H(R) в задаче регрессии и жёсткой классификации

Елена Кантонистова

Skillbox

H(R)

- На каждом шаге при построении решающего дерева выбирайте такой признак x_j и такой порог t, что при разбиении объектов на две группы условием $x_j > t$ значение Information Gain

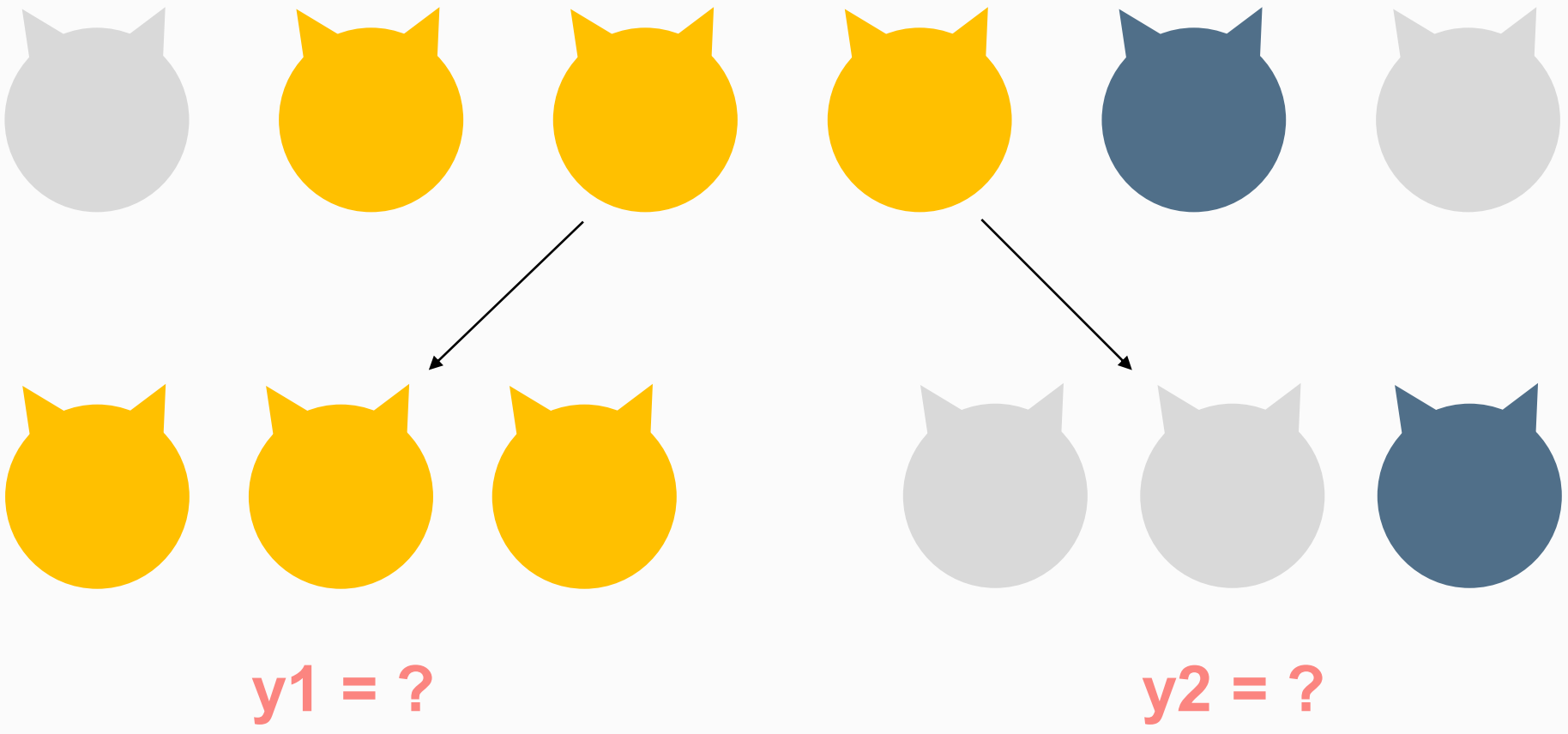
$$Q(R, j, t) = H(R) - \frac{|R_l|}{|R|} H(R_l) - \frac{|R_l|}{|R|} H(R_r)$$

будет максимально.

Как измерять неоднородность объектов $H(R)$?

