

Юстина Иванова

Программист, data scientist

Математическая статистика в python.

Кейс-стади 2.

Поиск болезни сердца.

Поиск СПАМа в тексте.

Анализ текста на тональность.

Спикер



Юстина Иванова,

- PhD в университете Больцано (Италия)
- Data scientist по компьютерному зрению в компании ОЦРВ, Сочи
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton (Англия)

Теорема Байеса

вероятность того, что событие В
истинно, если событие А истинно

↓

вероятность того, что
событие А истинно

↷

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

↑

вероятность того, что
событие А истинно, если
событие В истинно

↶

вероятность того, что
событие В истинно

Пример.

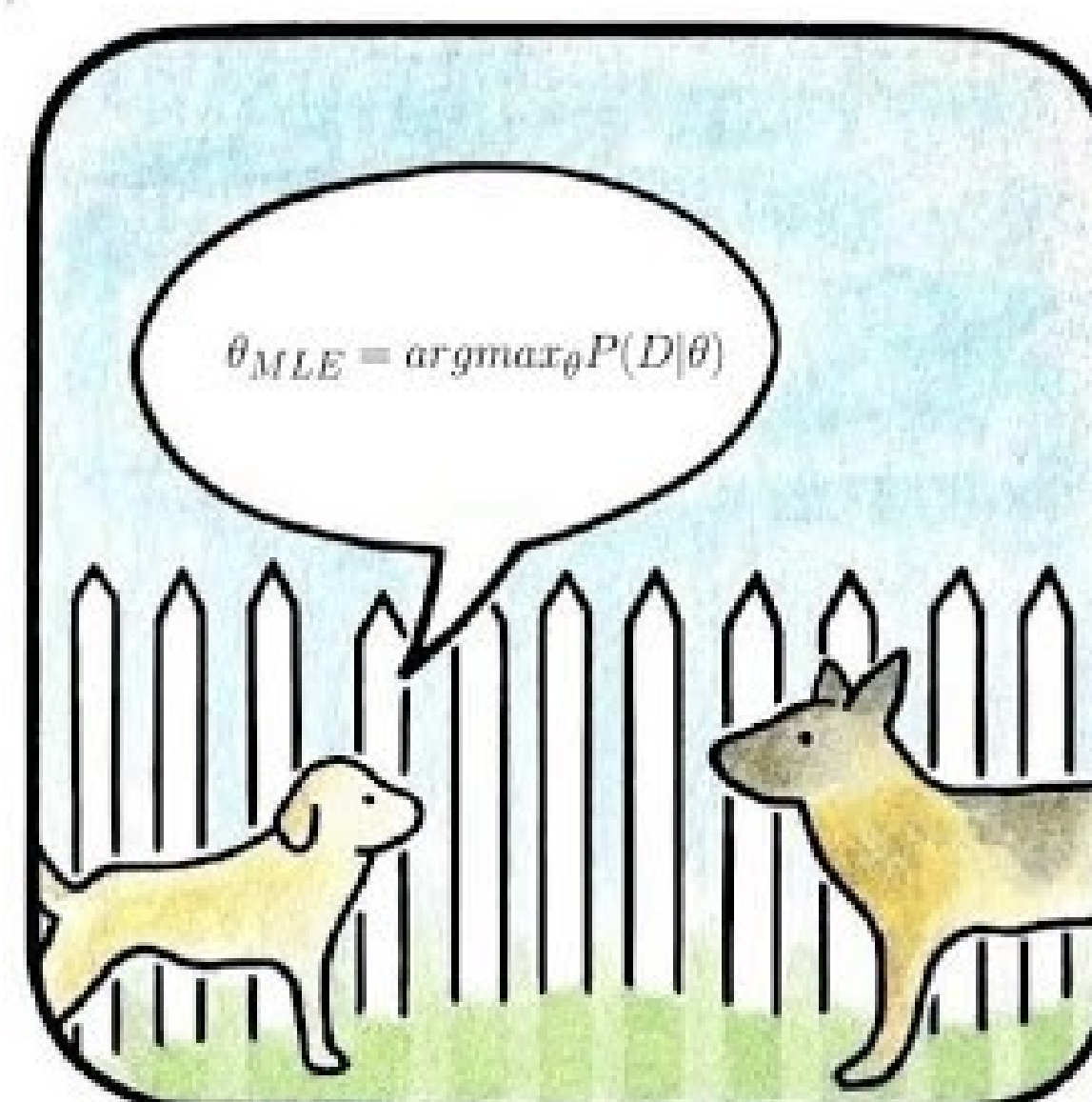
1% женщин в возрасте 40 лет, участвовавших в регулярных обследованиях, имеют рак груди. 80% женщин с раком груди имеют положительный результат маммографии. 9.6% здоровых женщин также получают положительный результат (маммография, как любые измерения, не дает 100% результатов). Женщина-пациент из этой возрастной группы получила положительный результат на регулярном обследовании. Какова вероятность того, что она фактически больна раком груди?

