Решающие деревья

Юлия Пономарева
Data Scientist

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

Цели занятия

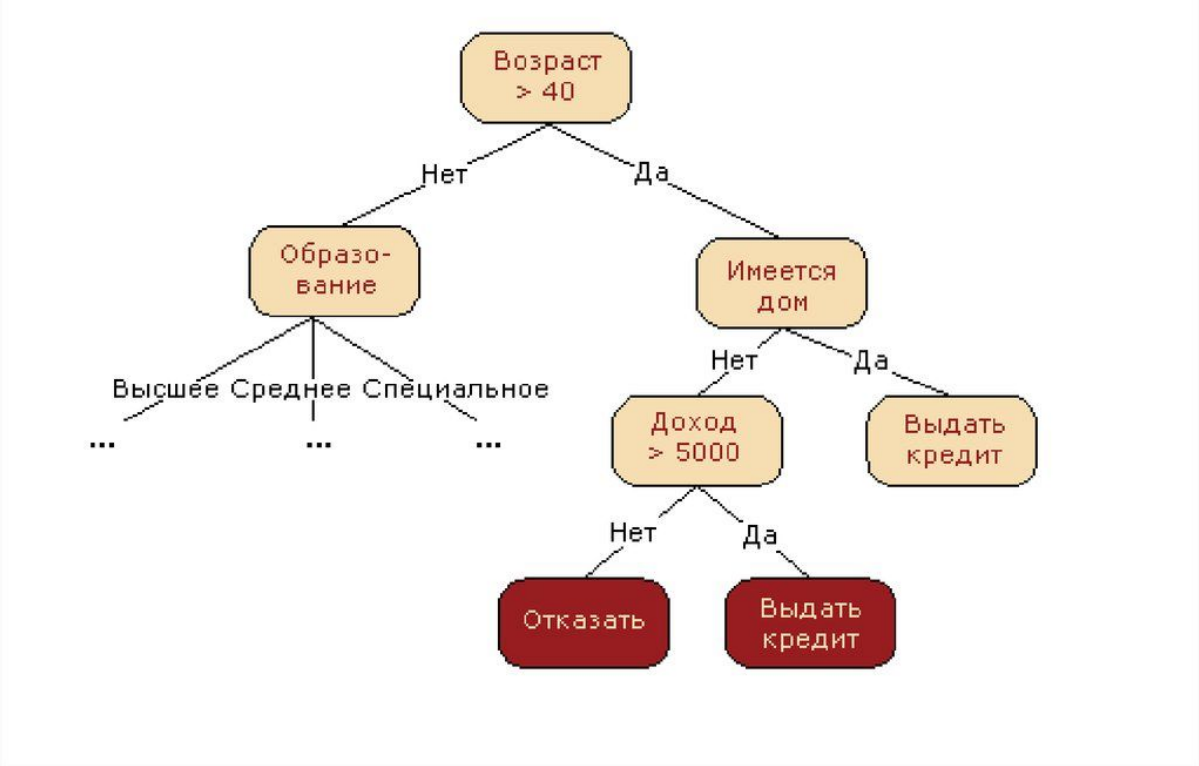
1. Изучим алгоритм построения дерева решений
2. Узнаем, какие есть информационные критерии
3. Познакомимся с критериями остнова
4. Подберем оптимальные гиперпараметры для дерева решений
5. Получим важность признаков для дерева решений

План занятия

1. Модель Дерево решений
2. Критерии информативности
3. Критерии останова
4. Важность признаков
5. Подбор гиперпараметров
6. Пример переобучения дерева решений
7. Итоги занятия

Дерево решений

Дерево решений



Дерево решений

Функция потерь - Информационный критерий - мера неопределённости в выборке.

