Машинное обучение с учителем

Андрей Куртасов
Системный аналитик

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

О чем поговорим сегодня

МФТИ

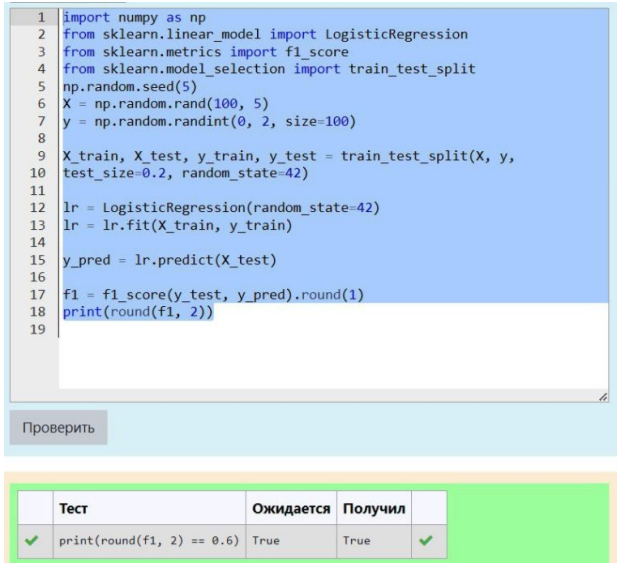
1. Разберем ваши вопросы.
2. Обсудим метрики качества МО с учителем.
3. Потренируемся в настройке конвейера для классификации на примере задачи анализа тональности текста.

Консультация и ответы на вопросы

Вопрос

Алексей Круглов:

В задании 4 последнего модуля не понятно на каком random.seed() тест генерирует выборку. Результат сильного будет зависеть от генерации чисел. Путем подбора нашел примерный random.seed(5).



```
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score
from sklearn.model_selection import train_test_split
np.random.seed(5)
X = np.random.rand(100, 5)
y = np.random.randint(0, 2, size=100)

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

lr = LogisticRegression(random_state=42)
lr = lr.fit(X_train, y_train)

y_pred = lr.predict(X_test)

f1 = f1_score(y_test, y_pred).round(1)
print(round(f1, 2))
```

