Фамилия, имя, номер группы:

. . . . . .

Внесите сюда ответы на тест:

Вопрос	1	2	3	4	5	6	7	8	9	10
Ответ										

Табличка для проверяющих работу:

Вопрос	11	12	13	14	15	16	17	Итого
Баллы								

«Если орел — я выиграла, если решка — ты проиграл»

Рейчел Грин, сериал "Друзья"

Работа состоит из трёх частей: тестовая, задачи и ответы на открытые вопросы. Работа пишется на раздаточном материале. Черновики можно использовать, но не сдавать - их не проверяем. Списывание карается обнулением работы. Удачи!

Часть первая: тестовая

Дайте ответ на 10 тестовых вопросов. Каждый вопрос стоит 3 балла. Никаких дополнительных пояснений в этой части работы от вас не требуются.

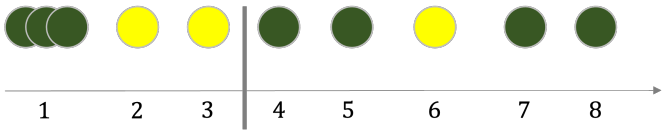
Вопрос 1. Для решения задачи регрессии не подходит алгоритм

[A] решающего дерева.
[B] случайного леса.
[C] логистической регрессии.
[D] градиентного бустинга.
[E] KNN.
[F] Нет верного ответа.

Вопрос 2. Джоуи не делится едой! Жадность ли это, мы не узнаем, а что верно для жадного построения дерева?

[A] Если разбиение сделано, оно не может быть изменено.
[B] Признаки выбираются случайно для разбиения узла.
[C] Дерево строится, пока не останется по 1 объекту в листьях.
[D] Дерево строится, пока не останутся объекты одного класса в листьях.
[E] Признаки выбираются из случайного подмножества.
[F] Нет верного ответа.

Вопрос 3. Посчитайте хаотичность разбиения вершины. В качестве критерия информативности используйте критерий Джини.



A 0.08
C 0.12
E 0.24
B 0.16
D 0.4
F Нет верного ответа.

Вопрос 4. Росс прогнозирует возраст останков динозавра. Обучает градиентный бустинг. Композиция из N - 1 модели оценила в 10.5 млн лет, хотя настоящий возраст - примерно 11 млн лет. Если Росс используется MAE в качестве функции потерь, то на какой таргет будет обучаться N-ная модель?

A +1
C 0.5
E -1.5
B -1
D -0.5
F Нет верного ответа.

Вопрос 5. Какой алгоритм не склонен к переобучению при увеличении соответствующего гиперпараметра?

A Решающее дерево (глубина дерева).
B Случайный лес (количество деревьев).
C Градиентный бустинг (количество деревьев).
D KNN (количество соседей).
E Нет верного ответа.

Вопрос 6. Какого гиперпараметра нет у случайного леса?

A Max Depth.
C Learning Rate.
E Min Samples Split.
B N Estimators.
D Max Features.
F Нет верного ответа.

Вопрос 7. Чендлер делает отчеты в экселе и решил изучить кластеризацию. На одном из шагов получилось такое распределение по кластерам. Где будет новый центр кластеров по координате x? Выберите правильную пару чисел.

кластер	1	1	1	2	2	2
x	−1.5	0.5	2	−1	4	6

A 0.5 и 4.
C 3 и 9.
E 4 и 9.
B -1 и 3.
D 0.5 и 9.
F Нет верного ответа.

Вопрос 8. Фиби, Рейчел, Моника и Дженис собираются на девичник. Девушки описали свои предпочтения по двум характеристикам, которые представлены в таблице. Чьи вектора интересов ближе друг к другу? Используйте косинусное расстояние как меру схожести.

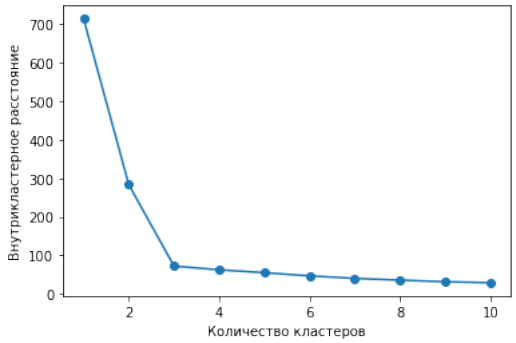
Имя	x_1	x_2
Моника	3	4
Рейчел	5	12
Фиби	6	8
Дженис	8	15

A Моника и Рейчел.
C Рейчел и Дженис.
E Фиби и Дженис.
B Моника и Фиби.
D Рейчел и Фиби.
F Нет верного ответа.

