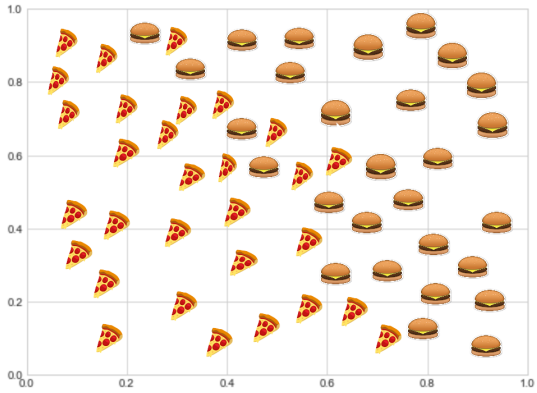
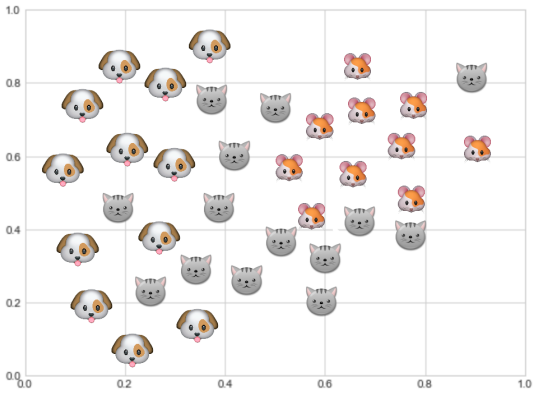
Семинар 7-8: Классификация

Задача 1 (классификация в картинках)

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?

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»?

Задача 4 (KNN, кросс-валидация)

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 используя метод одного ближайшего соседа.

б) Поделите плоскость на «зоны влияния» рыжих и чёрных 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 несовпадающих прогнозов.

Задача 5 (древо для классификации)

Машка пять дней подряд гадала на ромашке, а затем выкладывала очередную фотку «Машка с ромашкой» в инстаграмчик. Результат гадания — переменная y_i, количество лайков у фотки — переменная x_i. Постройте классификационное дерево для прогнозирования y_i с помощью x_i на обучающей выборке:

[1] деревня в Кадуйском районе Вологодской области

y_i	x_i
плюнет	10
поцелует	11
поцелует	12
к сердцу прижмёт	13
к сердцу прижмёт	14

Дерево строится до идеальной классификации. Критерий деления узла на два — минимизация числа допущенных ошибок[2]. Правило прогнозирования в каждой вершине: в качестве прогноза выдаем тот класс, представителей которого в вершине больше. Предположим, что под фоткой стоит 15 лайков, каков будет результат гадания?

Ещё задачи

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

Задача 6

Пятачок собрал данные о визитах Винни-Пуха в гости к Кролику. Здесь x_i - количество съеденного мёда в горшках, а y_i - бинарная переменная, отражающая застревание Винни-Пуха при входе

y_i	x_i
0	1
1	4
1	2
0	3
1	3
0	1

а) Пятачок собирается оценить дерево по всей выборке. Помогите очень маленькому существу сделать это.

б) Пятачок узнал у Иа-Иа, что оказывается выборку надо делить на тренировочную и тестовую Поэтому он отложил последние два наблюдения для теста. Оцените дерево по первым четырём наблюдениям и проверьте его работоспособность по последним двум.

[2]На самом деле на практике так не делают. Обычно для разбиения узла при строительстве классификационных деревьев используют энтропию. О том, что это такое, можно погуглить.

Задача 7

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 8

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

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.