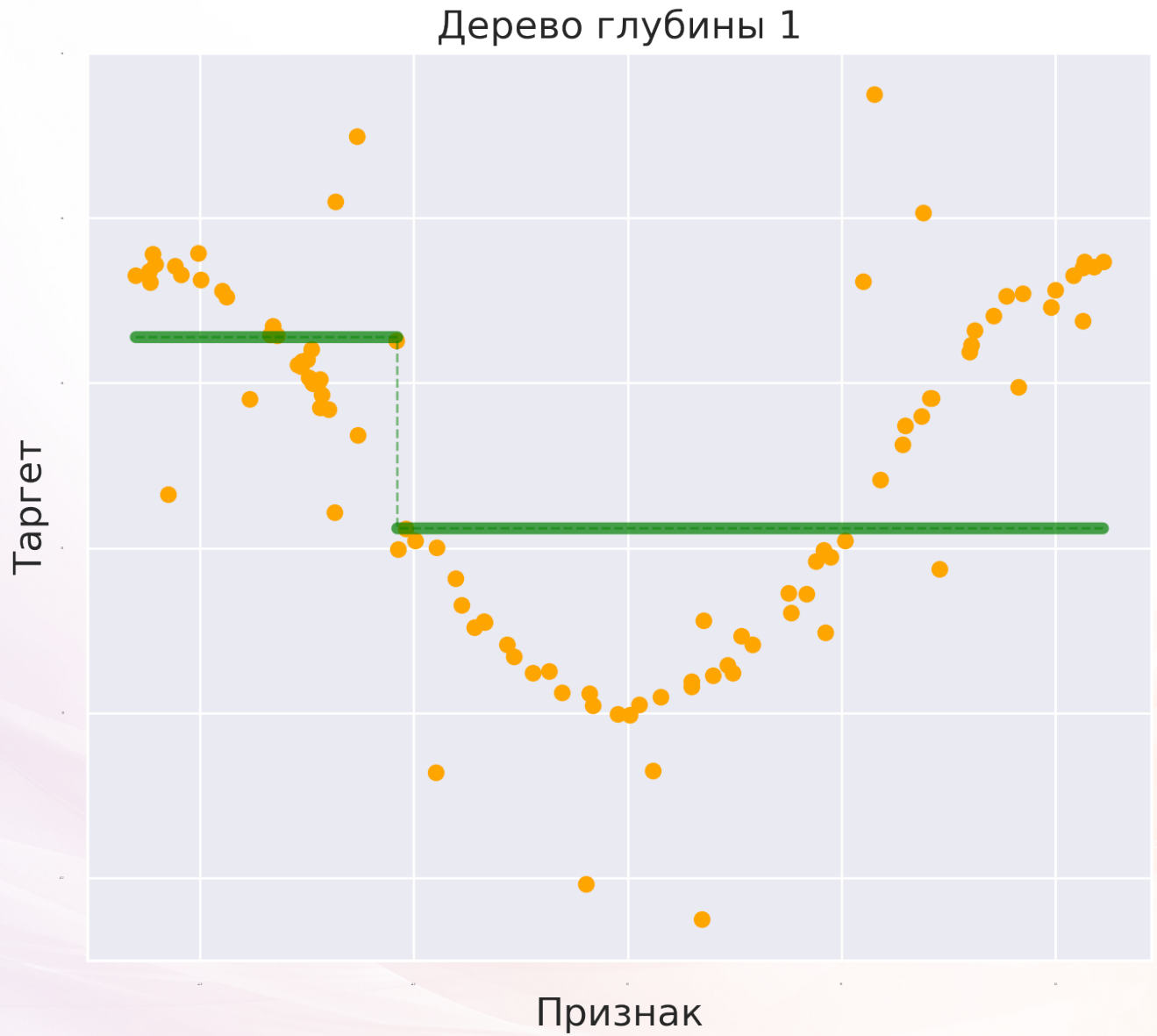
Пример 3:



