Семинар 8: Метрики классификации

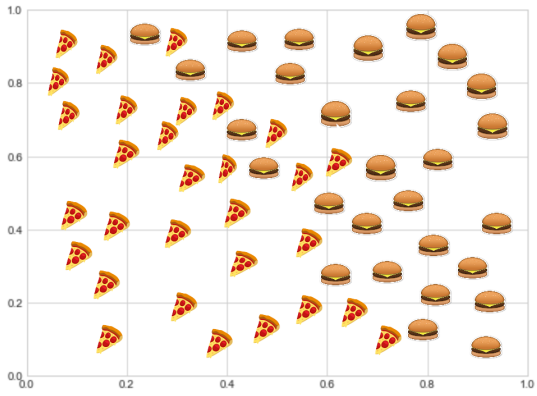
«Без бури нет величия!»
Винни-Пух перед тем как отправиться к
пчёлам в улей в виде тучки (1969)

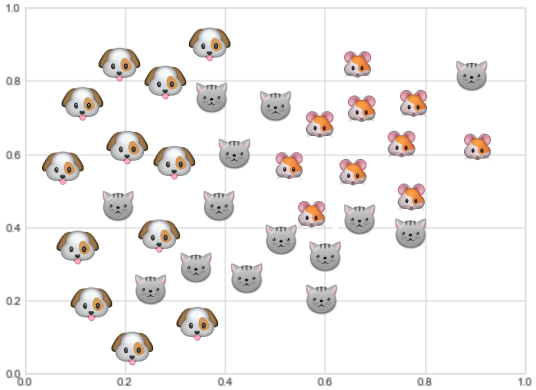
В этом семинаре мы подробнее поговорим про классификацию и метрики для неё. План будет таким:

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

Упражнение 1

Нам нужно научиться отделять пиццу от бургеров, а также котиков от пёсиков и от мышек. Проведите на картинках линии, которые отделят одни классы от других. Да, это и есть машинное обучение. Но обычно кривые рисуем не мы, а компуктер.





Почему нельзя провести между пиццей и бургерами слишком подробную и извилистую границу? В чём проблема самого правого верхнего котика? Что такое переобучение? Как понять переобучились ли мы?

Упражнение 2

Винни-Пух ищет неправильных пчёл. За долгие годы поиска он скопил довольно большую выборку и оценил на ней три модели: нейросеть, случайный лес и KNN. Он построил на тестовой выборке прогнозы и получил три матрицы ошибок:

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

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

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

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 Винни-Пух колектор. Пчела, по его мнению, неправильная, если она не возвращает кредит. Переменная y принимает значение 1, если пчела вернула кредит и 0, если не вернула. ВП хочет научиться прогнозировать платёжеспособность пчелы. Какую из первых двух моделей вы бы выбрали в таком случае?

г) Предположим, что Винни-Пух врач. Пчела, по его мнению, неправильная, если она умирает от болезни. Он хочет находить таких пчёл и лечить. Переменная y принимает значение 1, если пчела больна больной болезнью с болью и 0, если она здорова. ВП хочет спрогнозировать нужно ли пчеле пройти обследование. Какую из первых двух моделей вы б выбрали в этом случае?

д) Найдите для всех трёх моделей f1-меру.

Упражнение 3

Бандерлог из Лога[1] ведёт блог, любит считать логарифмы и оценивать модели[2]. С помощью нового алгоритма Бандерлог решил задачу классификации по трём наблюдениям и получил $\hat{p}_i = \hat{P}(y_i = 1|x_i)$.

y_i	$\hat{p}_i$
1	0.7
0	0.2
0	0.3
1	0.25
0	0.1

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...

Упражнение 4

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

[1] деревня в Кадуйском районе Вологодской области

[2] Читай больше про приключения Бандерлога тут: https://github.com/bdemeshev/mlearn_pro

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.

Упражнение 5

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

y_i	$\hat{p}_i$
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 вы бы выбрали?

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

Упражнение 6

Алгоритм бинарной классификации, придуманный Бандерлогом, выдаёт оценки вероятности $\hat{p}_i = \hat{P}(y_i = 1)$. Всего у Бандерлога 10000 наблюдений. Если ранжировать их по возрастанию $\hat{p}_i$, то окажется что наблюдения с $y_i = 1$ занимают ровно места с 5501 по 5600. Найдите площадь по ROC-кривой и площадь под PR-кривой.