Пример.

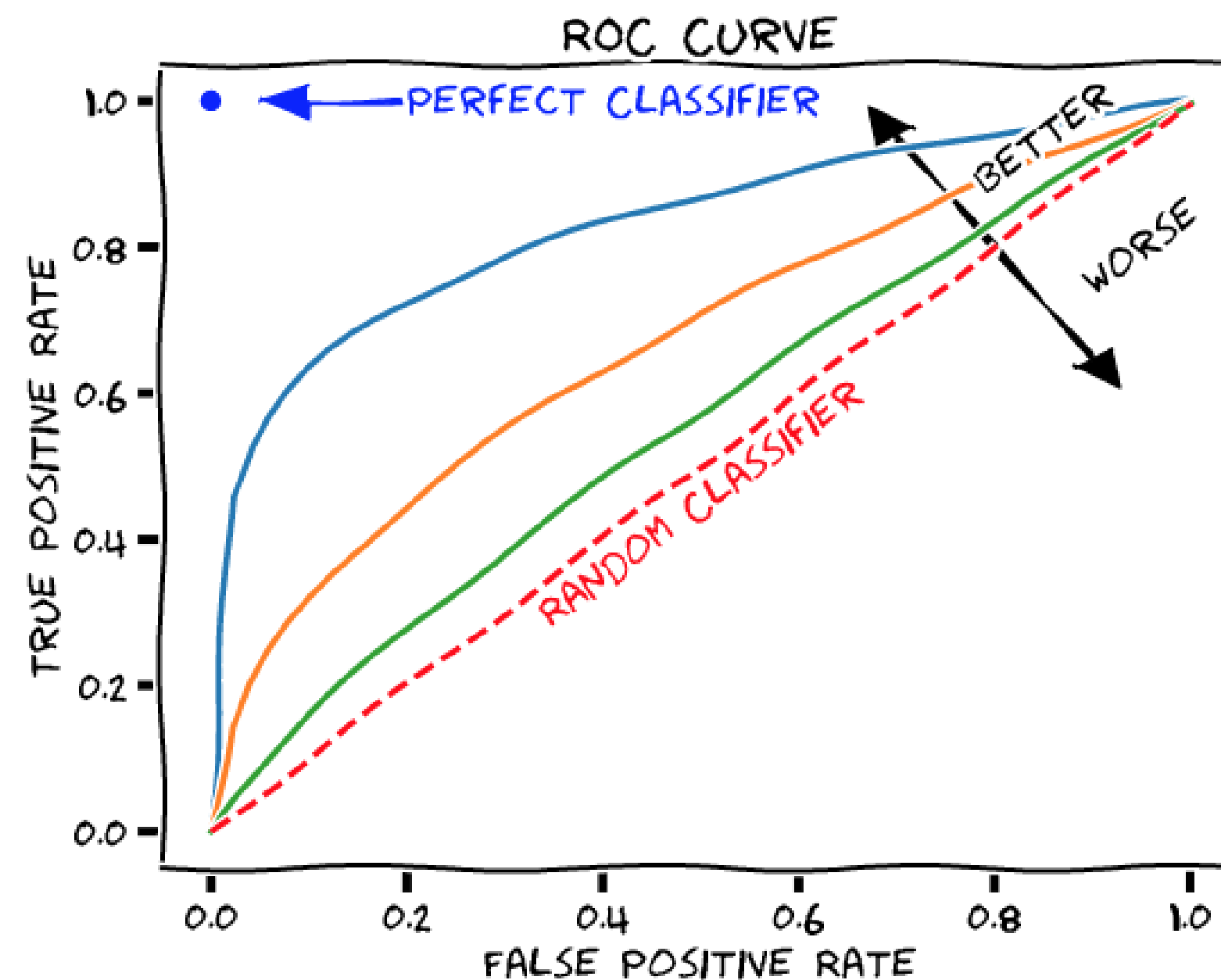
1% женщин в возрасте 40 лет, участвовавших в регулярных обследованиях, имеют рак груди. 80% женщин с раком груди имеют положительный результат маммографии. 9.6% здоровых женщин также получают положительный результат (маммография, как любые измерения, не дает 100% результатов). Женщина-пациент из этой возрастной группы получила положительный результат на регулярном обследовании. Какова вероятность того, что она фактически больна раком груди?