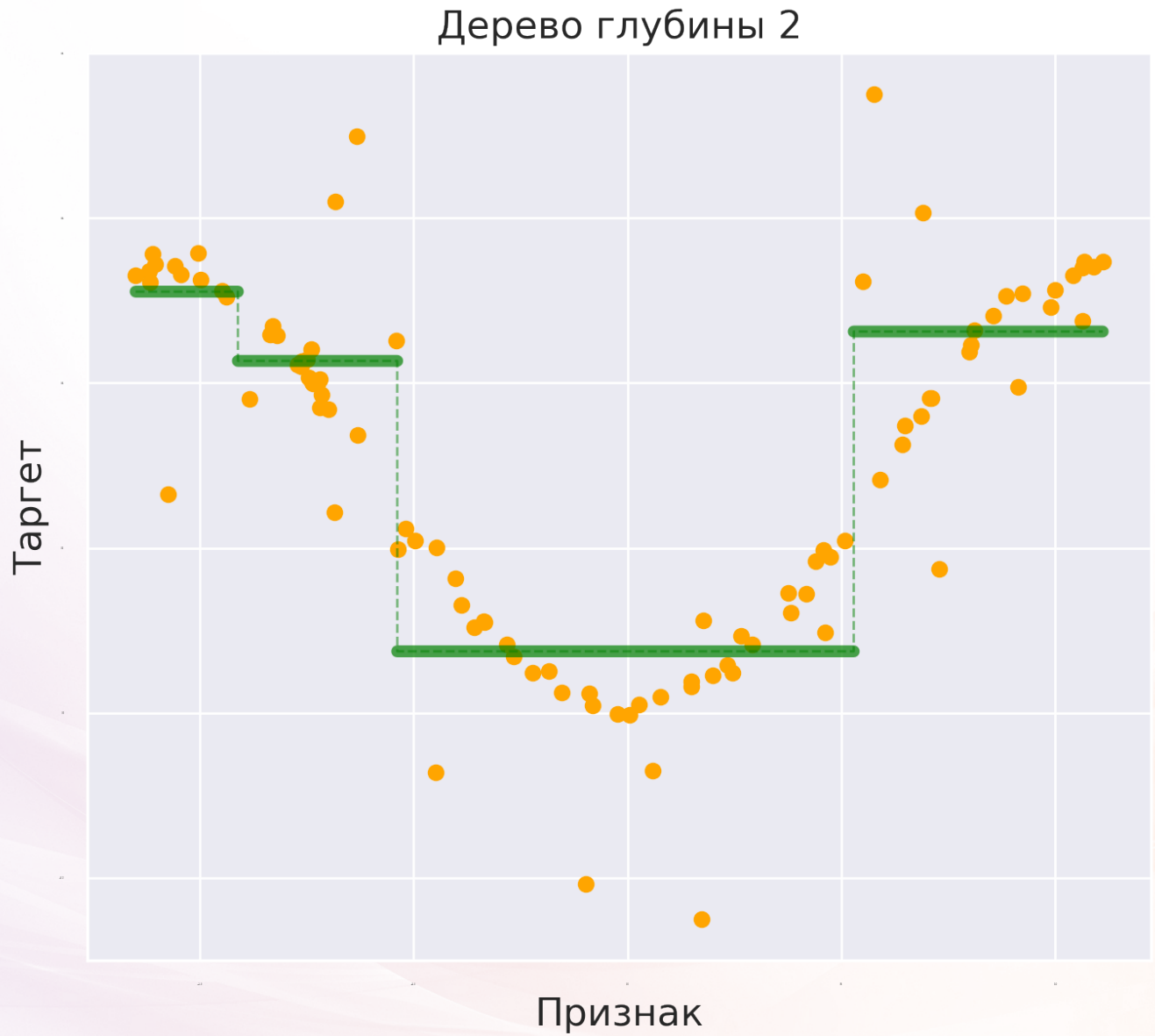
Регрессионное дерево и выбросы



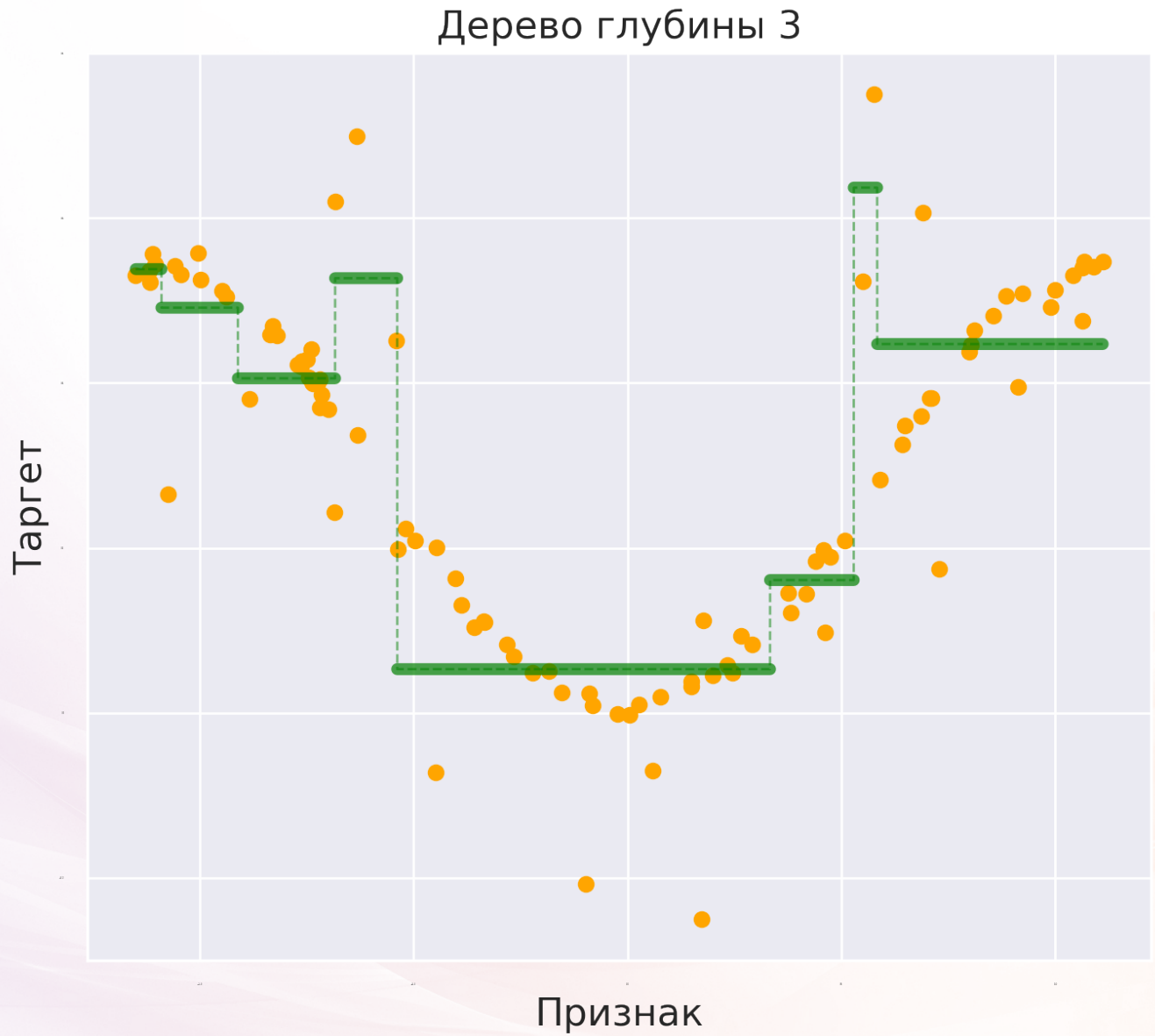
Регрессионное дерево и выбросы



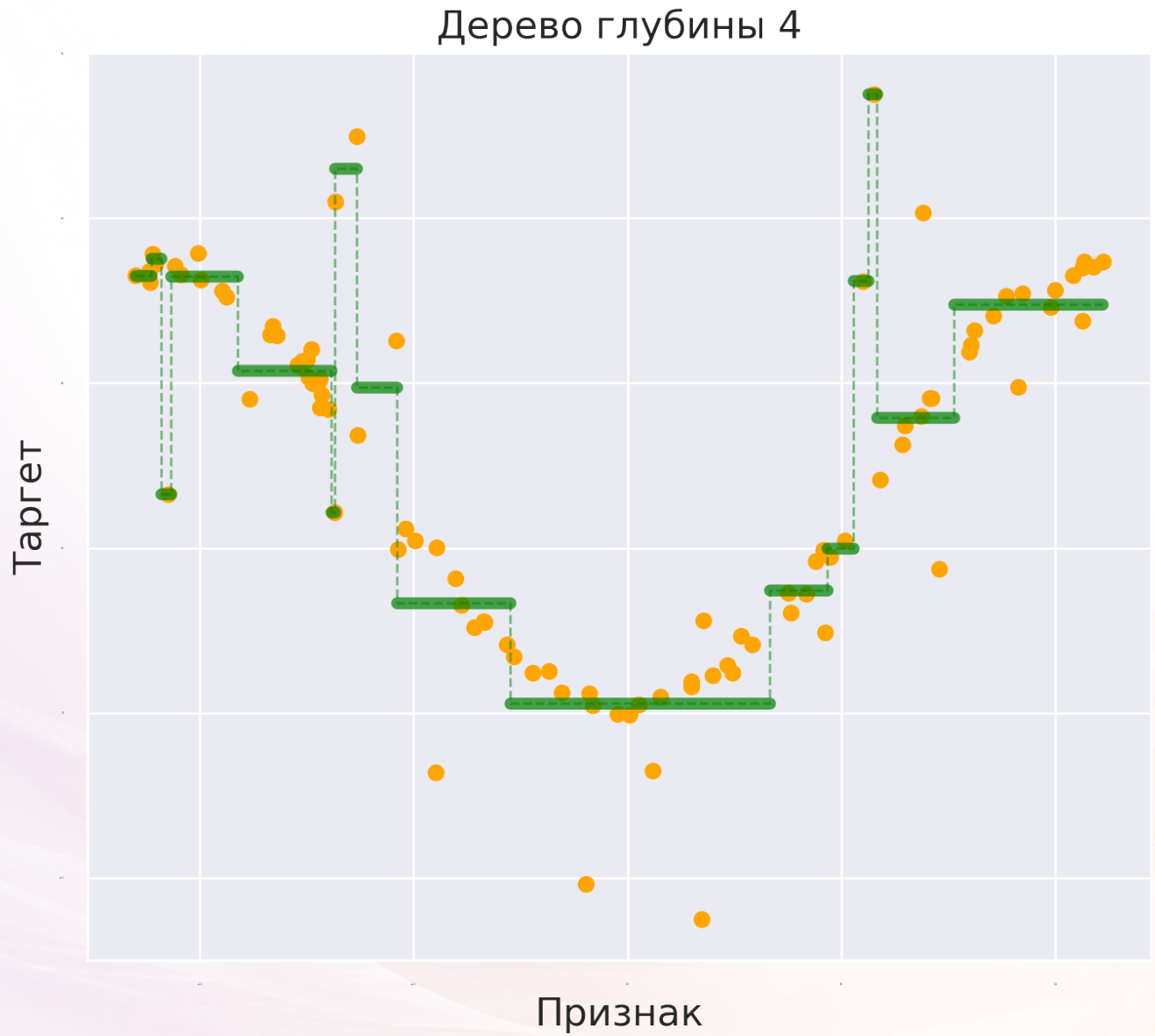
Регрессионное дерево и выбросы



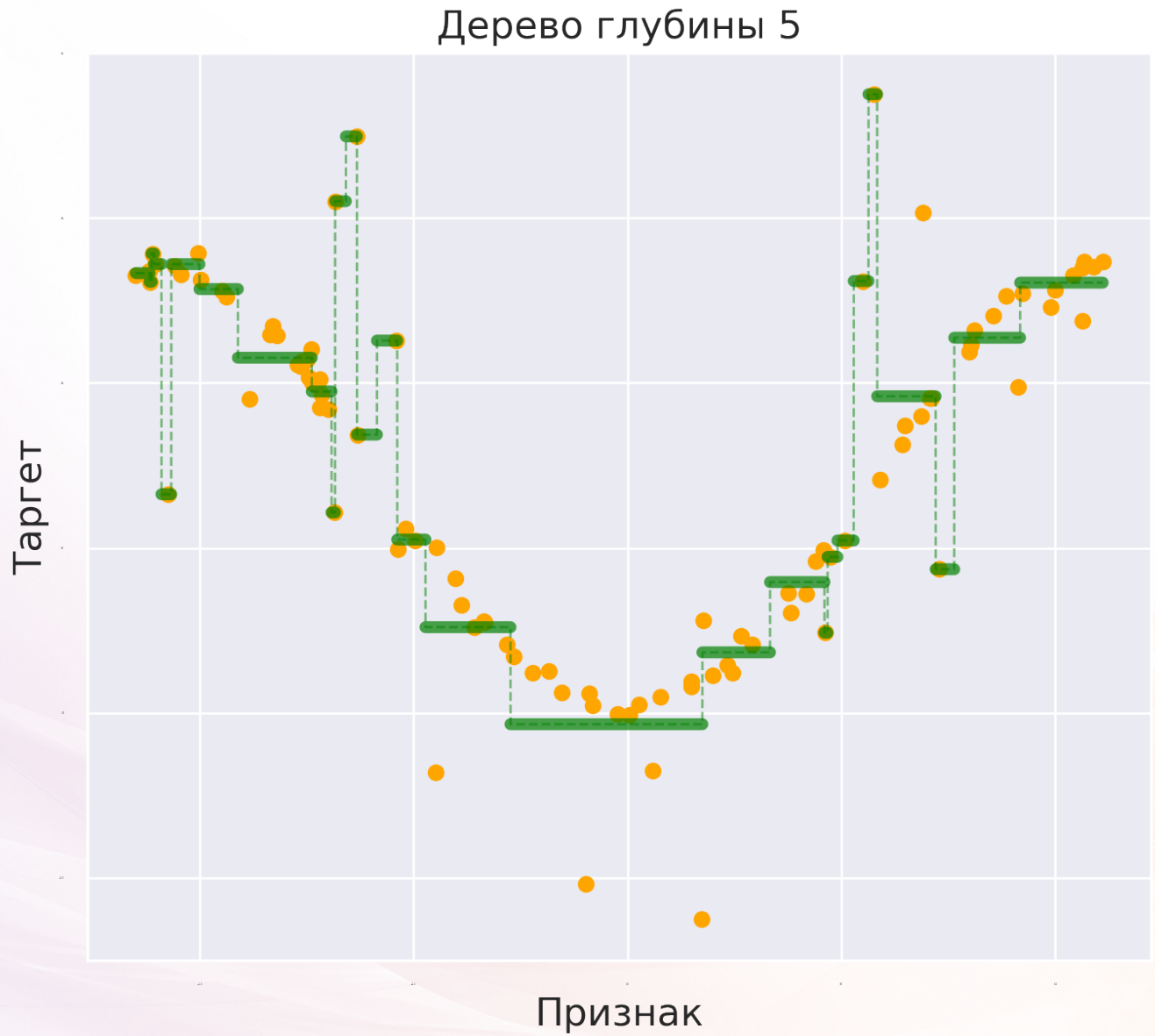
Регрессионное дерево и выбросы



Регрессионное дерево и выбросы



Регрессионное дерево и выбросы



Переобучение

Утверждение

Для любой обучающей выборки можно построить решающее дерево с нулевой ошибкой на обучении.

Переобучение

Утверждение