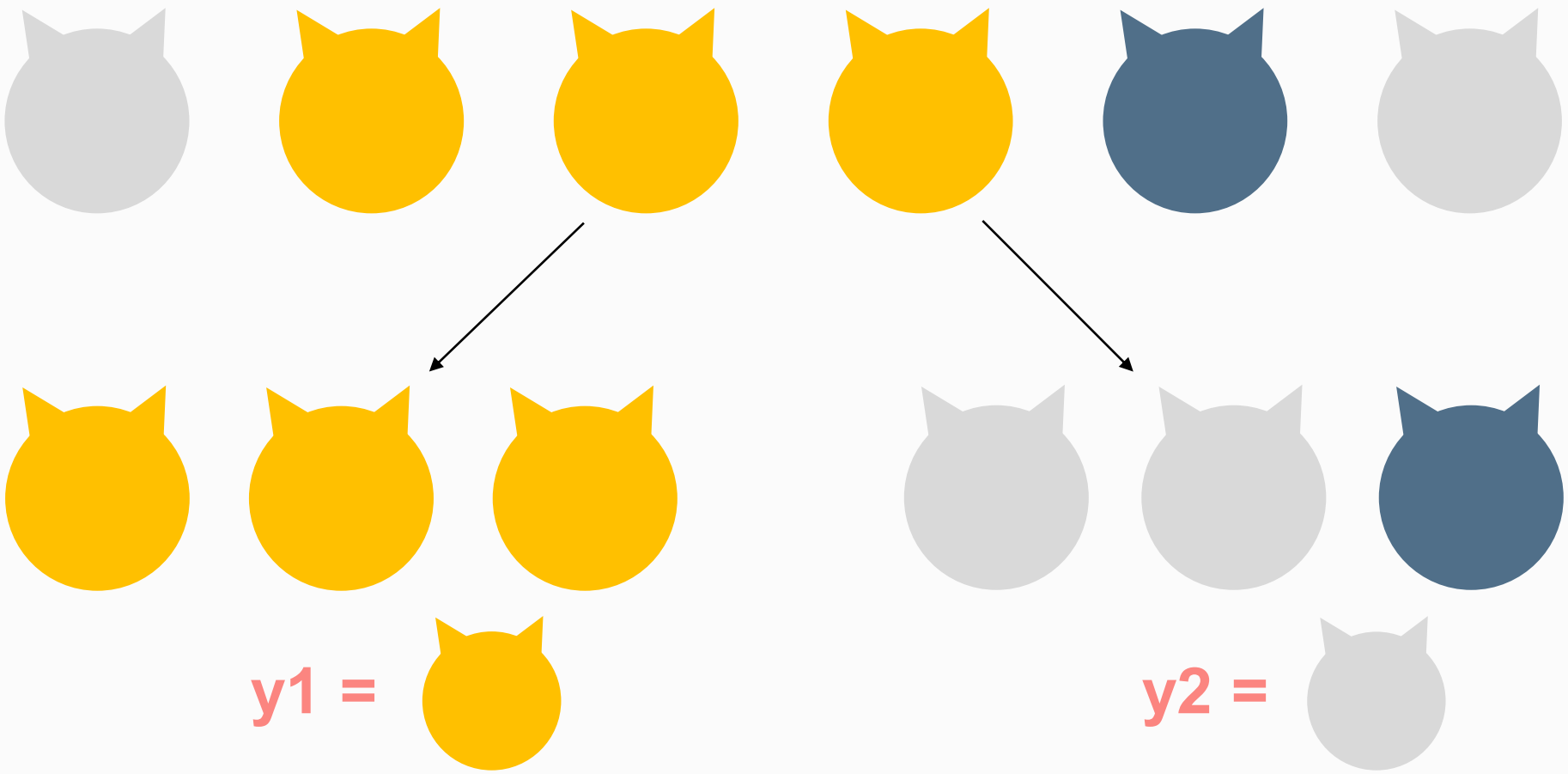
Жёсткая классификация

Пример: нужно определить цвет котиков.



Жёсткая классификация

В задаче классификации в вершине предсказывается самый популярный класс



H(R) в жёсткой классификации

- В качестве меры неоднородности $H(R)$ в задаче жёсткой классификации берут ошибку классификации, то есть долю неверно предсказанных классов

Регрессия

- Пусть решается задача предсказания стоимости квартир
- Для предсказания целевой переменной в вершине минимизируйте MSE

Регрессия

- Пусть решается задача предсказания стоимости квартир
- Для предсказания целевой переменной в вершине минимизируйте MSE

Кейс 1: в вершину попали квартиры стоимостью
10 млн, 100 млн, 30 млн

Кейс 2: вершину попали квартиры стоимостью
10 млн, 10.5 млн, 9.8 млн, 10.2 млн

Регрессия

- Пусть решается задача предсказания стоимости квартир
- Для предсказания целевой переменной в вершине минимизируйте MSE

Кейс 1: в вершину попали квартиры стоимостью
10 млн, 100 млн, 30 млн.

