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. В листьях деревьев записаны

[A] предикаты
[B] прогнозы
[C] вопросы
[D] условные выражения
[E] критерии информативности
[F] Нет верного ответа.

Вопрос 2. Выберите верное утверждение про измерение энтропии при разбиении вершины

[A] Энтропия считается для распределения признаков в вершине.
[B] Разумно выбирать разбиения, при котором энтропия в дочерних вершинах как можно больше.
[C] Чем меньше энтропия в вершине, тем больше элементов одного класса в вершине.
[D] Энтропия не используется для разбиения вершин.
[E] Высокая энтропия в вершине - критерий останова обучения.
[F] Нет верного ответа.

Вопрос 3. Джо, Чендлер и Росс выписывали недостатки алгоритмов. Какой недостаток они указали для решающих деревьев?

[A] Необходимость масштабирования признаков.
[B] Работа с ограниченным объемом данных.
[C] Работа с ограниченными типами данных.

D Долгое обучение.

E Склонность к переобучению.

F Нет верного ответа.

Вопрос 4. Ошибка модели складывается из

A трех компонент: шума, смещения и разброса.

B двух компонент: смещения и разброса.

C двух компонент: шума и разброса.

D трех компонент: шума, смещения и матожидания.

E трех компонент: матожидания, смещения и разброса.

F Нет верного ответа.

Вопрос 5. Выберите утверждение, которое неверно для случайного леса

A Базовые алгоритмы обучаются на бутстрапированной выборке.

B При разбиении узла используются все признаки.

C При обучении базового алгоритма могут использоваться одни и те же объекты.

D Алгоритмы обучаются независимо друг от друга.

E Базовые алгоритмы представляют собой неглубокие деревья.

F Нет верного ответа.

Вопрос 6. Выберите одно верное утверждение

A Решающие деревья устойчивы к изменениям в выборке.

B Случайный лес имеет меньшее смещение, чем решающее дерево той же глубины.

C Разброс случайного леса выше, чем решающего дерева той же глубины.

D Добавление дерева в лес уменьшает ошибку в N раз.

E Оценить качество случайного леса можно без тестовой выборки.

F Нет верного ответа.

Вопрос 7. Обучение без учителя, в отличие от обучения с учителем

A предполагает наличие целевой переменной.

B предполагает наличие тестовой выборки.

C не предполагает наличие целевой переменной.

D предполагает наличие валидационной выборки.

E не предполагает наличие тестовой выборки.

F Нет верного ответа.

Вопрос 8. Выберите верное утверждение про DBSCAN.

A Метод зависит от расположения центров кластера.

B Метод может не сойтись.

C Методу нужные правильные ответы для обучения.

D Метод сам определяет количество кластеров.

E Метод поделит на K кластеров, K надо задать заранее.

F Нет верного ответа.

Вопрос 9. PCA - это алгоритм для

A регрессии.

B классификации.

C понижения размерности.

D кластеризации.

E рекомендательных систем.

F Нет верного ответа.

Вопрос 10. Джо не делится едой, поэтому Чендлер 5 дней заказывал то же, что и Джо, чтобы понять его вкус. 1 - если еда понравилась, 0 - если нет. Если на Джо смотреть как на рекомендательную систему, чему ему равен p@5 для Чендлера?

Джо	1	1	1	1	1
Чендлер	1	0	1	1	0

$\boxed{A}$ 0.2

$\boxed{B}$ 0.6

$\boxed{C}$ 0.4

$\boxed{D}$ 0.8

$\boxed{E}$ 0.5

$\boxed{F}$ 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.