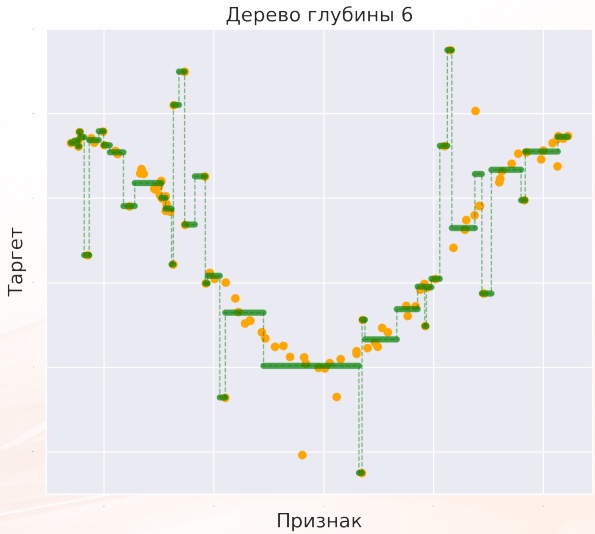
Для любой обучающей выборки можно построить решающее дерево с нулевой ошибкой на обучении.

Доказательство:

Построим решающее дерево, в котором каждый лист содержит только один объект.

Ответ в листе определятся только этим одним объектом.

$\Rightarrow$ предсказание для обучающей выборки не содержит ошибок.



Решающие деревья

Регрессионное дерево

Общий случай дерева

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Как строится дерево?

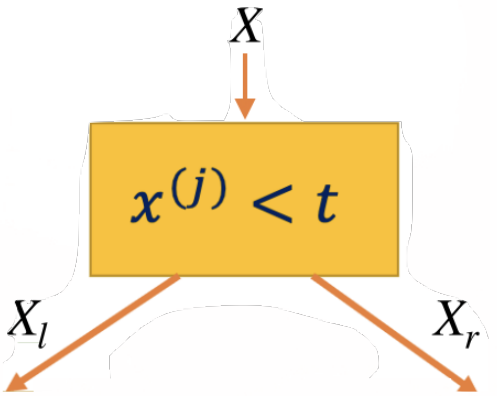
