

Максимальное количество баллов за олимпиаду — 600

Задание 1. Расстояние между словами

Будем называть словом любую последовательность букв русского алфавита. За одну операцию разрешается сделать одно из трёх действий:

- убрать из слова одну любую букву,
- добавить в любое место слова любую букву,
- заменить любую букву в слове на любую другую букву.

Расстоянием между словами будем называть наименьшее число операций, которое позволяет получить из одного слова другое. Сколько слов находятся на расстоянии 2 и от слова СТУК, и от слова КУСТЫ?

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение.

Задание 2. След степени матрицы

Матрицей $n \times m$ будем называть таблицу из чисел, состоящую из n строк и m столбцов.

Умножение матриц — это применение правила «строка на столбец». Если

$$M = \begin{pmatrix} p & q \\ r & s \end{pmatrix}, \quad N = \begin{pmatrix} u & v \\ w & x \end{pmatrix},$$

то

$$\begin{pmatrix} pu + qw & pv + qx \\ ru + sw & rv + sx \end{pmatrix}.$$

Обычно порядок важен: как правило, $MN \neq NM$. Возведением матрицы в степень называют многократное умножение её самой на себя: $A^1 = A$, а далее $A^{k+1} = A^k A$.

Дана матрица:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

Найдите сумму левого верхнего и правого нижнего элементов матрицы A^{20} .

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение.

Задание 3. ReLU

Напомним, что

$$\text{ReLU}(z) = \max(0, z).$$

Рассмотрим функцию действительного аргумента x :

$$F(x) = c + w_1 \cdot \text{ReLU}(x - t_1) + w_2 \cdot \text{ReLU}(x - t_2) + \dots + w_{2025} \cdot \text{ReLU}(x - t_{2025}),$$

где $t_1, \dots, t_{2025}$ — попарно различные действительные числа, каждый коэффициент w_i равен либо $+1$, либо -1 , а c — действительное число.

Какое наибольшее число различных решений может иметь уравнение $F(x) = 0$, если известно, что у него конечное число решений?

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение.

Задание 4. Чистка журнала заказов маркетплейса

Дан файл с журналом заказов маркетплейса, который вы можете скачать в форматах XLSX, ODS или CSV.

Каждая строка файла содержит три поля в порядке **delivery_hours** (вещественное число), **label** (строка) и **rating** (вещественное число).

Сначала оставьте только строки, где время доставки попадает в диапазон: $1 \leq \text{delivery_hours} \leq 96$. Затем в столбце **label** замените все пустые значения модой — самым частым значением этого столбца после фильтрации.

Далее нормируйте **rating**: пусть среди оставшихся строк **rating_{min}** — минимальный рейтинг, а **rating_{max}** — максимальный. Для любой строки её нормированный рейтинг вычисляется по формуле

$$\frac{\text{rating} - \text{rating}_{\min}}{\text{rating}_{\max} - \text{rating}_{\min}}.$$

Сколько строк после всех преобразований одновременно удовлетворяют следующим условиям:

- **label** равен моде;
- нормированный рейтинг не меньше 0.8;
- **delivery_hours** ≤ 48 ?

Ответ:

Критерий оценивания: точное совпадение ответа — 100 баллов

Максимальный балл за задание — 100

Решение.

Задание 5. Сжатие размерности

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

В машинном обучении часто возникает проблема: данных слишком много, а ресурсов слишком мало. Например, вы обучаете модель для распознавания изображений, и каждое изображение описывается миллионом признаков (по числу пикселей). Если хранить и обрабатывать все признаки, модель получится очень тяжёлой и будет работать медленно. Чтобы справиться с этим, используют приём сжатия размерности. Его смысл — это удалить или преобразовать те признаки, которые не дают новой информации. Например:

- если у всех объектов один и тот же цвет фона, то этот признак можно убрать;
- если признак не меняется, он не помогает отличать объекты друг от друга.

Рассмотрим упрощённый вариант этой идеи.

Дана выборка объектов a , состоящая из n элементов a_i . Каждый объект a_i представляется двоичной строкой длины 10. Получается таблица признаков: строки — объекты, столбцы — признаки (первый бит — первый признак, второй бит — второй признак и так далее). Все признаки бинарные: каждый бит равен либо 0, либо 1.

Если какой-то признак одинаков у всех чисел, он не несёт информации и должен быть удалён.

Сколько признаков останется после удаления одинаковых?

Формат входных данных

В первой строке входных данных даётся одно число n ($1 \leq n \leq 1000$) — количество объектов в выборке a .
В следующих n строках даётся по одному целому числу a_i ($0 \leq a_i < 1024$) — битовые представления объектов.

