

Библиотеки Python для Data Science:

продолжение

Урок 1. Введение в задачу классификации. Постановка задачи и подготовка данных.

1. Приведите по 2 примера, когда лучше максимизировать Precision, а когда Recall.

Precision - точность отражает какой процент объектов, отнесённых к "основному" классу, действительно являются таковыми.

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

- Если кредитная политика банка направлена на репутацию (макс. клиентов платежеспособны), имеет смысл максимизировать Precision.
- При оценке на сейсмоактивность лучше максимизировать Precision, причислив место к сейсмоактивным.

Recall - полнота отражает какой процент объектов "основного" класса было обнаружено алгоритмом.

$$recall = \frac{TP}{TP + FN}$$

- Если кредитная политика банка направлена на большую выручку, имеет смысл максимизировать Recall.
- Если при каком-либо серьезном заболевании речь идет о жизни человека, то выбор метода лечения определятся исходя из максимизации Recall.

2. Почему мы используем F-меру, почему, например, нельзя просто взять среднее от Precision и Recall?

F-score (*F-мера*, гармоническое среднее) - метрика объединяющая в себе точность и полноту, выдавая как результат их гармоническое среднее, а коэффициент β , позволяет сбалансировать вклад, вносимый точностью, тем самым акцентируя внимание на той метрике, максимизация которой для нас наиболее приоритетна.

Наиболее часто встречающаяся вариация данной метрики - это *F1-score*, когда $\beta = 1$, а влияние точности и полноты одинаково.

$$F_{\beta} = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

- F-мера позволяет оценить в комплексе как Precision, так и Recall.
- F-мера увеличивает или уменьшает влияние Precision и Recall в зависимости от выбранного приоритета