10-11 классы

Большие данные

Презентация занятия

Решающие деревья. Энтропия

16 занятие

инжинириум®
МГТУ им. Н.Э. Баумана

2023

Тема: Решающие деревья. Энтропия

На прошлом уроке мы знакомились с решающими деревьями. При создании решающего дерева мы использовали критерий энтропии. На этом уроке мы с вами поговорим о том, что такое энтропия и какая математика лежит в основе обучения решающих деревьев.

```
clf = tree.DecisionTreeClassifier(criterion='entropy')
```

Тема: Решающие деревья. Энтропия

Для построения решающего дерева нам необходимо задавать такие вопросы, которые будут уменьшать количество неопределенности наших данных.
Что подразумевается под неопределенностью, когда мы решаем задачу классификации?

Тема: Решающие деревья. Энтропия

Давайте рассмотрим некоторое множество объектов, которое мы хотим классифицировать. Допустим, что X - погибший пассажир, O – выживший

У нас есть выборка всех пассажиров Титаника - XXXXOOOO

Какова вероятность того, что следующий пассажир будет X?

Тема: Решающие деревья. Энтропия

А теперь мы узнали пол каждого пассажира.

Распределим наши данные по полу (1 = female, 0 = male)
Sex = 1 → OOOX (выжило 3 женщины, а погибла 1) → вероятность выживания женщины 3\4

Sex = 0 → XXXO (выжил 1 мужчина, погибло 3) → вероятность выживания мужчины 1\4

Таким образом, мы уже с большей вероятностью можем сказать выживет ли пассажир, зная его пол

Тема: Решающие деревья. Энтропия

А теперь мы узнали пол каждого пассажира.

Распределим наши данные по полу (1 = female, 0 = male)
Sex = 1 → OOOX (выжило 3 женщины, а погибла 1) → вероятность выживания женщины 3\4

Sex = 0 → XXXO (выжил 1 мужчина, погибло 3) → вероятность выживания мужчины 1\4

Таким образом, мы уже с большей вероятностью можем сказать выживет ли пассажир, зная его пол

Введем еще одно свойство - возраст.
Для Sex = 1 → OOOX
Age > 30 → X (1 погибший)
Age <30 → OOO (3 выживших)

Тема: Решающие деревья. Энтропия

Введем еще одно свойство - возраст.
Для Sex = 1 → OOOX
Age > 30 → X (1 погибший)
Age <30 → OOO (3 выживших)

Т.е. если у нас есть женщина и она старше 30 лет - то по нашей статистике она погибнет. А все женщины моложе 30 лет - выжили. Таким образом мы получили, что для этих состояний мы находимся в полной и максимальной определенности

В итоге мы начали с полной неопределенности, а добавив некоторых свойств и разделяя наши данные по этим свойствам мы пришли к полной определенности и смогли безошибочно классифицировать наши данные

Тема: Решающие деревья. Энтропия

Вернемся к нашему искусственному датасету
У нас будет две переменные X_1 и X_2, которые принимают количественные значения 0 или 1. И переменная Y, которая отвечает за некоторый итоговый класс.

Как нам измерить математически, что при включении информации о переменной X_1 наша неопределенность снижается?
Давайте введем важный термин, связанные с обучением решающих деревьев - термин **Энтропия**

```
In [70]: data
Out[70]:
```

	X_1	X_2	Y
0	1	0	1
1	1	0	1
2	1	0	1
3	0	1	1
4	0	0	0
5	0	0	0
6	0	0	0
7	1	1	0

Тема: Решающие деревья. Энтропия

Энтропия - уровень неопределенности наших данных

Неопределенность тем выше, чем хуже у нас получается разделять классы и относить наблюдения только к одному или другому классу.

$$H(A) = -\sum_{i=1}^{m} P_i \log_2 P_i,$$

Тема: Решающие деревья. Энтропия

Рассмотрим наше состояние, когда у нас 50 % крестиков и 50% кружочков. Вероятность того, что мы отнесем к крестикам - 1\2, вероятность того, что мы отнесем к кружочкам - 1\2

Тогда энтропия рассчитывается:

$$E = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} =$$

$$= \frac{1}{2} + \frac{1}{2} = 1$$

Тема: Решающие деревья. Энтропия

Рассмотрим наше состояние, когда у нас 100 % крестиков
Вероятность того, что мы отнесем к крестикам – 1, вероятность того, что мы отнесем к ноликам – 0