Упражнение 7

Для задачи классификации есть довольно много разных метрик, которые оценивают качество модели для её решения. На практике иногда оказывается, что по одной метрике наша модель может иметь более высокое качество, чем другая, а по другой метрике — более низкое.

Предположим, что у нас есть две модели, предсказывающие принадлежность к одному из классов: котики, 1 или собачки, 0. Качество моделей оценивается на выборке из пяти объектов. Оказывается, что по метрике A первая модель лучше, а по метрике B, вторая. Вам нужно придумать примеры тестовых меток y и предсказаний $\hat{y}_1$ и $\hat{y}_2$ для каждой пары метрик A и B:

а) A — precision (при пороге 0.5), B — auc-roc;

б) A — precision (при пороге 0.5), B — recall (при пороге 0.5);

в) A — F1-score (при пороге доставляющем максимум), B — auc-roc.

Упражнение 8

Винни-Пух пришёл в лесу к власти у становил свою медвежью диктатуру. Верные Винни-Пуху медведи случайно рассредоточились по лесу и обнюхивают его. Их главная задача — поиск правильных пчёл для изъятия у них мёда. После каждого изъятия медведи записывают в книжечку всех пчёл и обстоятельства, в которых они были найдены.

В один прекрасный день Винни-Пух решил обучить классификатор для поиска правильных пчёл. Он это сделал на сбалансированной обучающей выборке (взял по 5000 плохих и хороших пчёл), а после на сбалансированной тестовой выборке (по 500 пчёл) он посчитал метрики качества.

а) В качестве порога Винни взял 0.5. Точность (precision) получилась 0.9. Полнота (recall) получилась 0.7. Как выглядит матрица ошибок (confusion matrix)?

б) В природе всего лишь 10% пчёл правильные. Медведи зашли в гости к случайной 1000 пчёл. Сколько из них классификатор назовёт правильными? Сколько раз он ошибётся? Нарисуйте для это выборки матрицу ошибок и найдите precision и recall.

в) Правда ли, что ни precision ни recall не поменялись при переходе от сбалансированной тестовой выборки к природной? Как сделать так, чтобы на практике не задумываться о таких переходах?

Упражнение 9

Настя модерирует классификатор спама. Ей хочется понять когда он устареет, испортится и надо будет научить новый. Каждый день она берёт 100 забаненных наблюдений и 100 не забаненных, размечает их, рисует матрицу ошибок и считает по ней Precision и Recall. Правда ли, что у Насти получится таким образом адекватно оценить эти две метрики?

Упражнение 10

Винни-Пух наладил с бандерлогами торговлю мёдом. Бандерлоги любят мёд. Многие Бандерлоги стали постоянными клиентами Винни-Пуха. Иногда Бандерлогам надоедает мёд Винни-Пуха, они начинают унывать и отваливаются от его постоянной клиентуры.

Ситуация, когда Бандерлоги унывают называется **оттоком**. Чтобы бороться с ним, Винни-Пух позвал медведей-аналитиков и вместе они обучили классификатор, который прогнозирует вероятность оттока для каждого конкретного Бандерлога. Винни попробовал посмотреть что будет происходить с моделью при двух разных порогах (0.5 и 0.8):

	$y = 0$	$y = 1$
$\hat{y} = 0$	70	10
$\hat{y} = 1$	20	30

	$y = 0$	$y = 1$
$\hat{y} = 0$	80	20
$\hat{y} = 1$	10	20

Один горшочек мёда стоит 100\$. Если модель спрогнозировала отток, Бандерлогу будет даваться скидка в 10%. Средний доход от одного Бандерлога за месяц составляет 300\$. С помощью тестовой выборки ответьте на следующие вопросы:

а) Сколько баксов мог бы заработать Винни за месяц, если бы никто из Бандерлогов не уходил? Сколько баксов потерял бы Винни, если бы он давал бы Бандерлогам спокойно уходить?

б) Сколько баксов для обеих моделей Винни потратит на скидки? Сколько баксов будет потрачено впустую? Какой суммарный средний доход Винни сможет удержать благодаря модели? Выгодно ли это?

в) Придумайте на основе всех этих расчётов бизнес-метрику, на основе которой Винни-Пух мог бы выбрать порог для своей модели.

г) Как правильно выкатить такую модель в продакшн (включить работать в режиме реального времени) так, чтобы никто не пострадал?