Правильный ответ - 7.8%, и получается он так: Из 10,000 женщин, 100 больны раком груди; 80 из этих 100 имеют положительные маммограммы. Из тех же 10,000 женщин, 9,900 не имеют рака груди, и из этих 9,900 женщин 950 тоже получают положительные маммограммы. Таким образом, общее число женщин с положительными маммограммами $950 + 80$ то есть 1,030. Из этих 1,030 женщин с положительными маммограммами, 80 реально больны раком. Вычисляя, пропорцию, мы получаем $80/1,030$, или 0.07767, то есть 7.8%.

<http://scheigl2g.bget.ru/bayes/YudkowskyBayes.html>

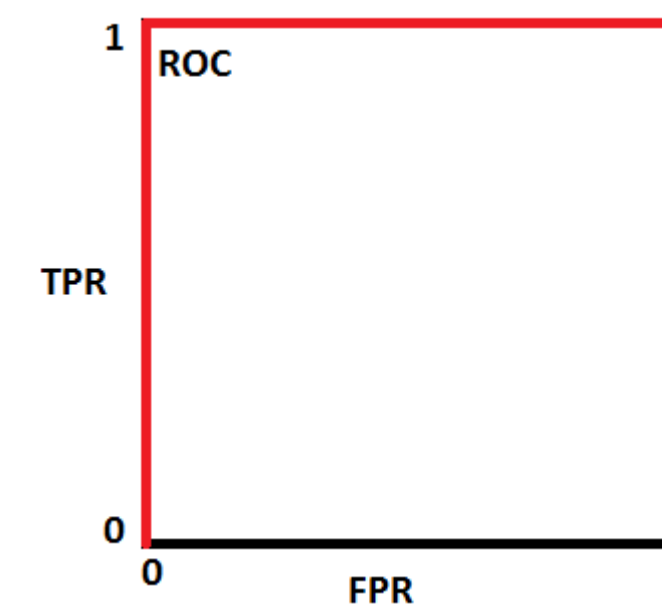
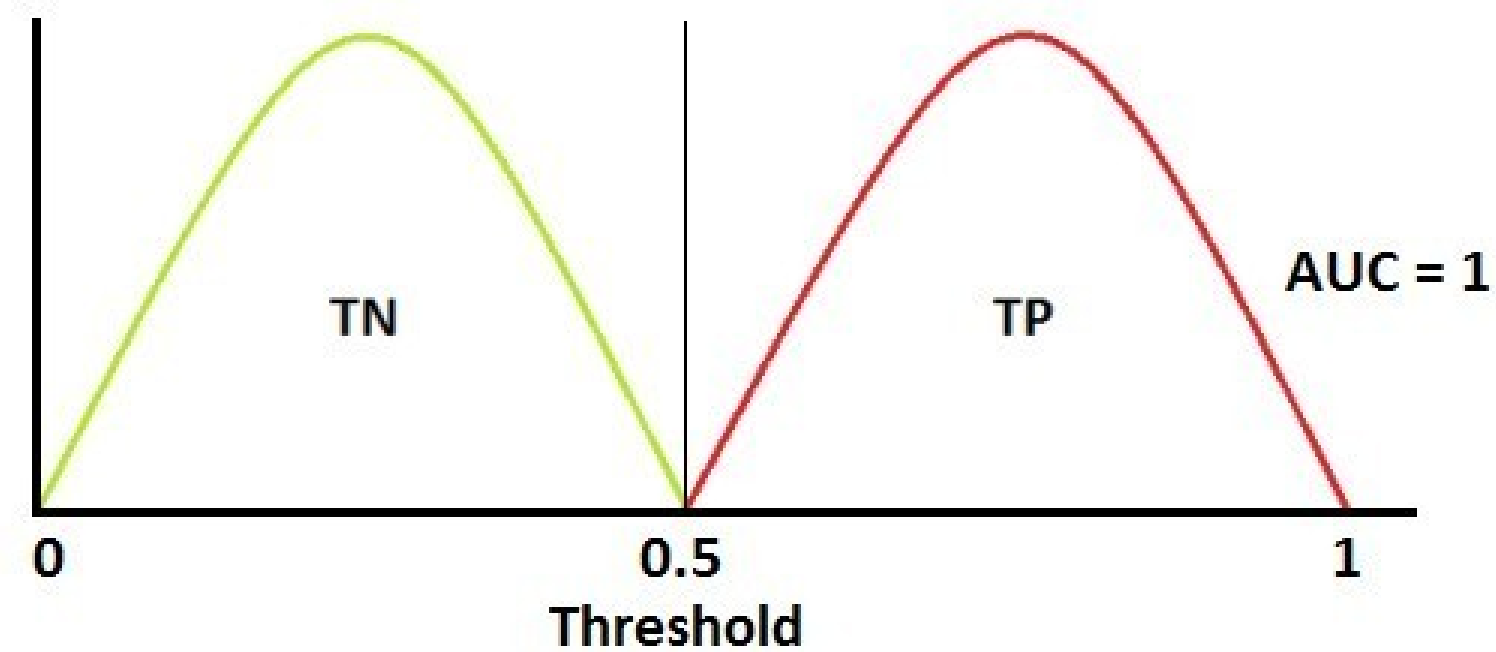