Вопрос 9. Выберите задачу, которую однозначно можно отнести к обучению с учителем.

A Понижение размерности.
C Классификация.
E SVD.
B Кластеризация.
D Рекомендательные системы.
F Нет верного ответа.

Вопрос 10. Какое количество кластеров надо выбрать, судя по графику?



A 2.
C 4.
E 10.
B 3.
D 7.
F Нет верного ответа.

Часть вторая: открытые вопросы

Эта часть состоит из открытых вопросов. На них необходимо дать краткие, но ёмкие ответы. За каждый ответ вы можете получить до 10 баллов.

Вопрос 11. Рассмотрим обучающую выборку для прогнозирования y с помощью x и z:

y_i	x_i	z_i
y_1	1	2
y_2	1	2
y_3	2	2
y_4	2	1
y_5	2	1
y_6	2	1
y_7	2	1

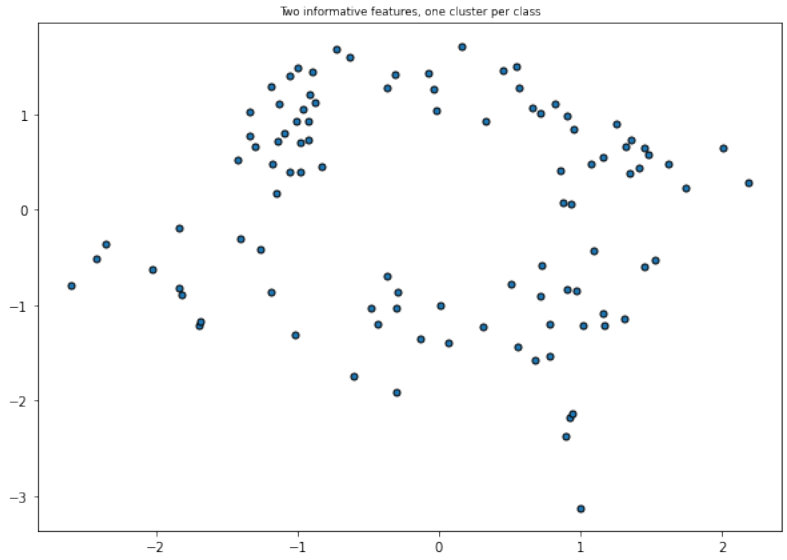
Будем называть деревья разными, если они выдают разные прогнозы на обучающей выборке. Сколько существует разных классификационных деревьев для текущего набора данных? Изобразите их.

Вопрос 12. Опишите пошагово, как обучается случайный лес, в том числе алгоритм построения дерева. Объясните, что такое out of bag ошибка.