Пусть X - обучающая выборка.

Найдем правило $I\{x^{(j)} < t\}$

оптимальное по некоторому критерию.

Данное правило разобъет X

на 2 части: X_l и X_r.



Как строится дерево?

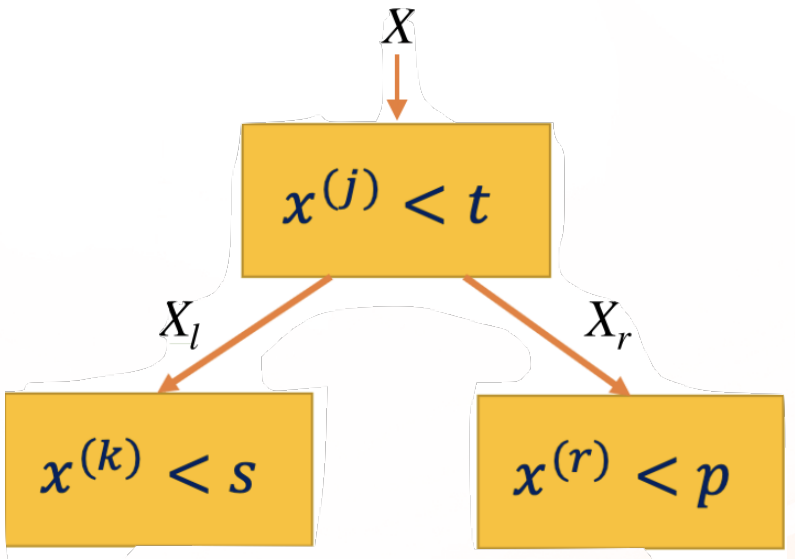
Пусть X - обучающая выборка.

Найдем правило $I\{x^{(j)} < t\}$

оптимальное по некоторому критерию.

Данное правило разобъет X

на 2 части: X_l и X_r.



Процедура вызывется рекурсивно от двух дочерних вершин

с обучающими выборками X_l и X_r соответственно.

Как строится дерево?

