Семинар 8: Метрики классификации

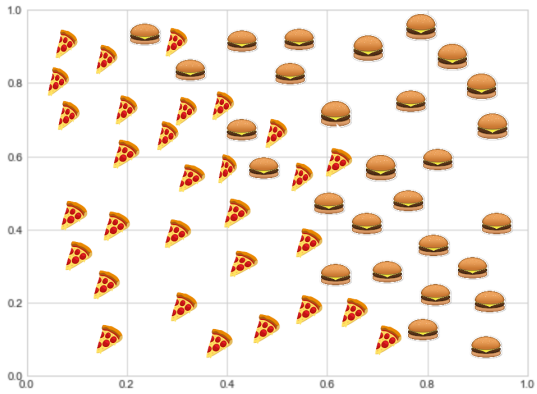
«Без бури нет величия!»
Винни-Пух перед тем как отправиться к
пчёлам в улей в виде тучки (1969)

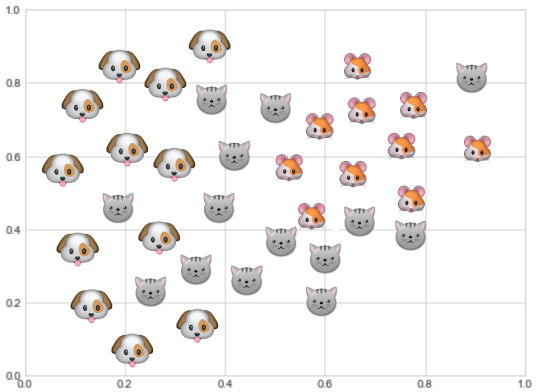
В этом семинаре мы подробнее поговорим про классификацию и метрики для неё. План будет таким:

- 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

(KNN) На плоскости расположены колонии рыжих и чёрных муравьёв. Рыжих колоний три и они имеют координаты $(-1,-1)$, $(1,1)$ и $(3,3)$. Чёрных колоний тоже три и они имеют координаты $(2,2)$, $(4,4)$ и $(6,6)$.

а) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод одного ближайшего соседа.

б) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод трёх ближайших соседей.

в) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей k перебрав $k \in \{1,3,5\}$. Целевой функцией является количество верных предсказаний (accuracy).