Формат выходных данных

Выведите одно целое число — количество оставшихся после удаления признаков.

Примеры

стандартный ввод	стандартный вывод
3 741 121 164	7

Замечание

В первом тестовом примере в массиве a было всего 3 числа: 741, 121, 164. В двоичном виде они выглядят как 1011100101, 0001111001, 0010100100. Удалены будут 2-й, 5-й и 9-й признаки. Останется 7 признаков.

Решение

Задание 6. Разнообразие разбиения

Решающие деревья — один из базовых методов машинного обучения. Они применяются для самых разных задач: от рекомендаций фильмов до диагностики по медицинским признакам. По сути, дерево — это последовательность вопросов о признаках, которые делят выборку объектов на группы. Классический алгоритм строится так, чтобы после каждого вопроса данные в ветвях становились как можно более однородными по классам — ведь чем однороднее ветка, тем проще в ней принять окончательное решение.

Однако у такого подхода есть недостаток: дерево может переобучаться, слишком подстраиваясь под конкретные данные. Чтобы изучить альтернативные принципы разбиения, исследователи решили рассмотреть противоположную идею — иногда выбирать разбиение, при котором обе ветви оказываются максимально разнообразными по составу классов. В этой задаче вам предстоит реализовать модуль, находящий порог признака, который даёт наибольшее разнообразие.

Дана обучающая выборка из n объектов. Каждый объект описывается одним числовым признаком x_i и меткой класса $y_i \in \{0, 1\}$. Рассмотрим целочисленный порог t и определим разбиение выборки на два блока:

- левый блок $L(t)$ — все объекты с номерами i , для которых $x_i \leq t$;
- правый блок $R(t)$ — все объекты с номерами i , для которых $x_i > t$.

Обозначим:

- $n_l = |L(t)|$ и $n_r = |R(t)|$ — количество объектов в левом и правом блоках;
- p_l — доля объектов класса 1 в левом блоке;
- p_r — доля объектов класса 1 в правом блоке;

Мера разнообразия одного блока определяется как

$$d = p(1 - p),$$

где p — доля объектов класса 1 в этом блоке. Общая оценка разбиения порогом t — это взвешенная по размерам блоков сумма мер разнообразия левого и правого блоков:

$$d_{\text{split}}(t) = \frac{n_l}{n} p_l(1 - p_l) + \frac{n_r}{n} p_r(1 - p_r),$$

где $n = n_l + n_r$.

Найдите такое целое значение порога t , при котором величина $d_{\text{split}}(t)$ максимальна. Если несколько порогов дают одинаковое значение разнообразия, выберите наименьший из них.

Формат входных данных

В первой строке входных данных дано единственное целое число n — количество объектов ($2 \leq n \leq 10^4$). В следующих n строках даны по два целых числа x_i и y_i — признак и класс i -го объекта ($|x_i| < 10^{12}$, $y_i \in \{0, 1\}$). Обратите внимание, что в данных могут встречаться повторяющиеся признаки и не гарантируется, что y_i будут совпадать при совпадающих x_i .

Формат выходных данных

Выведите одно целое число — оптимальное значение порога t , при котором величина $d_{\text{split}}(t)$ максимальна. Если несколько значений t дают одинаковое значение разнообразия, выведите наименьший из таких порогов.

Система оценки

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

Подзадача	Баллы	Дополнительные ограничения
1	20	Все метки классов одинаковы ($y_i = 0$ или $y_i = 1$)
2	80	$2 \leq n \leq 10^4$

Примеры

стандартный ввод	стандартный вывод
5 1 0 2 1 3 0 5 0	2

Замечание

Рассмотрим все возможные пороги $t \in 1, 2, 3, 4$. Для каждого порога вычислим меру разнообразия

$$d_{\text{split}} = \frac{n_l}{n} p_l(1 - p_l) + \frac{n_r}{n} p_r(1 - p_r),$$

где p_l и p_r — доли единиц в левой и правой группах соответственно.

t	Левый блок ($x_i \leqslant t$)	Правый блок ($x_i > t$)	p_l	p_r	d_{split}
1	[0]	[1, 0, 1, 0]	0	0.5	0.2
2	[0, 1]	[0, 1, 0]	0.5	$\frac{1}{3}$	0.2(3)
3	[0, 1, 0]	[1, 0]	$\frac{1}{3}$	0.5	0.2(3)
4	[0, 1, 0, 1]	[0]	0.5	0	0.2

Максимальное значение $d_{\text{split}} = 0.2333$ достигается при $t = 2$ и $t = 3$, но по условию при равенстве выбирается наименьший порог: $t = 2$.

Решение