- **Для классификации:** критерий Джини, энтропийный критерий
- **Для регрессии**: MSE, MAE

Предсказание

- **Для классификации:** Класс, который встречается чаще всего в листе/ вероятность классов
- **Для регрессии:** Среднее арифметическое целевых значений объектов, которые попали в итоговый лист

МФТИ

Построение дерева решений

Алгоритм построения дерева решений

1. Перебираем все признаки:
   - сортируем выбранный признак по возрастанию
   - перебираем пороги разделения выборки на две части, считая информационный критерий
2. Выбираем лучшее разбиение с точки зрения значения **прироста информации**

Алгоритм построения дерева решений

$$IG(R) = H(R) - q_{left} * H(R_{left}) - q_{right} * H(R_{right})$$

```
['cat', 'cat', 'cat', 'cat', 'cat', 'dog', 'dog', 'dog']
```

$$IG(R) = H(R) - 5/8 * 0 - 3/8 * 0$$

$$IG(R) = H(R)$$

Энтропийный критерий

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

$$H(R) = -(\frac{4}{9} \cdot log2(\frac{4}{9}) + \frac{5}{9} \cdot log2(\frac{5}{9})) = 0.991$$

$$H(Rleft) = -(\frac{3}{4} \cdot log2(\frac{3}{4}) + \frac{1}{4} \cdot log2(\frac{1}{4})) = 0.81$$

$$H(Rright) = -(\frac{1}{5} \cdot log2(\frac{1}{5}) + \frac{4}{5} \cdot log2(\frac{4}{5})) = 0.72$$

$$IG(R) = 0.991 - \frac{4}{9} \cdot 0.811 - \frac{5}{9} \cdot 0.722 = 0.22$$

Критерий Джини

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

$$H(R) = \frac{4}{9} \cdot (1 - \frac{4}{9}) + \frac{5}{9} \cdot (1 - \frac{5}{9}) = 0.494$$

$$H(Rleft) = \frac{3}{4}(1 - \frac{3}{4}) + \frac{1}{4} \cdot (1 - \frac{1}{4}) = 0.375$$

$$H(Rright) = \frac{1}{5}(1 - \frac{1}{5}) + \frac{4}{5} \cdot (1 - \frac{4}{5}) = 0.32$$

$$IG(R) = 0.494 - \frac{4}{9} \cdot 0.375 - \frac{5}{9} \cdot 0.32 = 0.15$$

Для регрессии

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - c)^2$$

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_j, y_j) \in R} y_j \right)^2$$

МФТИ

Практика
(построение дерева решений)