Вывод: совершенно разные квартиры, их надо развести по разным веткам.

Кейс 2: вершину попали квартиры стоимостью
10 млн, 10.5 млн, 9.8 млн, 10.2 млн.

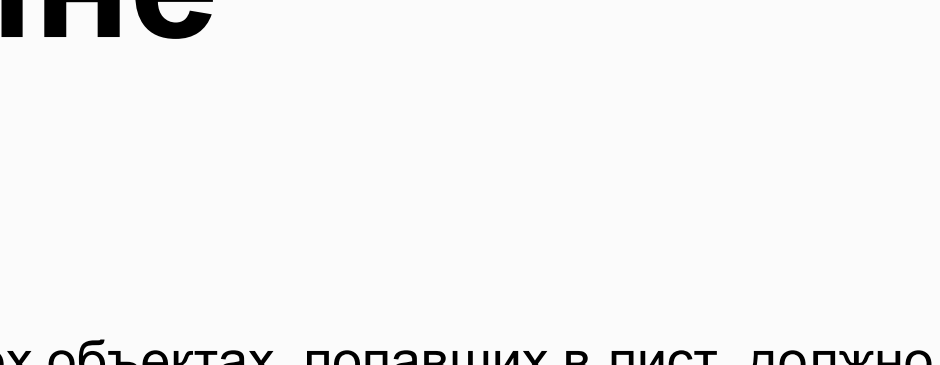
Вывод: похожие по ответу квартиры, их можно оставить в одной вершине и назвать её листом.

Предсказание в вершине

- Предсказание на всех объектах, попавших в лист, должно быть одно и то же
- При минимизации MSE предсказывается среднее значение целевой переменной по всем объектах в вершине

Кейс 1:

Кейс 2:



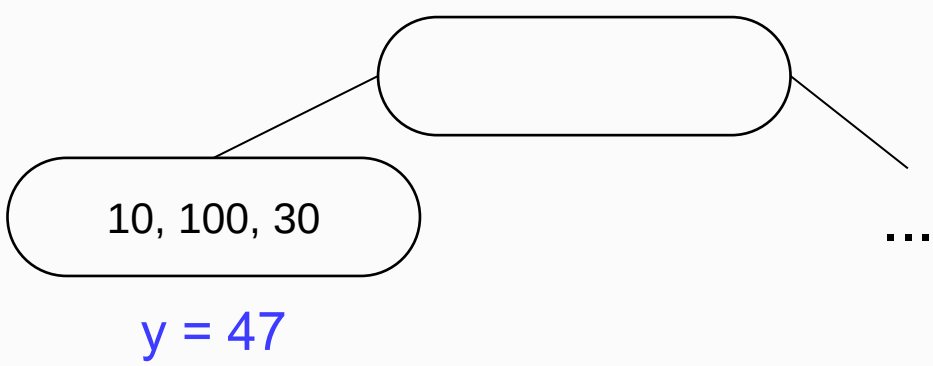
Автор: Кантонистова Елена

Дисперсия

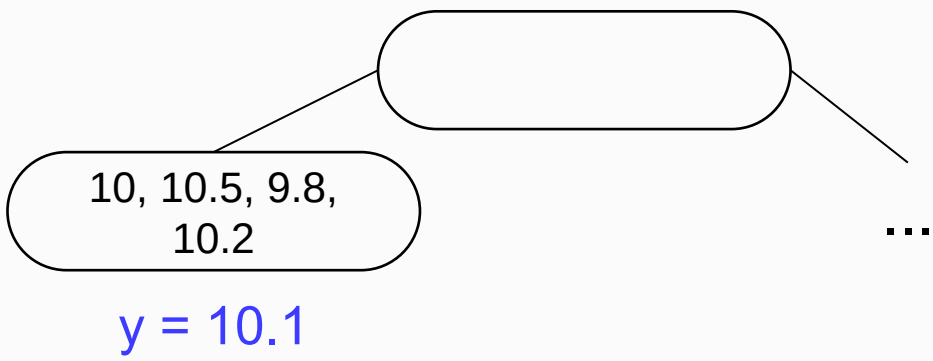
$$D(R) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2,$$

где $\hat{y}$ — среднее значение целевой переменной в вершине R

Кейс 1: $D(R) = 1489$



Кейс 2: $D(R) = 0.075$



Автор: Кантонистова Елена

H(R) для регрессии

В качестве критерия информативности в задаче регрессии берут **дисперсию**

$$H(R) = D(R)$$

Итоги

- $H(R)$ в задаче регрессии — дисперсии целевой переменной на объектах в вершине
- $H(R)$ в задаче жёсткой классификации — ошибка классификации, т. е. доля неверно предсказанных классов в вершине