Метрики классификации: ROC-кривая

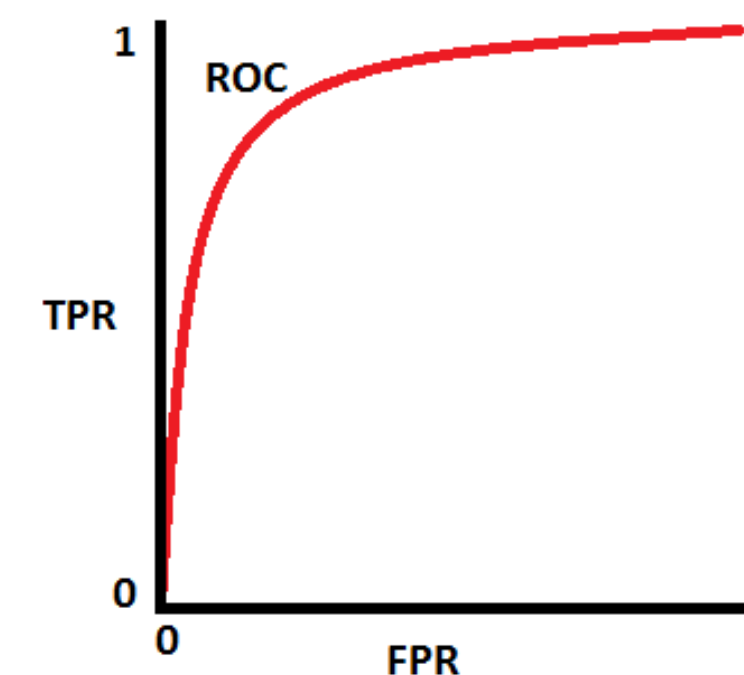
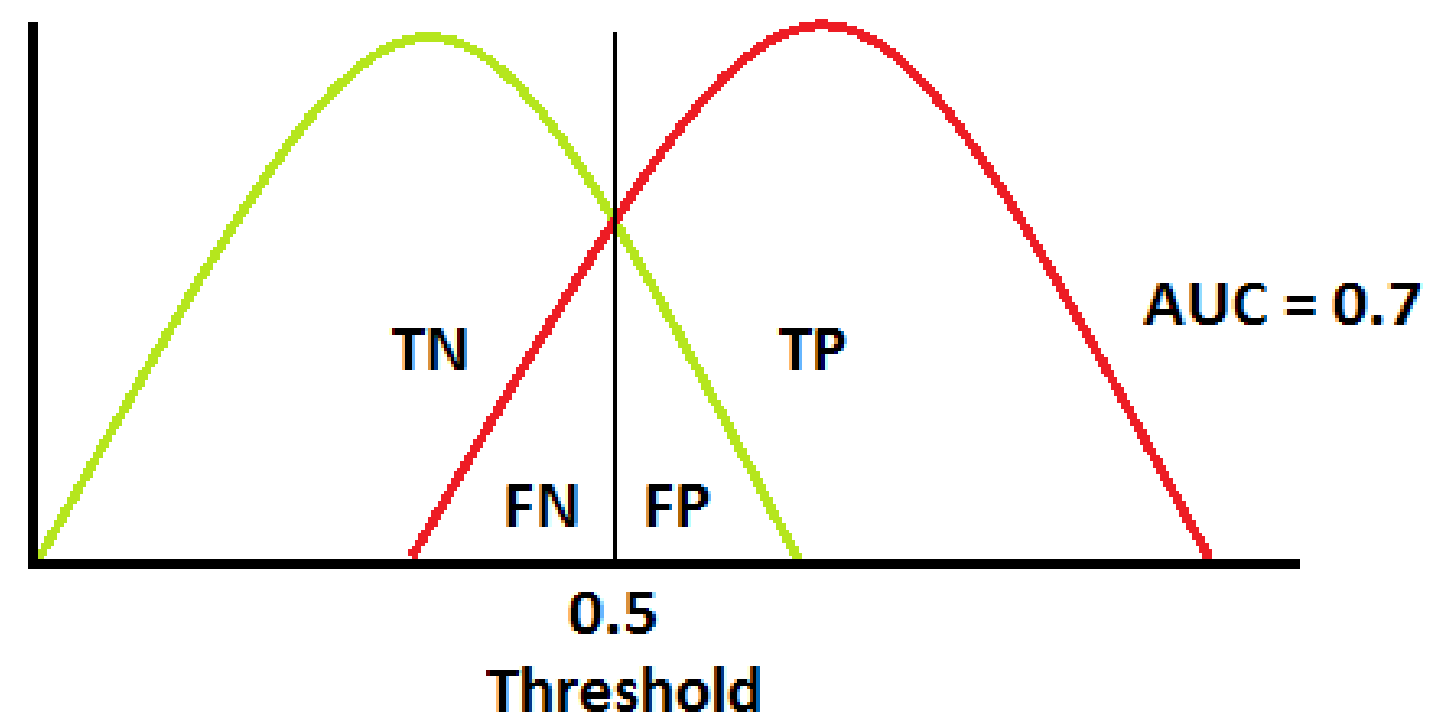


Позволяет определить порог,
при котором мы будем отделять один класс от другого

ROC-кривая



Идеальная модель — порог 50%



Модель с некоторыми ошибками — порог выбирается в зависимости от допускаемых ошибок

BAG of words - Мешок слов

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

15



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Спасибо за внимание

Вопросы?