Критерии останова

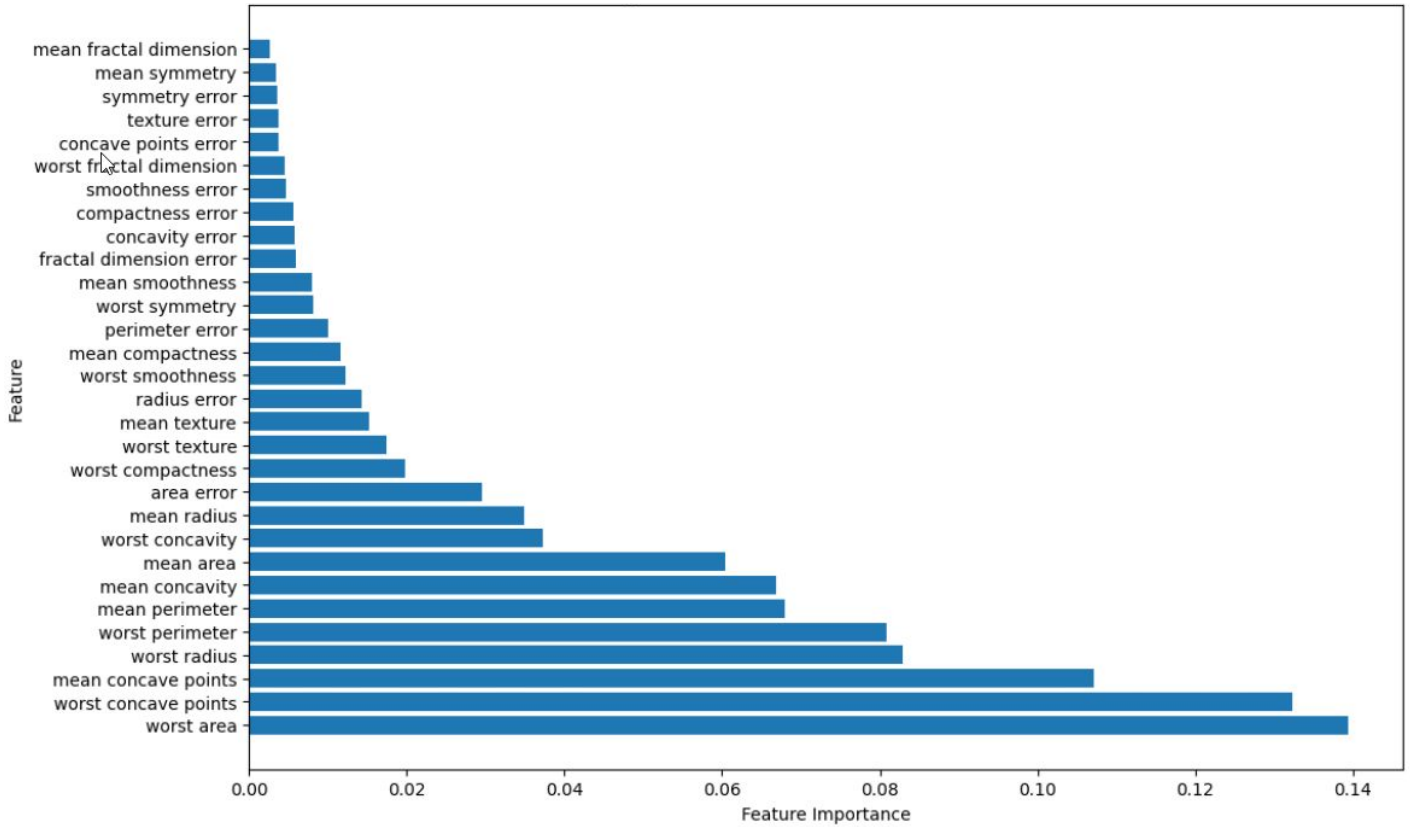
Критерии останова

МФТИ

- Ограничение максимальной глубины дерева
- Ограничение минимального числа объектов в листьях

Влияние признаков

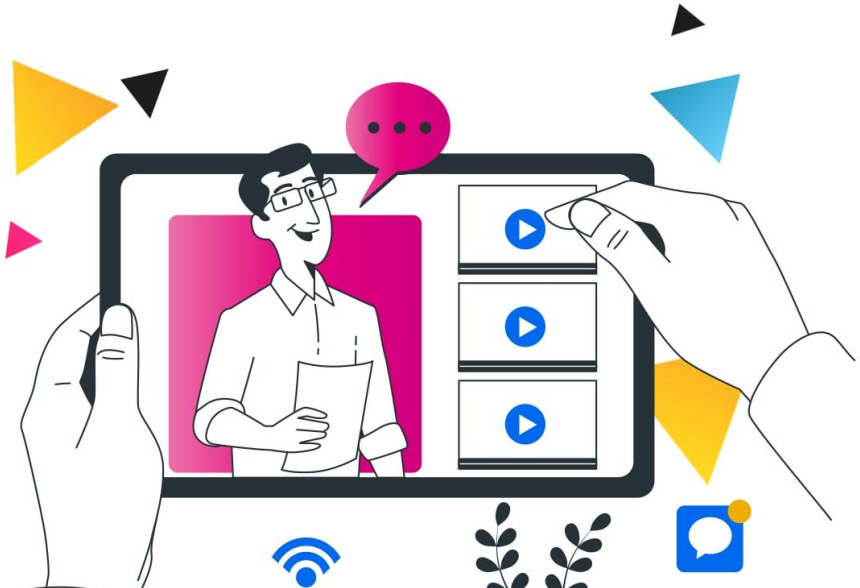
Feature_importances_



МФТИ

Практика
(DecisionTree из sklearn)

Ваши вопросы?



Итоги занятия

Итоги занятия

1. Изучили алгоритм построения дерева решений
2. Узнали, какие есть информационные критерии
3. Познакомились с критериями остнова
4. Подобрали оптимальные гиперпараметры для дерева решений
5. Получили важность признаков для дерева решений

Дополнительные материалы

МФТИ

1. Дерево решений для задачи регресии https://youtu.be/0mMeaC3gjNI
2. Дерево решений для задачи классификации https://youtu.be/j8L07nuns2Y
3. Критерии останова дерева решений https://youtu.be/aWEdaXAZ01M
4. Классификация, деревья решений https://habr.com/ru/company/ods/blog/322534/
5. Энтропия и деревья принятия решений https://habr.com/ru/post/171759/

МФТИ

Пожалуйста, оставьте свой отзыв о семинаре

До встречи!