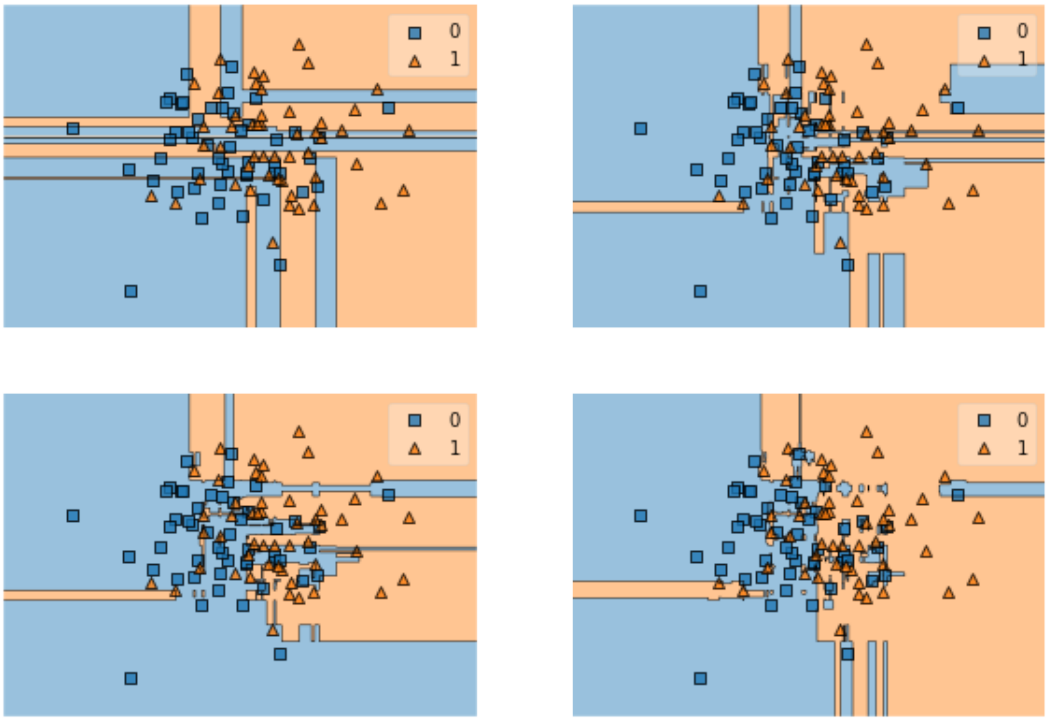
Вопрос 13. Рейчел сходила на свидание с доктором Митчеллом и услышала про загадочную болезнь ALS*. Она ничего не запомнила, поэтому просит помочь и описать, что это такое - ALS. Начинайте с самого начала: как ставится задача для применения этого алгоритма, что нужно сделать, какая функция потерь оптимизируется, какие шаги работы алгоритма.

(* ALS - Amyotrophic lateral sclerosis, все совпадения случайны)

Вопрос 14. Джоуи готовится к роли профессора по компьютерным наукам и учит текст. По сценарию он должен объяснить своим студентам, почему использовать K-means для таких данных - плохая затея, а затем расскажет, какой алгоритм нужно использовать. Объясните, почему k-means не стоит использовать для этих данных. Предложите альтернативный алгоритм и подробно опишите, как он работает.



Вопрос 15. Пока Моника ждет оффера в Twins Garden, она изучает машинное обучение. На картинке представлены 4 случайных леса, обученных Моникой на одних и тех же данных. Варьируется один гиперпараметр. Какой гиперпараметр это может быть? Предположите, какие значения гиперпараметра могут соответствовать каждой картинке. К чему приводит очень большое значение этого гиперпараметра? Предположите, какое значение и картинку стоит выбрать.



Часть третья: задачки

Решите все задания. Все ответы должны быть обоснованы. Решения должны быть прописаны для каждого пункта. Рисунки должны быть чёткими и понятными. Все линии должны быть подписаны. За решение каждой задачи вы можете получить до 10 баллов.

Вопрос 16. После второго скандального развода Росса Моника, Чендлер, Рейчел, Фиби и Джоуи спорят, каковы шансы, что Росс разведется и в третий раз. y - это прогноз каждого из друзей, где "1" - развод, а "0" - долгая и счастливая жизнь. Росс знает, какие признаки используют друзья. Он хочет обучить случайный лес, прогнозирующий по этим данным, разведется ли он в следующем браке или нет. Возможно, он сможет поработать над собой и снизить вероятность третьего развода.

y	f_1	f_2	f_3	f_4
1	5	7	6	9
1	3	1	8	9
0	0	0	0	5
0	1	1	4	7
1	2	4	3	2

Обучите 5 деревьев глубины 1. Используйте Accuracy в качестве критерия разбиения. Для каждого дерева используйте следующие пары признаков для поиска наилучшего разбиения:

Номер	Признак 1	Признак 2
Дерево 1	f_1	f_2
Дерево 2	f_2	f_4
Дерево 3	f_2	f_3
Дерево 4	f_1	f_3
Дерево 5	f_3	f_4

Сделайте прогноз для нового объекта с признаками $(3, 5, 0, 6)$

Вопрос 17. Рейчел прочитала в Vogue, что все модные дома используют технологии машинного обучения. Она решила отсортировать склад одежды в магазине, где работает, на две части, пользуясь методом k-means. Помогите Рейчел рассортировать одежду! Стартовые центры кластеров заданы треугольниками.