Задача ($+\varepsilon$)

Может ли метрика Precision принимать значение $\log_{10} 3$?

Метрика Precision считается следующим образом:

$$Precision = \frac{TP}{TP + FP}$$

TP и FP являются целыми положительными числами, т.к. подсчитывают количество исходов.

Попробуем пойти от обратного и доказать, что Precision *может* принимать такое значение. Сумма двух целых чисел также будет целым числом, поэтому мы можем представить $\frac{TP}{TP + FP}$ как $\frac{m}{n}$, где $n \geqslant m > 0$.

В таком случае должно быть верным следующее: $log_{10}3 = \frac{m}{n}$

Т. е. возведя 10 в степень $\frac{m}{n}$ мы должны получить три:

$10^{\frac{m}{n}} = 3$

Однако получаем противоречение: $10^m = 3^n$

Такое уравнение в целых числах не имеет корней кроме нулей, а в этом случае невозможно посчитать и сам Precision

Ответ: метрика precision не может быть равна $log_{10}3$

Задача (+2ε)

Найдите наилучшее константное предсказание (или покажите, что его не существует) для метрик:

- ROC AUC
- Accuracy
- Recall
- Precision
- Mean Absolute Error

- **ROC AUC**
  В данном случае главным будет являться разделяющая способность модели, т.е. порядок объектов, отсортированных по таргету, а т.к. он в нашем случае всегда будет одинаков у всех объектов, то и не важно, какое константное предсказание мы выберем, от конкретного его значения ничего не изменится. Итоговое значение метрики будет зависеть только от истинных значений и заранее его угадать невозможно.
- $\text{Accuracy} = \dfrac{\text{Количество правильных предсказаний}}{\text{Количество предсказаний}}$
  Для данной метрики лучшим предсказанием будет самый популярный класс объектов, а лучшим значением метрики доля этого класса. Если у самого популярного объекта было 70% в выборке, то мы сможем получить accuracy 0.7
- $\text{Recall} = \dfrac{TP}{TP + FN}$
  Лучшим предсказанием будут единицы, ведь мы максимизируем количество True Positive, не увиличивая False Negative ни на единицу.
- $\text{Precision} = \dfrac{TP}{TP + FP}$
  Лучшим предсказанием снова будут все единицы, потому что в противном случае значение метрики будет равно нулю, но в данном случае, в отличие от Recall, мы не добьёмся значения метрики 1.0, т.к. в знаменателе добавятся ложноположительные случаи.
- **MAE**
  Можем показаться, что лучшим вариантом было бы среднее значение, однако на самом деле это не так: даже одного примера в случае с данными [1, 5, 1000] заметно, что среднее 335 даст плохой результат по сравнению с медианой. Например, у нас есть n + m объектов. Поставим после объекта n некоторую точку, слева от которой окажется n объектов, а справа - m. Если мы сдвинем точку на небольшое расстояние d, не перехода при этом границу крайних объектов, то средняя ошибка изменится на $\dfrac{d * (n - m)}{(n + m)}$. Средняя ошибка будет

уменьшаться тем сильнее, чем больше мы двигаемся к большему количество точек и будет оптимальной, когда слева и справа окажется одинаковое их количество, следовательно, медиана является лучшим константным предсказанием.

Задача ($+\varepsilon$)

Согласно стандартным настройкам (в sklearn), ExtraTrees не использует bagging. Почему?

В отличие от случайного леса, где при помощи жадного алгоритма выбирается набор признаков, на основе которых производится разбиение, в ExtraTrees это происходит случайно, поэтому нет прямой необходимости добавлять еще больше случайности, также добавляя бэггинг. Как результат, ExtraTrees будет иметь чуть более слабую обобщающую способность, но чуть более высокую точность.