Тогда энтропия рассчитывается:

$$E = -1 \cdot \log_2 1 - 0 \cdot \log_2 0 =$$

$$= -1 \cdot 0 - 0 = 0$$

Тема: Решающие деревья. Энтропия

Вернемся к нашему датасету. Условно заменим 1 на X, а 0 на O

	X_1	X_2	Y	
0	1	0	1	X
1	1	0	1	X
2	1	0	1	X
3	0	1	1	X
4	0	0	0	O
5	0	0	0	O
6	0	0	0	O
7	1	1	0	O

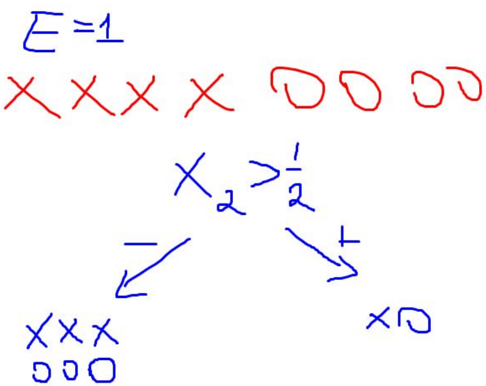
Тема: Решающие деревья. Энтропия

Давайте разделим наши данные по признаку X_2. Так как вариантов у нас всего 2 (X_2 = 0 и X_2 = 1), то условие X_2 > 1 / 2

У нас есть 2 записи, для которых выполняется условие и 6 записей, для которых условие не выполняется.

Рассчитайте энтропию для двух новых полученных групп

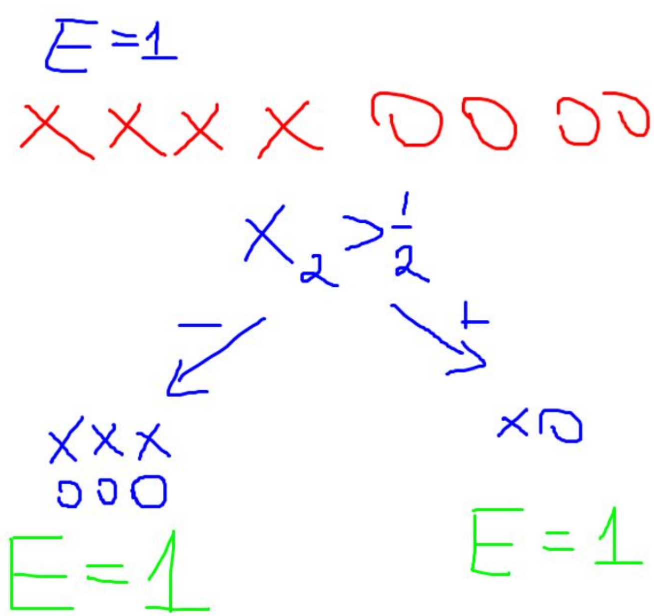
	X_1	X_2	Y	
0	1	0	1	X
1	1	0	1	X
2	1	0	1	X
3	0	1	1	X
4	0	0	0	O
5	0	0	0	O
6	0	0	0	O
7	1	1	0	O



Тема: Решающие деревья. Энтропия

В результате расчёта мы получим, что энтропия двух новых групп также равна 1. Что получили? Получили, что первое разделение по переменной X_2 не привнес нам никакой дополнительной неопределенности

	X_1	X_2	Y	
0	1	0	1	X
1	1	0	1	X
2	1	0	1	X
3	0	1	1	X
4	0	0	0	O
5	0	0	0	O
6	0	0	0	O
7	1	1	0	O

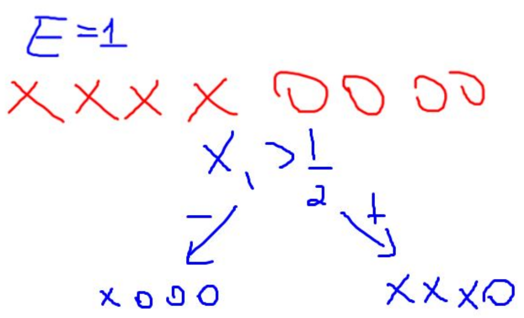


Тема: Решающие деревья. Энтропия

Давайте попробуем разделить наши данные по первому признаку X_1.