Проверить

	Тест	Ожидается	Получил	
✔	print(round(f1, 2) == 0.6)	True	True	✔

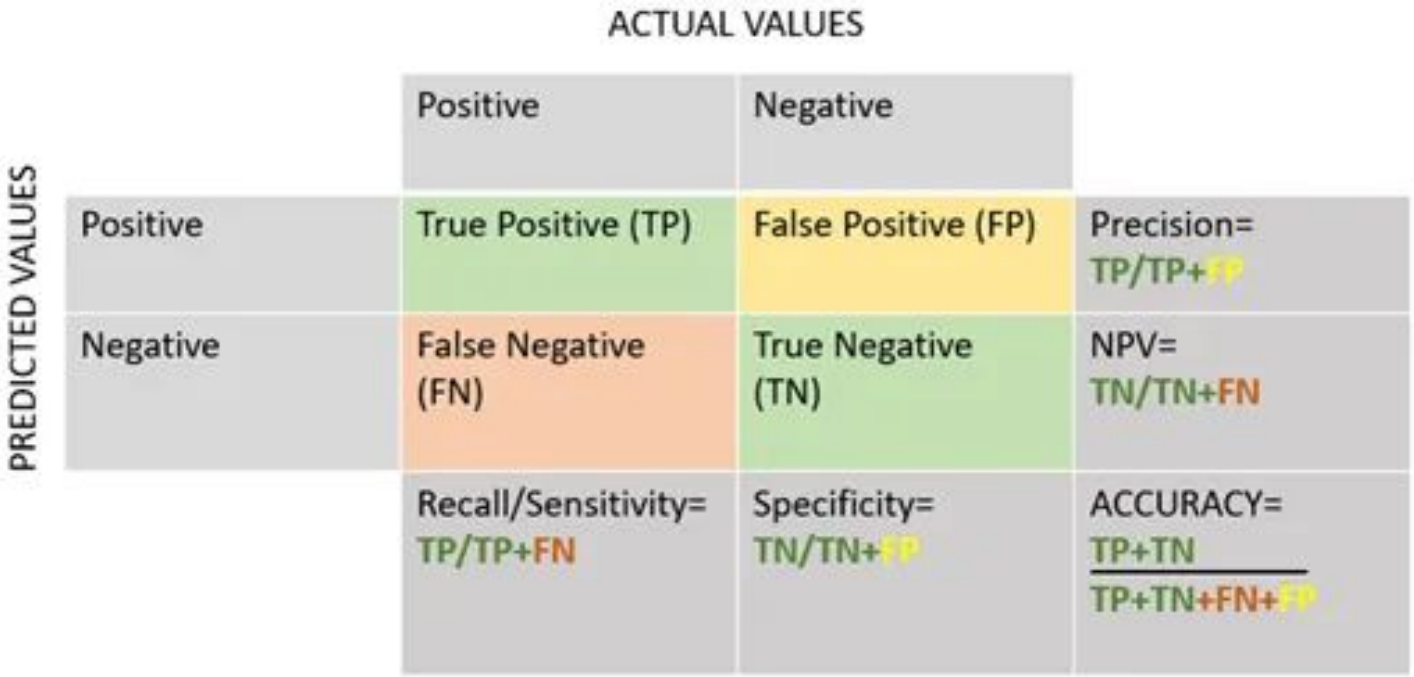
Метрики качества

Регрессия: меры разности между прогнозами и фактическими значениями целевой переменной

МФТИ

- Mean Squared Error (MSE) — средняя квадратичная ошибка;
- Root Mean Squared Error (RMSE) — среднеквадратичное отклонение;
- Mean Absolute Error (MAE) — среднее значение модуля разностей между фактическими и прогнозируемыми значениями;
- R-squared (R2) — коэффициент детерминации.

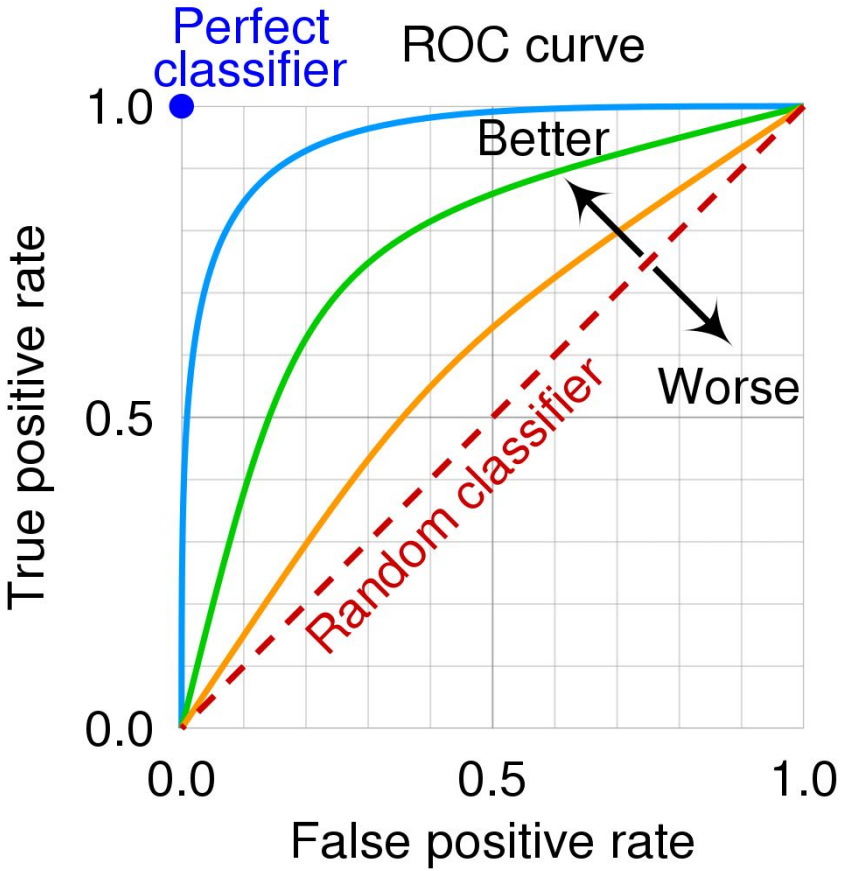
Классификация: ошибки I-го и II-го рода



$$Fmeasure = \frac{2Precision \cdot Recall}{Precision + Recall}$$

Источник изображения:
towardsdatascience.com/baffling-concept-of-true-positive-and-true-negative-bffbc340f107

ROC & AOC



На заметку:
https://github.com/catboost/tutorials/blob/master/metrics/AUC_tutorial.ipynb

Источник изображения:
en.wikipedia.org/wiki/Receiver_operating_characteristic

Упражнение 1

Постройте ROC-кривую и определите площадь AOC для задачи логистической регрессии.

Можно использовать библиотеку `plot-metric`.

Упражнение 2: sentiment analysis

1. Воспользуемся датасетом с отзывами о фильмах: https://www.kaggle.com/c/word2vec-nlp-tutorial
2. В качестве признаков используем векторные представления текстов (отзывов) – можно использовать CountVectorizer.
3. Векторизатор объединяем с лог-регрессией (или др. классификатором через Pipeline).

МФТИ

Свободная дискуссия

Ваши вопросы? Пожелания?

До встречи!