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								

> Братан никогда не обижается, если другой Братан не перезвонил, или не ответил на сообщение, своевременно.
>
> *Кодекс Братана: Статья 145*

Это нулевой вариант экзамена. Он нужен для того, чтобы его формат не стал для вас сюрпризом. Работа состоит из трёх частей: тестовая, задачи и ответы на открытые вопросы. Списывание карается обнулением работы. Удачи!

Часть первая: тестовая

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

Вопрос 1. Выберите одно верное утверждение про решающие деревья

- [A] В каждой внутренней вершине дерева проверяется некоторое условие
- [B] В каждой внутренней вершине дерева выдаётся некоторый прогноз
- [C] В каждой листовой вершине дерева проверятеся некоторое условие
- [D] Каждая листова вершина дерева связана как минимум ещё с двумя вершинами
- [E] С помощью решающего дерева можно идеально решить задачи только с линейно разделимой выборкой
- [F] Нет верного ответа.

Вопрос 2. Выберите одно верное утверждение про случайный лес

- [A] В случайном лесу каждое новое дерево исправляет ошибки предыдущих
- [B] Случайный лес переобучается с ростом числа деревьев
- [C] При обучении случайного леса используют **НЕ** глубокие деревья
- [D] Случайный лес непригоден для задачи классификации
- [E] Случайный лес позволяет оценить обобщающую способность без тестовой выборки
- [F] Нет верного ответа.

Вопрос 3. Маршал смотрит хоккей. Все предыдущие сезоны он записывал количество игр, которые выиграли «Викинги». Также он записывал количество пинт пива, которое он выпил в баре. Теперь Маршал хочет обучить решающее дерево предсказывать число победных игр по выпитым пинтам. В качестве критерия разбиения вершины на две он использует MSE. Какое значение порога будет выбрано для разбиения на первом уровне дерева?

пинты	4	5	6	7	8
победы	2	4	3	5	10

[A] 4.5
[B] 5.5
[C] 6.5
[D] 7.5
[E] 8.5
[F] Нет верного ответа.

Вопрос 4. Лили обучает алгоритм машинного обучения предсказывать стоимость картин в долларах. Какой из алгоритмов, перечисленных ниже, может выдать отрицательные прогнозы, несмотря на то, что в обучающей выборке нет отрицательных цен?

[A] Глубокое решающее дерево
[B] Неглубокое решающее дерево
[C] Случайный лес
[D] Градиентный бустинг
[E] Метод ближайших соседей
[F] Нет верного ответа.

Вопрос 5. Стелла обучает градиентный бустинг. В качестве функции потерь она использует MAE. Как будут выглядеть сдвиги s_i?

[A] $y_i - a(x_i)$
[B] $|y_i - a(x_i)|$
[C] $\mathrm{sign}(y_i - a(x_i))$
[D] $\mathrm{sign}(|y_i - a(x_i)|)$
[E] $(y_i - a(x_i))^2$
[F] Нет верного ответа.

Вопрос 6. Робин обучила случайный лес из трёх деревьев для регрессии. Они предсказали $4, -2, 7$. Каким будет итоговое предсказание леса?

[A] 3
[B] 11
[C] 4
[D] 6
[E] -3
[F] Нет верного ответа.

Вопрос 7. Выберите одно верное утверждения про смещение и разброс

[A] У глубоких деревьев высокий разброс и низкое смещение
[B] У деревьев не бывает смещения и разброса
[C] Алгоритмы с маленьким смещением не подвержены переобучению
[D] У глубоких деревьев высокое смещение и низкий разброс
[E] Более сложные модели обычно обладают более высоким смещением
[F] Нет верного ответа.

Вопрос 8. Выберите одно верное утверждения про K-means:

- [A] Метод сам выбирает необходимое число кластеров.
- [B] Метод зависит от выбора начального положения центров кластеров
- [C] Метод гарантированно сходится за 1000 итераций
- [D] K-means и DBSCAN это один и тот же метод
- [E] Метод находит шумовые объекты и выбросы и не учитывает их при кластеризации
- [F] Нет верного ответа.

Вопрос 9. Тэд Мозби, архитектор, хочет выбрать место для своего нового здания. Каждое место в городе описывается 1000 признаков. Тэд хочет сжать пространство этих признаков до двух и изобразить все места на плоскости. Какой алгоритм может ему в этом помочь?

- [A] Градиентный бустинг
- [B] ALS
- [C] Случайный лес
- [D] DBSCAN
- [E] Логистическая регрессия
- [F] Нет верного ответа.