Рассчитайте энтропию новых получившихся групп

	X_1	X_2	Y	
0	1	0	1	X
1	1	0	1	X
2	1	0	1	X
3	0	1	1	X
4	0	0	0	O
5	0	0	0	O
6	0	0	0	O
7	1	1	0	O

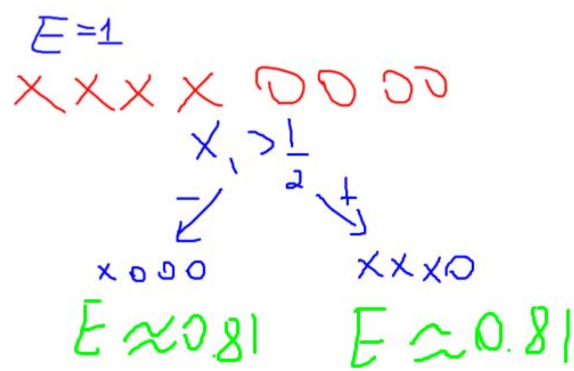


Тема: Решающие деревья. Энтропия

Давайте попробуем разделить наши данные по первому признаку X_1.

Рассчитайте энтропию новых получившихся групп

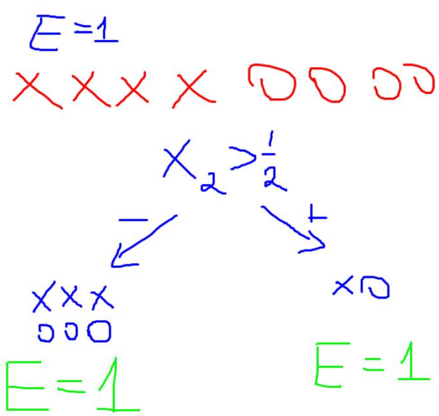
	X_1	X_2	Y	
0	1	0	1	X
1	1	0	1	X
2	1	0	1	X
3	0	1	1	X
4	0	0	0	O
5	0	0	0	O
6	0	0	0	O
7	1	1	0	O

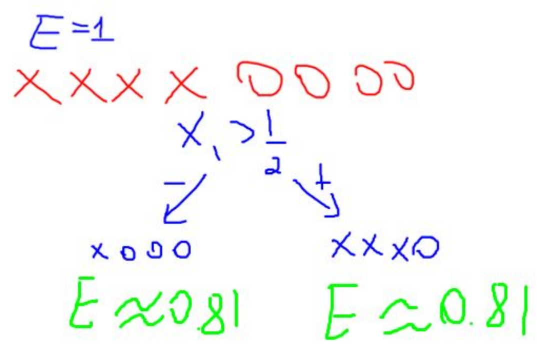


Тема: Решающие деревья. Энтропия

Что получили в итоге? Если делаем разделение по X_2 – энтропия не меняется. Если делаем по X_1 – меняется.

Далее необходимо понять, какая из переменных внесла больший вклад в снижение неопределенности наших данных



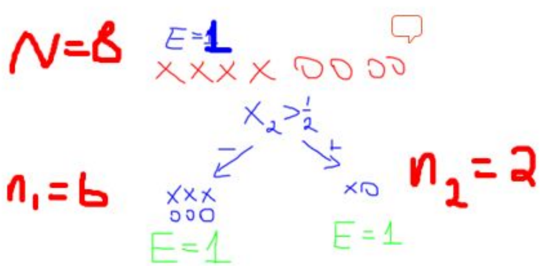


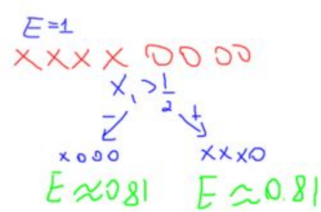
Тема: Решающие деревья. Энтропия

Как понять, какой «сплит» принес нам больше «пользы». Введем понятие IG (information gain)

IG = E(Y) – E(Y / x)
IG = полная энтропия – энтропия с учетом разделения по переменной

E(Y / x) = (n_1 / N) * E1 + (n_2 / N) * E2, где
n_1 и n_2 – количество элементов, вошедших в первую и вторую выборку соответственно, E1 и E2 – энтропия первой и второй выборки и N – общее количество элементов





