

Максимальное количество баллов за олимпиаду — 600

Задание 1. Конференция по ИИ

На конференцию по применению ИИ съехались 15 учёных из 15 разных стран, один из них — из России. У каждого иностранного учёного есть совместные статьи ровно с 13-ю гостями конференции. Сколько участников конференции могут являться соавторами нашего соотечественника? Укажите все подходящие варианты. Каждый ответ записывайте в отдельное поле, добавляя их при необходимости.

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов. За ответ $\{2, 4, 6, 8, 10, 12, 14\}$ — 60 баллов

Максимальный балл за задание — 100

Решение.

Задание 2. Ошибающийся чат-бот

Модель ИИ по запросу «Какое наибольшее число прямоугольников из 4 клеток можно вырезать из квадрата 25×25 ?» выдала ответ на запрос «Какое наибольшее число квадратиков из 4 клеток можно вырезать из прямоугольника 25×25 ?». На сколько ответ модели отличается от верного?

Ответ:

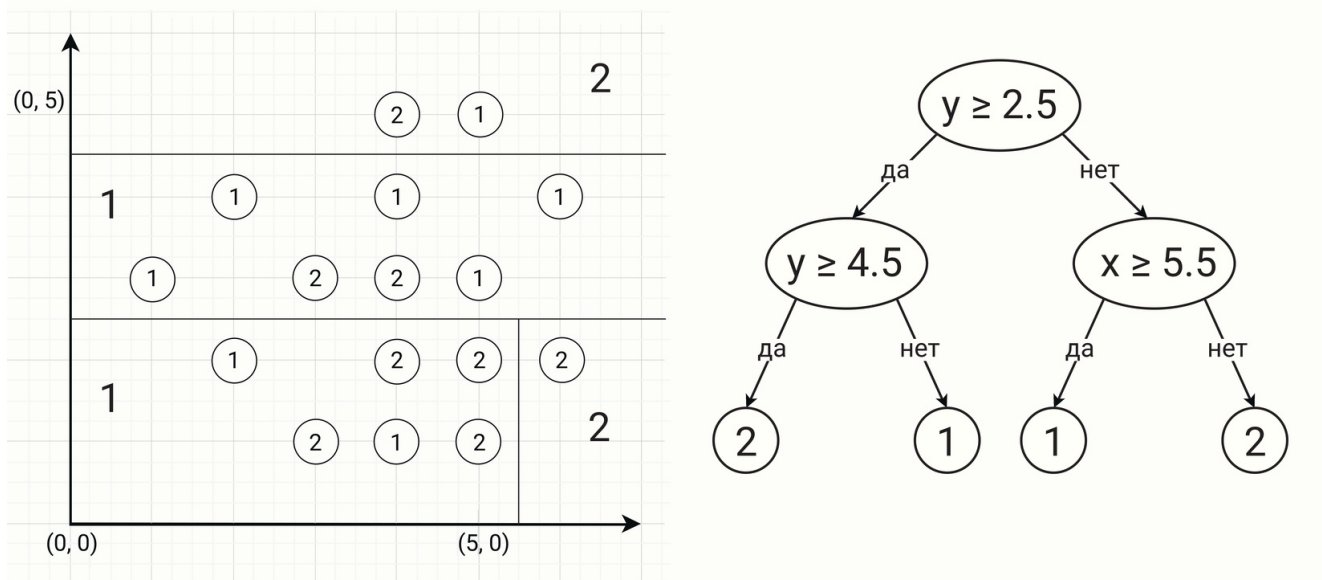
Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