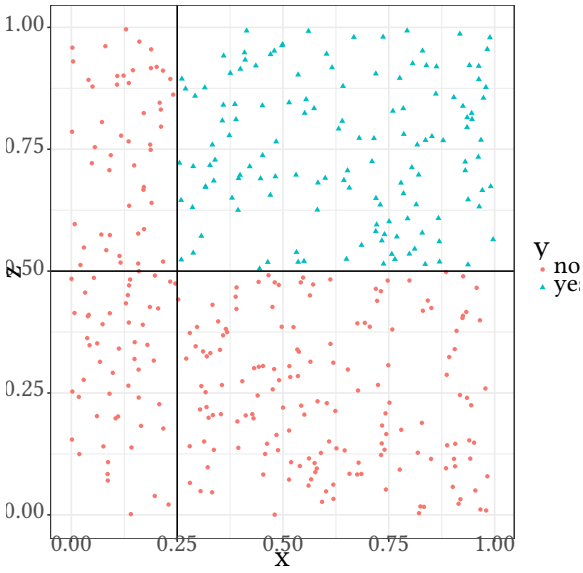
Вопрос 10. Лили рекомендует Тэду заказать себе бургер. Маршал пробовал бургер, ему понравилось. А ещё им обоим до этого понравилась пицца. Как называется такой тип рекомендаций?

- [A] Матричная факторизация
- [B] Item based
- [C] ALS
- [D] Косинусные
- [E] User based
- [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 10 баллов.

Вопрос 11. У Маршала есть диаграмма рассеяния. Постройте по ней классификационное дерево для зависимой переменной y:



Вопрос 12. При обучении случайного леса можно оценивать важность признаков. Опишите здесь один из алгоритмов, с помощью которого это можно сделать.

Вопрос 13. Виктория рассматривает следующий способ обучения базовой модели $b_N(x)$ в градиентном бустинге для функции потерь $L(y, x)$:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(s_i - a_N(x_i)) \to \min_{a_N}; \qquad s_i = \frac{\partial}{\partial z}L(s_i, z)\Big|_{z=b_{N-1}(x_i)}$$

Найдите в формулах все ошибки. Объясните, почему это ошибки и исправьте их.

Вопрос 14. Объясните, как работает метод главных компонент. Для решения каких задач он может использоваться?

Вопрос 15. Братаны собрались на площади и собираются разбиться на группы с помощью алгоритма DBSCAN. В качестве параметра ε братаны используют радиус синей окружности, нанесённой на рисунок. Параметр *min_samples* равен 3. Сам братан, в окрестности которого мы ищем других братанов, тоже входит в эту тройку.

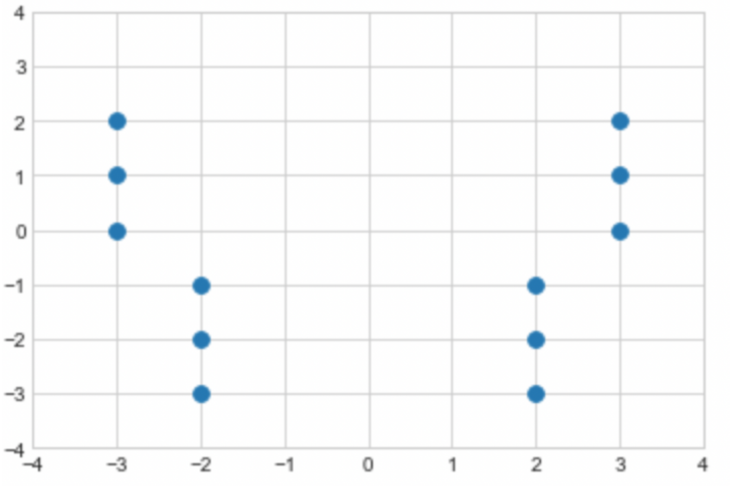
Сколько кластеров и почему выделит DBSCAN? Отметьте их на картинке. Приведите пример объектов, которые окажутся граничными. Отметьте их на рисунке и объясните почему так произойдёт.

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 10 баллов.

Вопрос 16. На картинке ниже синими точками изображены бары. Ребята хотят кластеризовать их с помощью K-means на 4 кластера. Помогите им.

Для поиска расстояний используется манхеттенская метрика. В качестве начальных точек используются $(0, 2)$, $(0, 1)$, $(-1, -3)$, $(-1, -2)$.



Вопрос 17. Барни и Тэд зашли на youtube со своих компьютеров. Они получили одинаковую выдачу из четырёх видосов:

1. History of the empire state building
2. Robin Sparkles-Let's Go To The Mall'
3. How to use playbook correctly
4. Barney Stinson - Nothing Suits Me Like A Suit

Для Барни оказались релевантны второй и третий ролилки. Для Тэда первый и четвёртый. Посчитайте значение метрики $\mathrm{map@4}$. Объясните в чём заключается её смысл.