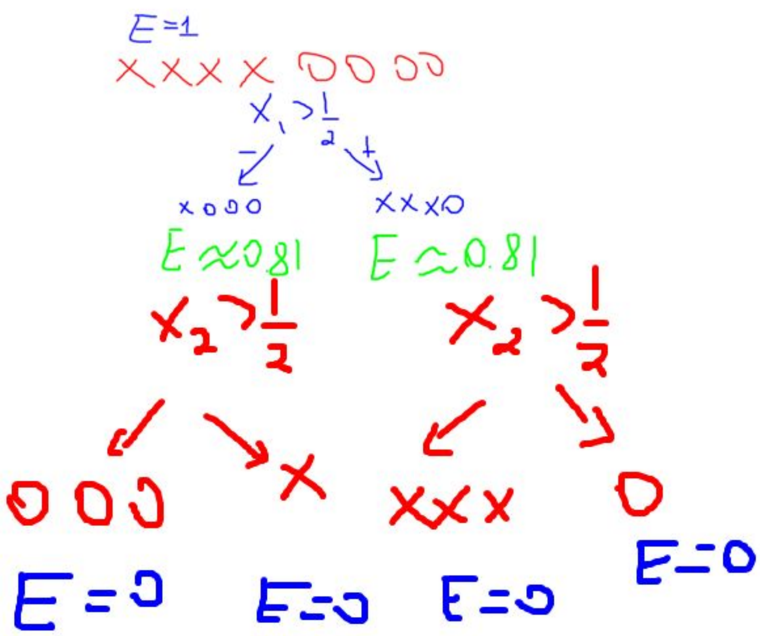
$$IG = 1 - \left(\frac{6}{8}\cdot 1 + \frac{2}{8}\cdot 1\right) = 0$$

$$IG = 1 - \left(\frac{4}{8}\cdot 0{,}81 + \frac{4}{8}\cdot 0.81\right) = 0.19$$

Тема: Решающие деревья. Энтропия

Из расчёта IG мы понимаем, что сплит по переменной X_1 лучше, чем сплит по переменной X_2

Однако полной определенности мы не получили. Продолжаем разбивать «успешный» сплит по переменной X_2



Тема: Решающие деревья. Энтропия

Как правильно делить данные, если у нас не бинарное свойство (1 или 0), а количественное? Например возраст.

Тогда делаем сплит по всем возможным вариантам, считаем для каждого IG и делим данные по максимальному IG.

Если признаков несколько, например: пол (1 и 0), возраст (20, 30, 40, 50) – делаем сплит по всем вариантам и признакам, считаем максимальный IG – делим по нему. Для новых получившихся групп рекурсивно повторяем, пока не получим максимальную определенность.

Тема: Решающие деревья. Энтропия

1. Переменная Лазает по деревьям позволяет идеально различить 2 вида по исходным данным
2. Обе переменные Гавкает и Лазает по деревьям дают одинаковый Information Gain, если поместить их в вершину дерева
3. Переменная Шерстист позволяет идеально различить 2 вида по исходным данным
4. Для различения котиков от собачек, по этим данным, хватит всего 1-ой переменной
5. Переменная Гавкает позволяет идеально различить 2 вида по исходным данным
6. Все переменные одинаково хороши для разделения видов

	Шерстист	Гавкает	Лазает по деревьям	Вид
0	1	1	0	собачка
1	1	1	0	собачка
2	1	1	0	собачка
3	1	1	0	собачка
4	1	0	1	котик
5	1	0	1	котик
6	1	0	1	котик
7	1	0	1	котик

Тема: Решающие деревья. Энтропия

1. Рассчитайте энтропию при разделении по свойству Шерстист в группах (где Шерстист = 0 и 1)
2. Рассчитайте энтропию при разделении по свойству Гавкает в группах (где Гавкает = 0 и 1)
3. Рассчитайте энтропию при разделении по свойству Лазает по деревьям в группах (где Лазает = 0 и 1)
4. Рассчитайте IG для свойств Шерстист, Гавкает, Лазает по деревьям (не забудьте, что полная энтропия не обязательно равна 1)

Округляйте до 2х знаков после запятой

	Шерстист	Гавкает	Лазает по деревьям	Вид
0	1	1	0	собачка
1	1	1	0	собачка
2	1	1	0	собачка
3	1	1	0	собачка
4	1	0	1	котик
5	1	0	1	котик
6	1	0	1	котик
7	1	0	1	котик
8	1	1	1	котик
9	0	0	1	котик

Тема: Решающие деревья. Энтропия

Задача на программирование

1) Напишите функцию, которая на входе получает массив вероятностей и возвращает энтропию системы (проверьте, что суммарная вероятность = 1)
2) Напишите функцию, которая на входе получает массив различных чисел и возвращает энтропию системы