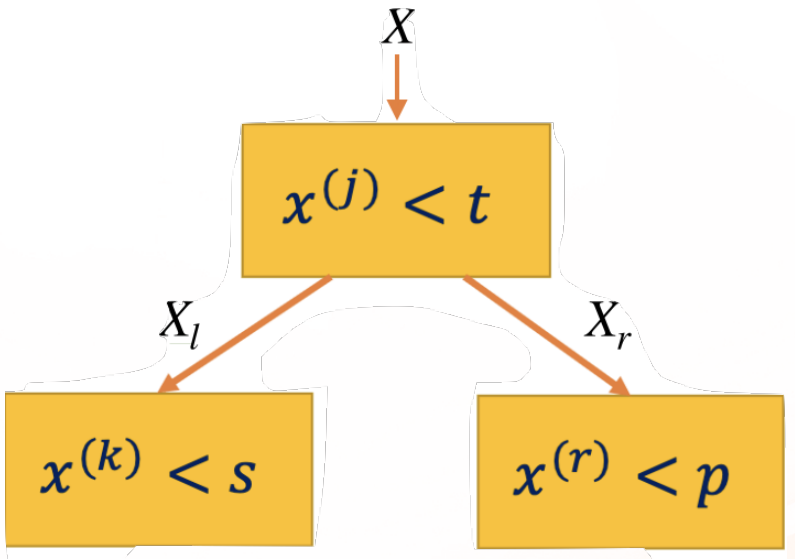
Пусть X - обучающая выборка.

Найдем правило $I\{x^{(j)} < t\}$

оптимальное по некоторому критерию.

Данное правило разобъет X

на 2 части: X_l и X_r.



Процедура вызывется рекурсивно от двух дочерних вершин

с обучающими выборками X_l и X_r соответственно.

Если выполнен некий критерий останова —

не делить текущую вершину.

Выбор разбиения

Как выбрать оптимальное разбиение $I\{x^{(j)} < t\}$*?*

Выбор разбиения

Как выбрать оптимальное разбиение $I\{x^{(j)} < t\}$*?*

Пусть в вершине m оказалась выборка X_m.

Выбор разбиения

Как выбрать оптимальное разбиение $I\{x^{(j)} < t\}$*?*

Пусть в вершине m оказалась выборка X_m.

Пусть $Q(X_m, j, t)$ - критерий ошибки условия $I\{x^{(j)} < t\}$.

Выбор разбиения

Как выбрать оптимальное разбиение $I\{x^{(j)} < t\}$*?*

Пусть в вершине m оказалась выборка X_m.

Пусть $Q(X_m, j, t)$ - критерий ошибки условия $I\{x^{(j)} < t\}$.

Ищем лучшие параметры (номер признака j и порог t) перебором$^{(*)}$:

$$Q(X_m, j, t) \to \min_{j,t}$$

(*) — Перебираем все возможные признаки, для каждого все возможные его пороги.

Выбор разбиения

Выбор разбиения

Вид критерия:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

Выбор разбиения

Вид критерия:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X)$ - **критерий информативности (impurity).**

Выбор разбиения

Вид критерия:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X)$ - **критерий информативности (impurity).**

Показывает разброс ответов в вершине, т.е. качество подвыборки X.

Чем меньше разброс ответов в вершине, тем меньше значение $H(X)$.

Выбор разбиения

Вид критерия:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X)$ - **критерий информативности (impurity).**

Показывает разброс ответов в вершине, т.е. качество подвыборки X.

Чем меньше разброс ответов в вершине, тем меньше значение $H(X)$.

Хорошее разбиение: после него больше уверены в ответе в вершине.

Т.е. хотим разбить вершину на две так,
чтобы полученные две вершины были более однородны по ответам.

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов, которые идут вправо и влево.

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов, которые идут вправо и влево.

Зачем это нужно?

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов, которые идут вправо и влево.

Зачем это нужно?

Пусть $|X_m| = 1000, \quad |X_l| = 990\ , \quad |X_r| = 10$

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов, которые идут вправо и влево.

Зачем это нужно?

Пусть $|X_m| = 1000, \quad |X_l| = 990\ , \quad |X_r| = 10$

В X_l все объекты имеют один класс $\Rightarrow$ $H(X_l)$ маленький.

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов,
которые идут вправо и влево.

Зачем это нужно?

Пусть $|X_m| = 1000, \quad |X_l| = 990\ , \quad |X_r| = 10$

В X_l все объекты имеют один класс $\Rightarrow$ $H(X_l)$ маленький.

X_r содержит объекты всех возможных классов $\Rightarrow$ $H(X_r)$ большой.

Выбор разбиения

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

$H(X_l)$ и $H(X_r)$ нормируются на доли объектов,
которые идут вправо и влево.

Зачем это нужно?

Пусть $|X_m| = 1000, \quad |X_l| = 990\ , \quad |X_r| = 10$

В X_l все объекты имеют один класс $\Rightarrow$ $H(X_l)$ маленький.

X_r содержит объекты всех возможных классов $\Rightarrow$ $H(X_r)$ большой.

Не так страшно, что X_r получилось плохим,
при том, что 990 попали в правильную вершину.

Критерий информативности

Неформальное определение:

- $H(X)$ зависит от меток в выборке X
- Показывает разброс ответов в X
- Чем меньше разброс ответов в X, тем меньше $H(X)$

Критерий информативности

Неформальное определение:

- $H(X)$ зависит от меток в выборке X
- Показывает разброс ответов в X
- Чем меньше разброс ответов в X, тем меньше $H(X)$

Формальное определение:

$H(X)$ показывает насколько хорошо целевые переменные предсказываются константой при оптимальном выборе константы:

$$H(X) = \min_{c \in Y} \frac{1}{|X|} \sum_{x_i \in X} \mathcal{L}(Y_i, c),$$

где $\mathcal{L}(y, c)$ - некоторая функция потерь.

Критерий информативности

Неформальное определение:

- $H(X)$ зависит от меток в выборке X
- Показывает разброс ответов в X
- Чем меньше разброс ответов в X, тем меньше $H(X)$

Формальное определение:

$H(X)$ показывает насколько хорошо целевые переменные предсказываются константой при оптимальном выборе константы:

$$H(X) = \min_{c \in Y} \frac{1}{|X|} \sum_{x_i \in X} \mathcal{L}(Y_i, c),$$

где $\mathcal{L}(y, c)$ - некоторая функция потерь.