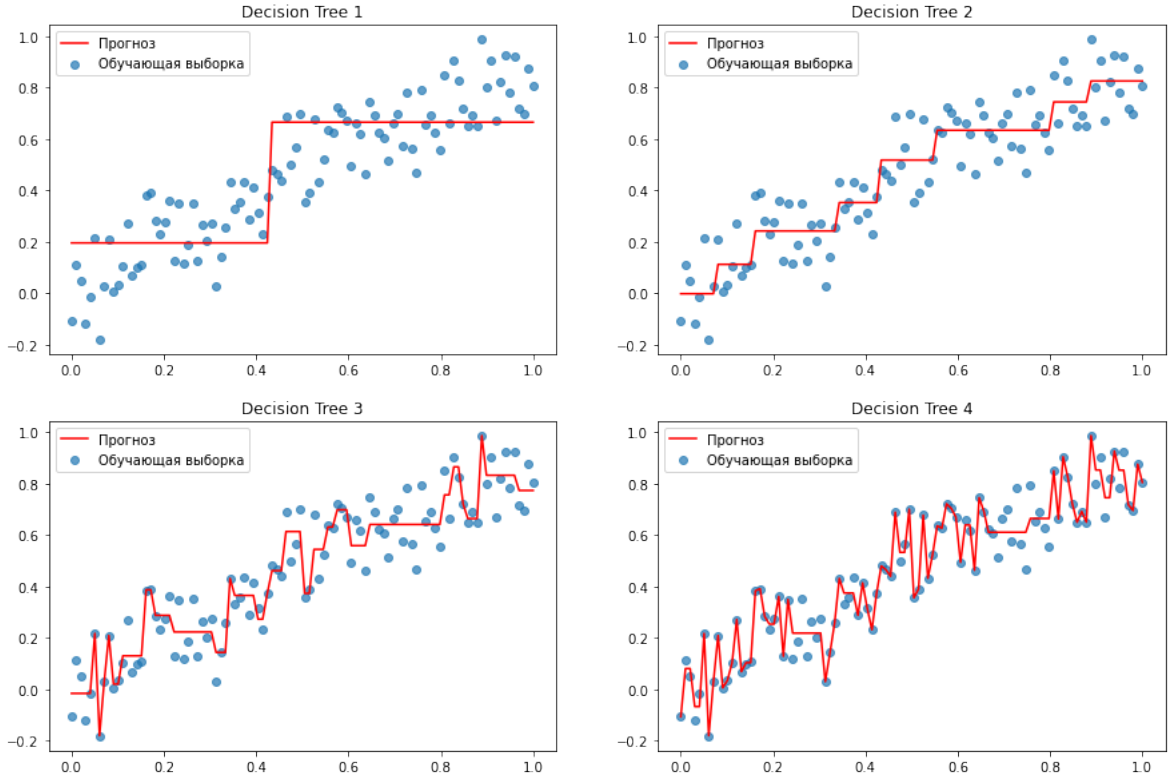
Вопрос 11. Опишите пошагово, как обучается градиентный бустинг для произвольной функции потерь. Фиби планирует написать научно-просветительскую песню, и ей очень нужные доходчивые объяснения.

Вопрос 12. Чендлер задумался о композициях над решающими деревьями. Допустим, думает он, мы обучили две композиции решающих деревьев: случайный лес и градиентный бустинг. В обоих случаях мы остановили добавление деревьев, когда ошибка на валидационной выборке перестала убывать. Что произойдёт с ошибкой этих моделей на валидационной выборке, если мы продолжим обучение и добавим в каждую из них по 10 000 деревьев? Обоснуйте ваш ответ.

Вопрос 13. Рейчел прочитала про алгоритм K-means и решила использовать его для сортировки вещей на складе модного дома, где она работает, но не знает, на сколько кластеров разбить данные. Как подобрать оптимальное число кластеров в алгоритме K-means? Опишите подробно метод.

Вопрос 14. Джо пытается найти новую работу и пользуется сервисом по подбору вакансий, которые используют рекомендательные системы для подборки подходящих вакансий. Как работает коллаборативная фильтрация? Какие виды коллаборативной фильтрации вы знаете? Опишите, как работают известные вам подходы.

Вопрос 15. Моника со скрупулезностью разрабатывает новый рецепт. Чтобы успокоить нервы, она обучает модели машинного обучения в перерывах. На картинке представлены 4 регрессионных дерева, обученных 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. Гантер утомился видеть постоянно бездельничающих друзей на диване и решил себя развлечь, записывая, сколько чашек кофе они заказывают в неделю. y - количество чашек кофе, x_1 - присоединился ли Росс к друзьям или нет, x_2 - погода в градусах по Цельсию.

Гантер построил решающее дерево, прогнозирующее количество заказанных чашек кофе, и хотел было применить его, но Рейчел разлила на листок с деревом кофе. Помогите Гантеру восстановить дерево.

Дерево строится до идеального прогноза. В качестве критерия для разбиения узла Гантер использовал MSE.

y	x_1	x_2
12	0	5
3	1	6
6	1	7
9	1	8
18	0	12
21	0	13

Вопрос 17. Росс задумался, можно ли поделить посетителей кофейни Central Perk на кластеры. В статье он прочитал, что алгоритм DBSCAN получил в этом году премию. Он решил использовать его с параметрами eps=2 и minsamples=1, чтобы выделить кластеры.

Росс пришел к вам за проверкой своих изысканий. Кластеризуйте точки по алгоритму DBSCAN с заданными параметрами.

Сколько кластеров получилось?