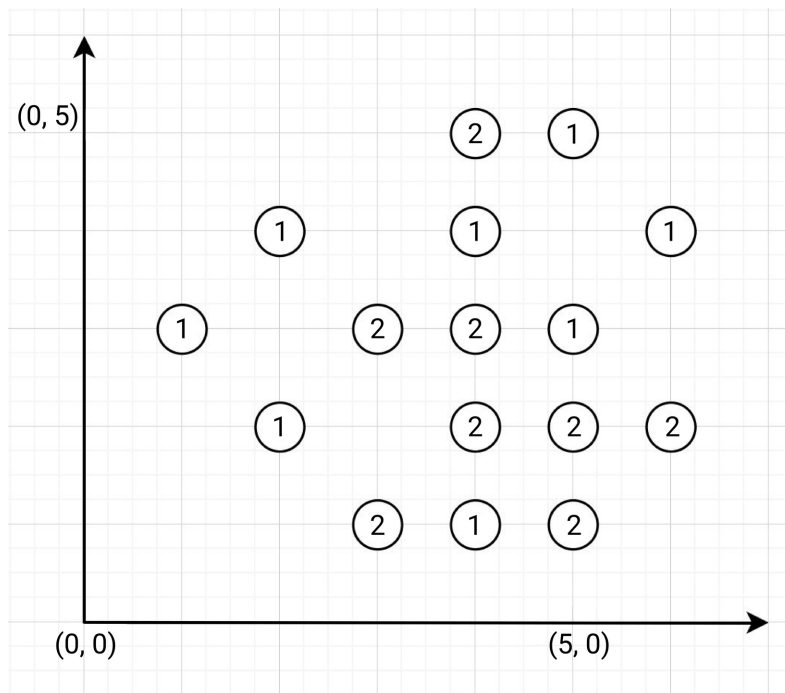
Решение.

Задание 3. Решающее дерево глубины 2

Решающее дерево глубины 2 — это полное бинарное дерево с 7 вершинами: в 4 листьях стоят предсказанные классы, в 3 внутренних вершинах — пороговые условия на одну переменную, например, $y \geq 2.5$ (то есть проверки вида $x \geq t$ или $y \geq t$). Пример такого решающего дерева приведён ниже.



Даны 16 точек на плоскости, каждая принадлежит одному из классов 1, 2.



Заполните пропуски в решающем дереве глубины 2 так, чтобы доля правильно предсказанных ответов оказалась наибольшей.

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение.

Задание 4. Тикеты поддержки

В службе поддержки каждый тикет автоматически помечается вероятностью того, что он критический. Есть разные издержки: если пометим критическим обычный тикет, придётся зря отвлекать команду, а если не заметим критический тикет, понесём серьёзные потери.

Дан файл, который вы можете скачать в форматах [XLSX](#), [ODS](#) или [CSV](#). В каждой строке файла записаны четыре значения:

- `prob` — предсказанная вероятность того, что тикет критический, число из интервала $[0,1]$;
- `label` — истинная метка: 0 (обычный тикет) или 1 (критический);
- `costFP` — штраф за ложное срабатывание;
- `costFN` — штраф за пропуск критического тикета.

Для каждой строки применяется следующее правило классификации:

$$\hat{y} = \begin{cases} 1, & \text{если } p \geq \frac{\text{costFP}}{\text{costFP} + \text{costFN}}, \\ 0, & \text{иначе.} \end{cases} \quad (1)$$

Для скольких строк предсказание совпадёт с истинной меткой, то есть $\hat{y} = \text{label}$?

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Задание 5. Кросс-валидация

Однажды Саша решил обучить классификатор. К сожалению, ресурсов на сложную нейронную сеть у него не было, поэтому он решил, что классификатор будет предсказывать класс, который чаще всего встречался в обучающих данных. Если таких было несколько, он будет предсказывать класс с наименьшим номером. Саша недавно узнал о кросс-валидации и поэтому решил оценить свою модель именно с помощью неё.

Всего у Саши есть n объектов, i -й из которых имеет класс c_i . Для кросс-валидации Саша разделил эти данные на части по t объектов (последняя часть при этом могла оказаться короче). Для каждой из частей классификатор обучается на данных из всех остальных частей и считает число правильных ответов модели по данным из текущей части.

В итоге он суммирует число правильных ответов для всех разбиений. Таким образом, получается некоторая оценка точности модели. Помогите Саше: найдите значение точности модели, посчитанное с помощью кросс-валидации.

Формат входных данных

В первой строке вводятся два целых числа n, t ($1 \leq n \leq 200000, 1 \leq t \leq 1000$).