Чтобы получить вид критерия при конкретной $\mathcal{L}(y, c)$,
нужно найти оптимальное значение c и подставить его в формулу.

Критерий информативности: задача регрессии с MSE

От неформального определения:

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Какая оптимальная константа для предсказания?

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Какая оптимальная константа для предсказания?

$$c = \frac{1}{|X|} \sum_{x_i \in X} y_i = \overline{y}$$

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Какая оптимальная константа для предсказания?

$$c = \frac{1}{|X|} \sum_{x_i \in X} y_i = \overline{y}$$

Какой вид имеет критерий информативности?

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Какая оптимальная константа для предсказания?

$$c = \frac{1}{|X|} \sum_{x_i \in X} y_i = \overline{y}$$

Какой вид имеет критерий информативности?

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - c)^2 =$$

Критерий информативности: задача регрессии с MSE

От неформального определения:

Как показать разброс ответов для задачи регрессии?

Возьмем выборочную дисперсию ответов в качестве $H(X)$:

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

От формального определения:

Какая оптимальная константа для предсказания?

$$c = \frac{1}{|X|} \sum_{x_i \in X} y_i = \overline{y}$$

Какой вид имеет критерий информативности?

$$H(X) = \frac{1}{|X|} \sum_{x_i \in X} (y_i - c)^2 = \frac{1}{|X|} \sum_{x_i \in X} (y_i - \overline{y})^2$$

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Энтропийный критерий

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

Считаем, что $0 \ln 0 = 0$

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Энтропийный критерий

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

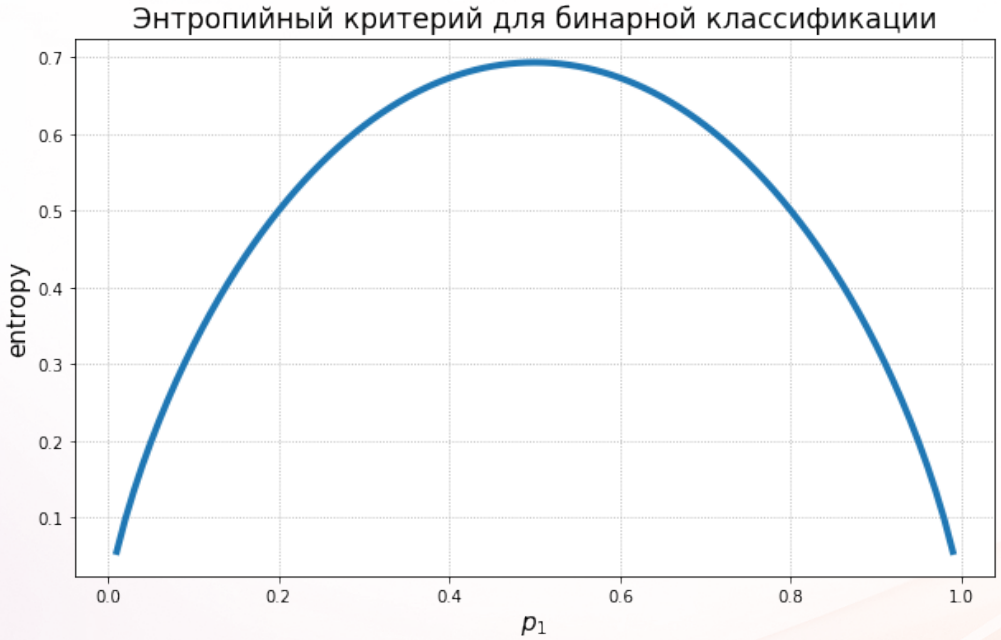
Считаем, что $0 \ln 0 = 0$

- $H(X) \geq 0$
- При $p_1 = 1, p_2 = 0, ..., p_K = 0 : H(X) = 0$

Критерий информативности: задача классификации

Энтропийный критерий

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



Интерпретация:

Мера отличия распределения классов от вырожденного.

Вырожденное - всегда знаем, что получим,

равномерное - непредсказуемо

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Критерий Джини

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

Критерий информативности: задача классификации

Пусть решаем задачу классификации на K классов.

p_1, ..., p_k - доли объектов классов $1, ..., K$ в X:

$$p_j = \frac{\sum\limits_{x_i \in X} I\{y_i = j\}}{|X|}$$

Критерий Джини

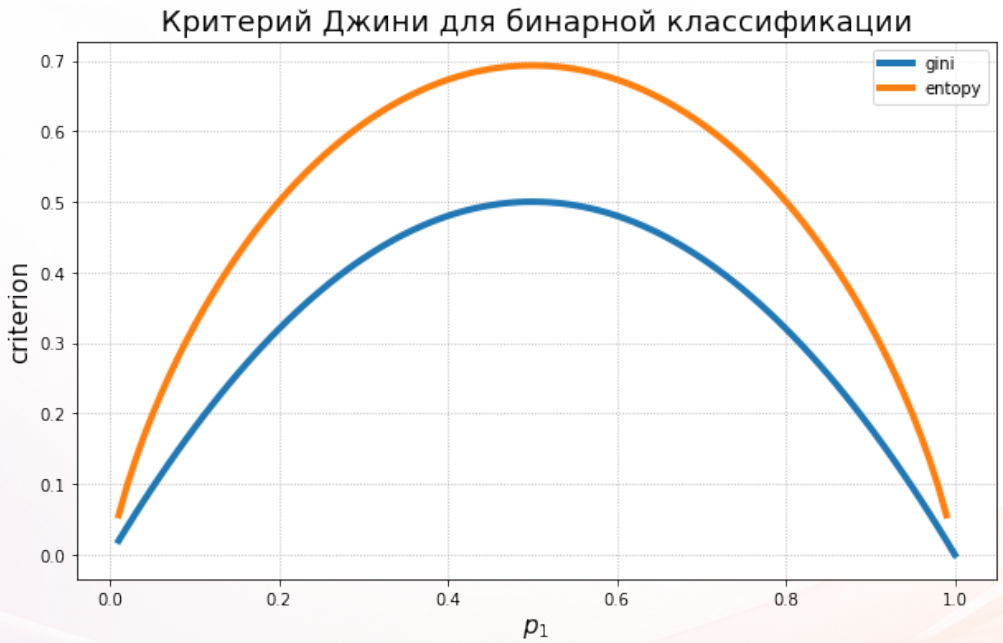
$$H(X) = \sum_{k=1}^{K} p_k(1 - p_k)$$

- $H(X) \geq 0$
- При $p_1 = 1, p_2 = 0, ..., p_K = 0 : H(X) = 0$

Критерий информативности: задача классификации

Критерий Джини

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



Интерпретация:

Вероятность ошибки случайного классификатора, который выдает ответы пропорционально p_k.

