Семинар 7-8 (часть первая): Метрики классификации

В этом семинаре мы подробнее поговорим про классификацию и метрики для неё. План будет таким:

- сформулируем задачу и поймём её специфику;
- поймём с помощью каких метрик можно оценить качество прогнозирования;
- попробуем разобраться какой смысл стоит за этими метриками.

Задача 1 (формулируем задачу)

Вася — меломан. Каждый день он смотрит на youtube музыкальные клипы и читает комментарии к ним. В один прекрасный день ему стало интересно, можно ли определить жанр клипа (рэп или попса) по его характеристикам: число комментов, лайков, характер комментов и тп.

1. К какому типу относится такая задача: классификация или регрессия? Почему?
2. Какие факторы из профилей вы бы использовали, чтобы её решить? Как бы вы оценивали качество итогового прогноза?

Задача 2 (метрики)

Бандерлог оценил три модели: нейросеть, случайный лес и KNN. Он построил на тестовой выборке прогнозы и получил три матрицы ошибок:

	$y = 1$	$y = 0$
$\hat{y} = 1$	80	20
$\hat{y} = 0$	20	80

	$y = 1$	$y = 0$
$\hat{y} = 1$	48	2
$\hat{y} = 0$	52	98

	$y = 1$	$y = 0$
$\hat{y} = 1$	10	20
$\hat{y} = 0$	90	10000

а) Найдите для всех трёх моделей долю правильных ответов. Чем плоха эта метрика?

б) Найдите для всех трёх моделей точность (precision) и полноту (recall)

в) Предположим, что целевая переменная y принимает значение 1, если заемщик вернул кредит и 0, если не вернул. Вы хотите научиться прогнозировать платежеспособность клиента. Какую из первых двух моделей вы бы выбрали в таком случае?

г) Предположим, что целевая переменная y принимает значение 1, если человек болен больной болезнью с болью и 0, если он здоров. Вы хотите спрогнозировать нужно ли человеку обследование. Какую из первых двух моделей вы б выбрали в этом случае?

Задача 3 (ещё немного метрик)

Бандерлог из Лога[1] ведёт блог, любит считать логарифмы и оценивать модели. С помощью нового алгоритма Бандерлог решил задачу классификации по трём наблюдениям и получил $b_i = \hat{P}(y_i = 1|x_i)$.

y_i	b_i
1	0.7
0	0.2
0	0.3
1	0.25

а) Найдите ROC AUC.

б) Постройте ROC-кривую.

в) Постройте PR-кривую (кривая точность-полнота).

г) Найдите площадь под PR-кривой.

д) Как по-английски будет «бревно»?

Ещё задачи!

Тут лежит ещё несколько задач для самостоятельного решения. Возможно, похожие будут в самостоятельной работе...

Задача 4

Бандерлог начинает все определения со слов «это доля правильных ответов»:

а) accuracy — это доля правильных ответов...

б) точность (precision) — это доля правильных ответов...

в) полнота (recall) — это доля правильных ответов...

г) TPR — это доля правильных ответов...

Закончите определения Бандерлога так, чтобы они были, хм, правильными.

Задача 5

Бандерлог обучил модель для классификации и получил вектор предсказанных вероятностей принадлежностей к классу 1.

[1] деревня в Кадуйском районе Вологодской области

y_i	b_i
1	0.9
0	0.1
0	0.75
1	0.56
1	0.2
0	0.37
0	0.25

а) Бинаризуйте ответ по порогу t и посчитайте точность и полноту для $t = 0.3$ и для $t = 0.8$.

б) Какой порог бы вы выбрали?

в) Постройте ROC-кривую и найдите площадь под ней.