Во второй строке вводятся n целых чисел c_i — номера классов ($1 \leq c_i \leq 100$).

Формат выходных данных

Выведите одно целое число: суммарное количество правильных ответов по всем разбиениям в кросс-валидации.

Замечание

В первом тестовом примере данные будут разбиты на части $\{1, 1, 2\}, \{2, 2, 3\}, \{3\}$:

- при тестовом блоке $\{1, 1, 2\}$, самым частым среди оставшихся классов будет 2-й (по количеству он совпадает с 3-м, но модель выберет наименьший по номеру), поэтому для этого блока будет 1 правильное предсказание;
- при тестовой части $\{2, 2, 3\}$ самым частым среди оставшихся классов будет 1-й, поэтому для этой части будет 0 правильных предсказаний;
- при тестовой части $\{3\}$ самым частым среди оставшихся классов будет 2-й, поэтому для этой части будет 0 правильных предсказаний.

Примеры

стандартный ввод	стандартный вывод
7 3 1 1 2 2 2 3 3	1

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение

Задание 6. Шаг K-Means

Ограничение по времени: 1 секунда
Ограничение по памяти: 256 мегабайт

Современные сервисы, такие как карты или маркетплейсы, часто используют кластеризацию — разбиение множества объектов на группы по сходству. Например, можно сгруппировать магазины по районам или фотографии по похожим цветам. Один из самых простых и популярных методов кластеризации — k-means.

Этот метод ищет группы точек, расположенных близко друг к другу. Основная идея состоит в том, что каждая группа описывается своим центром (средней точкой), а обучение происходит итерационно: точки относят к ближайшим центрам, а затем центры пересчитывают как среднее значение всех точек, попавших в группу. Эта простая идея лежит в основе многих систем машинного обучения и анализа данных.

В этой задаче вам предстоит реализовать один шаг обновления центров кластеров. Имеется множество точек на плоскости и несколько текущих центров. Необходимо выполнить стандартную операцию алгоритма k-means:

1. Каждая точка относится к ближайшему центру.
2. Координаты каждого центра заменяются на среднее арифметическое координат всех точек, отнесённых к нему.

Если к центру не относится ни одна точка, то его положение не меняется. Если точка одинаково близка к нескольким центрам, она относится к тому, у которого первая координата меньше, а при их равенстве — вторая координата меньше.

Координаты новых центров необходимо вывести в том же порядке, в котором заданы исходные центры.

Формат входных данных

Первая строка содержит два целых числа n и k — количество точек и количество кластеров ($1 \leq k \leq n \leq 10^5$). Далее даны n строк по два целых числа x_i, y_i — координаты точек. Затем следуют k строк по два вещественных числа $c_{x,j}, c_{y,j}$, — координаты текущих центров ($|x_i|, |y_i|, |c_{x,j}|, |c_{y,j}| \leq 10^6$).

Формат выходных данных

Выведите k строк по два вещественных числа с точностью 10^{-4} — координаты новых центров кластеров после пересчёта — в том же порядке, что и во входных данных.

Система оценки

Баллы за каждую подзадачу начисляются только в случае, если все тесты для этой подзадачи успешно пройдены.

Подзадача	Баллы	Доп. ограничения
1	20	$k = 1$
2	80	$1 \leq nk \leq 10^6$

Замечание

Первые три точки ближе к первому центру (0,0) оставшиеся две — ко второму (10,0). После шага алгоритма центры пересчитываются как средние значения точек, относящихся к каждому кластеру: (1.0, 0.3333), (8.5, 2.5).

Замечание

Первые три точки ближе к первому центру (0,0) оставшиеся две — ко второму (10,0). После шага алгоритма центры пересчитываются как средние значения точек, относящихся к каждому кластеру: (1.0, 0.3333), (8.5, 2.5).

Примеры

стандартный ввод	стандартный вывод
5 2	1.0000 0.3333
0 0	8.5000 2.5000
1 1	
2 0	
8 3	
9 2	
0.0 0.0	
10.0 0.0	

Решение