Критерий информативности

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Все объекты в вершине относятся к одному классу

  Простой критерий, хорош для простых выборок.
  Если выборка сложная, то сработает только когда
  в вершине осталось 1-2 объекта.

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Все объекты в вершине относятся к одному классу

  Простой критерий, хорош для простых выборок.
  Если выборка сложная, то сработает только когда
  в вершине осталось 1-2 объекта.

- В вершину попало $\leq k$ объектов.

  k - гиперпараметр. Нужно выбирать таким, что по k объектам в листе можно построить надежный прогноз.

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Все объекты в вершине относятся к одному классу

  Простой критерий, хорош для простых выборок.
  Если выборка сложная, то сработает только когда
  в вершине осталось 1-2 объекта.

- В вершину попало $\leq k$ объектов.

  k - гиперпараметр. Нужно выбирать таким, что по k объектам в листе можно построить надежный прогноз.

- Глубина дерева превысила порог.

  Грубый критерий, не зависит ни от распределения классов,
  ни от числа объектов.

  Хорошо работает в композициях
  (много моделей объединяются в одну сложную модель).

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Число листьев в дереве превысило порог.

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Число листьев в дереве превысило порог.
- Функционал ошибки при делении вершины не уменьшился.

  Если лучшее из разбиений приводит к росту функционала ошибки, не разбиваем эту вершину.

Критерий останова

Как понять, разбивать ли вершину далее или сделать ее листовой?

- Число листьев в дереве превысило порог.
- Функционал ошибки при делении вершины не уменьшился.

  Если лучшее из разбиений приводит к росту функционала ошибки, не разбиваем эту вершину.
- Функционал ошибки при делении вершины на две не уменьшился на s процентов.

  Если лучшее из разбиений не уменьшает функционал на $s\%$, не разбиваем эту вершину.

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max\limits_{y \in \mathbf{Y}} \sum\limits_{i \in X_m} I\{Y_i = y\}$

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max_{y \in Y} \sum_{i \in X_m} I\{Y_i = y\}$
  2. Оценки вероятности классов $\widehat{p_m} = [\widehat{p_m^1}, ..., \widehat{p_m^K}]$,где $\widehat{p_m^k} = \frac{1}{|X_m|} \sum_{i \in X_m} I\{Y_i = k\}$

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max_{y \in Y} \sum_{i \in X_m} I\{Y_i = y\}$
  2. Оценки вероятности классов $\widehat{p_m} = [\widehat{p_m^1}, ..., \widehat{p_m^K}]$,где $\widehat{p_m^k} = \frac{1}{|X_m|} \sum_{i \in X_m} I\{Y_i = k\}$
- Регрессия с MSE:

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max_{y \in Y} \sum_{i \in X_m} I\{Y_i = y\}$
  2. Оценки вероятности классов $\widehat{p_m} = [\widehat{p_m^1}, ..., \widehat{p_m^K}]$,где $\widehat{p_m^k} = \frac{1}{|X_m|} \sum_{i \in X_m} I\{Y_i = k\}$
- Регрессия с MSE:

$$\widehat{y_m} = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$$

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max_{y \in Y} \sum_{i \in X_m} I\{Y_i = y\}$
  2. Оценки вероятности классов $\widehat{p_m} = [\widehat{p_m^1}, ..., \widehat{p_m^K}]$,где $\widehat{p_m^k} = \frac{1}{|X_m|} \sum_{i \in X_m} I\{Y_i = k\}$
- Регрессия с MSE:

$$\widehat{y_m} = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$$

Почему это оптимально?

Ответ в листе

Дерево построено.

Какое предсказание выдавать для объекта, попавшего в лист?

- Классификация:
  1. Самый популярный класс в листе: $\widehat{y_m} = \arg\max_{y \in Y} \sum_{i \in X_m} I\{Y_i = y\}$
  2. Оценки вероятности классов $\widehat{p_m} = [\widehat{p_m^1}, ..., \widehat{p_m^K}]$,где $\widehat{p_m^k} = \frac{1}{|X_m|} \sum_{i \in X_m} I\{Y_i = k\}$
- Регрессия с MSE:

$$\widehat{y_m} = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$$

Почему это оптимально?

Узнаем в домашке!

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Пропуски в данных

Типы пропусков:

1. случайные (напр., сломалось оборудование)
2. неслучайные (напр., признак не применим к объекту)

В решающих деревьях обрабатываются одинаково.

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.
  Игнорируем пропущенные значения.

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.
  Игнорируем пропущенные значения. Приближение функцинала:

$$Q(X_m, j, t) \approx \frac{|X_m \setminus X_m^o(j)|}{|X_m|} Q(X_m \setminus X_m^o(j), j, t)$$

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.
  Игнорируем пропущенные значения. Приближение функцинала:

$$Q(X_m, j, t) \approx \frac{|X_m \setminus X_m^o(j)|}{|X_m|} Q(X_m \setminus X_m^o(j), j, t)$$

- Если правило $I\{x^{(j)} < t\}$ оказалось оптимальным

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.
  Игнорируем пропущенные значения. Приближение функцинала:

$$Q(X_m, j, t) \approx \frac{|X_m \setminus X_m^o(j)|}{|X_m|} Q(X_m \setminus X_m^o(j), j, t)$$

- Если правило $I\{x^{(j)} < t\}$ оказалось оптимальным
  1. Отправляем $X_m^o(j)$ в оба поддерева.

