Занятие 4
Деревья решений

Энтропия

$p_1 = \frac{4}{6}$

$p_2 = \frac{2}{6}$

$$E(Цвет) = -\sum_{c=1}^{n} p_c \cdot log_2(p_c)$$

n — количество классов

p_c — вероятность того, что случайно выбранный элемент принадлежит с-му классу

$$E = -\frac{5}{8} \cdot log_2\left(\frac{5}{8}\right) + \frac{3}{8} \cdot log_2\left(\frac{3}{8}\right) \approx 0.95$$

😉Можно показать, что данный расчёт энтропии полностью эквивалентен расчёту по формуле кросс-энтропии, которую мы применяли на занятии по логистической регрессии.

Условная энтропия

группа 1

группа 2

Разделим данные на группы по размеру и посчитаем энтропию в каждой.

$$E_1 = -\left[\frac{4}{5}log_2\left(\frac{4}{5}\right) + \frac{1}{5}log_2\left(\frac{1}{5}\right)\right] \approx 0.72$$

$$E_2 = -\left[\frac{1}{3}log_2\left(\frac{1}{3}\right) + \frac{2}{3}log_2\left(\frac{2}{3}\right)\right] \approx 0.91$$

$$E(Цвет \mid размер) = \frac{n_1}{N}E_1 + \frac{n_2}{N}E_2$$

Энтропия по цвету при условии разделения по размеру

$$E_{new} = \frac{5}{8} \cdot 0.72 + \frac{3}{8} \cdot 0.91 \approx 0.80$$

Насколько хорошим было разделение?

Регрессия: Выигрыш в среднеквадратичной ошибке

$$MSE_G = MSE - MSE_{new}$$

посчитано по выборке до разделения

взвешенная сумма MSE по группам после разделения (то же самое, что и MSE, посчитанное сразу для всех элементов с учётом разделения)

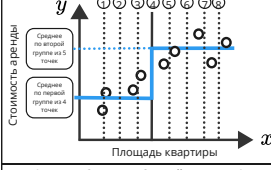
Классификация: Информационный выигрыш (Information Gain)

$$I_G = E - E_{new}$$

Разделение по признаку "размер" в нашем примере снижает энтропию на 0.15:

$$I_G = 0.95 - 0.8 = 0.15$$

Возможные разделения выборки

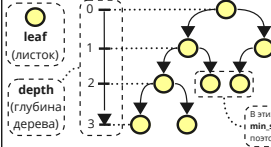


Выбирается одно из 8 разделений, которое обеспечивает максимальный выигрыш в MSE.

Возможные разделения выборки

Выбирается одно из 7 разделений, которое обеспечит максимальный информационный выигрыш.

Параметры решающего дерева



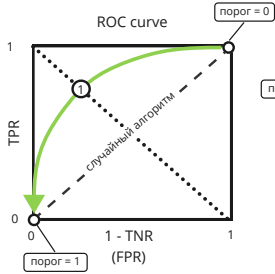
Остальные параметры:
criterion(default='gini') — критерий выбора наилучшего разделения.
max_depth(default=None) - максимальная глубина дерева
min_impurity_decrease (default=None) — минимальный выигрыш в энтропии для разделения.

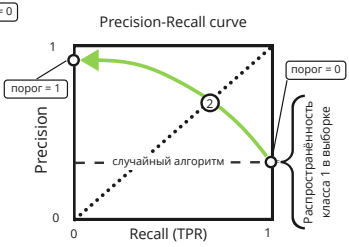
😉 Я прошу вас обратить внимание, что дерево для регресии строится так, что функция потерь MSE на обучающей выборке будет минимальной из возможных. Дерево для классификации строится так, что функция потерь CE на обучающей выборке будет минимальной из всех возможных. То есть **как деревья, так и линейные методы, изученные на занятиях 2 и 3 делают одно и то же, только разными способами.**

Занятие 4

Деревья решений

Автоматический выбор порога для задачи классификации





На графике представлен один из способов выбора порога. Этот подход называется «нормалищзацией алгоритма», потому что при таком выборе порога доля ложноположительных среди класса 0 равна доле ложноотрицательных ответов среди класса 1

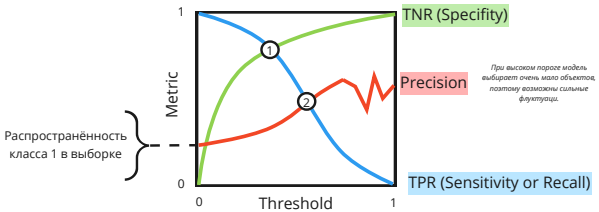
На графике представлен второй метод выбора порога, при котором Чувствительность будет равна точности. Если Precision-Recall-кривая выпукла вправо вверх (как на картинке), то данный метод будет максимизировать F1 score (среднее гармоническое между Precision и Recall).

① $TPR = TNR$

Алгоритмически выбирают точку, в которой TPR - TNR —> min, поскольку тождественного равенства при выборе порога не наблюдается.

② $Precision = Recall \quad F_1 \to max$

Алгоритмически выбирают точку, в которой Precision - Recall —> min, поскольку тождественного равенства при выборе порога не наблюдается.



На этой картинке представлен график метрик от порога напрямую, что является более наглядным представлением.

Примеры прикладных задач по выбору порога

1. Модель, которая ищет рак на снимке. Требуется высокий Recall, чтобы было **как можно меньше ложноотрицательных**. Низкий Precision - это не проблема.
2. Модель, которая ищет золото или нефть для постройки добывающего предприятия. Требуется очень высокий Precision, чтобы было **как можно меньше ложноположительных**. Если мы потратим деньги на постройку шахты, а нефти в скважине не окажется, мы впустую потратили огромное количество денег.