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.

  Игнорируем пропущенные значения. Приближение функцинала:

$$Q(X_m, j, t) \approx \frac{|X_m \setminus X_m^o(j)|}{|X_m|} Q(X_m \setminus X_m^o(j), j, t)$$

- Если правило $I\{x^{(j)} < t\}$ оказалось оптимальным

  1. Отправляем $X_m^o(j)$ в оба поддерева.
  2. Оцениваем вероятности попадания объекта в каждое поддерево:

$$\widehat{p}_{m,l} = \frac{|X_l \setminus X_m^o(j)|}{|X_m \setminus X_m^o(j)|}, \quad \widehat{p}_{m,r} = \frac{|X_r \setminus X_m^o(j)|}{|X_m \setminus X_m^o(j)|}$$

Этап обучения

Деление узла m

Пусть в узле m оказалась выборка X_m.

$X_m^o(j)$ — объекты из X_m, для которых неизвестен $x^{(j)}$.

- Рассматриваем правило $I\{x^{(j)} < t\}$.

  Игнорируем пропущенные значения. Приближение функцинала:

$$Q(X_m, j, t) \approx \frac{|X_m \setminus X_m^o(j)|}{|X_m|} Q(X_m \setminus X_m^o(j), j, t)$$

- Если правило $I\{x^{(j)} < t\}$ оказалось оптимальным

  1. Отправляем $X_m^o(j)$ в оба поддерева.
  2. Оцениваем вероятности попадания объекта в каждое поддерево:

$$\widehat{p}_{m,l} = \frac{|X_l \setminus X_m^o(j)|}{|X_m \setminus X_m^o(j)|}, \quad \widehat{p}_{m,r} = \frac{|X_r \setminus X_m^o(j)|}{|X_m \setminus X_m^o(j)|}$$

Листья

Считаем оценки вероятностей классов.

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.

l и r — узлы левого и правого поддеревьев.

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.

l и r — узлы левого и правого поддеревьев.

Обозначим "$x_0 \Rightarrow m$" событие "x_0 попал в узел m"

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathsf{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) =$$

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$
$$+ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$
$$+ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$
$$= \widehat{p}_{m,l} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) + \widehat{p}_{m,r} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r)$$

Этап применения дерева для задачи классификации

Пусть x_0 — новый объект. Нужно оценить его отклик Y_0.

Рассмотрим узел m с правилом $I\{x^{(j)} < t\}$.
l и r — узлы левого и правого поддеревьев.
Обозначим "$x_0 \Rrightarrow m$" событие "x_0 попал в узел m"

- Значение $x_0^{(j)}$ известно $\Longrightarrow$ отправляем его в l или r

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \begin{cases} \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \\ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \end{cases}$$

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$
$$+ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$
$$= \widehat{p}_{m,l} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) + \widehat{p}_{m,r} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r)$$

Итоговая оценка $\widehat{\mathrm{P}}_{x_0}(Y_0 = y) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow root)$

Этап применения дерева для задачи классификации

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$

$$+ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$

$$= \widehat{p}_{m,l} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) + \widehat{p}_{m,r} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r)$$

Итоговая оценка $\widehat{\mathrm{P}}_{x_0}(Y_0 = y) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow root)$

Этап применения дерева для задачи классификации

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_\mathbf{0}}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$

$$+ \widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_\mathbf{0}}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$

$$= \widehat{p}_{m,l} \cdot \widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow l) + \widehat{p}_{m,r} \cdot \widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow r)$$

Итоговая оценка $\widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y) = \widehat{\mathrm{P}}_{x_\mathbf{0}}(Y_0 = y \mid x_0 \Rrightarrow root)$

Смысл операции: считаем оценки вероятностей классов в поддеревьях и усредняем их с весами, равными оценке вероятности попасть в конкретное поддерево.

Этап применения дерева для задачи классификации

- Значение $x_0^{(j)}$ неизвестно $\Longrightarrow$ отправляем его в оба поддерева

$$\widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow m) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow l \mid x_0 \Rrightarrow m) +$$

$$+ \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r) \cdot \widehat{\mathrm{P}}_{x_0}(x_0 \Rrightarrow r \mid x_0 \Rrightarrow m) =$$

$$= \widehat{p}_{m,l} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow l) + \widehat{p}_{m,r} \cdot \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow r)$$

Итоговая оценка $\widehat{\mathrm{P}}_{x_0}(Y_0 = y) = \widehat{\mathrm{P}}_{x_0}(Y_0 = y \mid x_0 \Rrightarrow root)$

Смысл операции: считаем оценки вероятностей классов в поддеревьях и усредняем их с весами, равными оценке вероятности попасть в конкретное поддерево.

Для регрессии нужно заменить условную вероятность на УМО. Т.е. считаем оценку отклика в поддеревьях и усредняем его с весами.

Обработка пропусков

Решающие деревья

Общий случай дерева

Регрессионное дерево

Построение дерева

Критерий останова

Ответ в листе

Пропуски в данных

Плюсы и минусы деревьев

Плюсы и минусы деревьев

Плюсы

1. Интерпретируемая структура
2. Восстанавливают сложные нелинейные зависимости
3. Умеет обрабатывать категориальные признаки
4. Умеет обрабатывать пропущенные значения
5. Не требует нормализации и масштабирования признаков

Плюсы и минусы деревьев

Плюсы

1. Интерпретируемая структура
2. Восстанавливают сложные нелинейные зависимости
3. Умеет обрабатывать категориальные признаки
4. Умеет обрабатывать пропущенные значения
5. Не требует нормализации и масштабирования признаков

Минусы

1. Легко переобучаются: маленькое изменение с выборке может сильно изменить дерево
2. Требуют больше вычислений, чем линейные модели
3. Решающее правило всегда параллельно осям признаков
4. Плохо обрабатывают линейные зависимости
5. Жадный выбор разбиения в вершине
6. Чем дальше вершина от корня, тем меньше обучающая выборка


ВСЁ!