

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?

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} .

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?

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 ?

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.

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 ($x_i \leq t$)	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: $t = 2$.

Максимальное количество баллов за олимпиаду — 600

Задание 1. Картинки в папке

В папке на компьютере размещено 3000 изображений трёх типов: с кошкой, с собакой, с кошкой и собакой одновременно. Изображений всех типов одинаковое количество. Модель ИИ ответила на два вопроса к каждой картинке: есть ли на ней кошка и есть ли на ней собака. На первый вопрос было получено 2790 ответов «да», а на второй — 2861. Удалили все изображения, для которых хотя бы один ответ ИИ был неверным. Какое наименьшее число изображений могло остаться в папке?

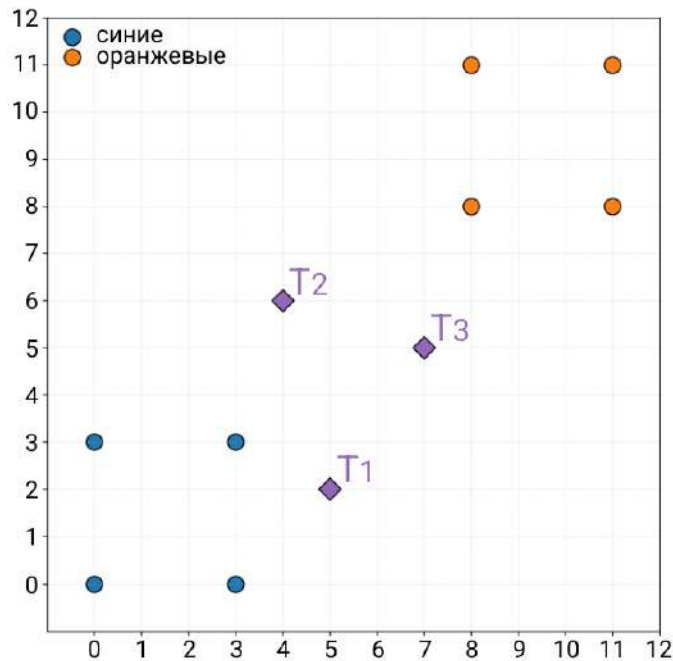
Задание 2. Сколько расстановок дают положительный определитель

Матрицей $n \times m$ будем называть таблицу из чисел, состоящую из n строк и m столбцов. Определитель матрицы 2×2 — это число, вычисляемое по формуле. Для $\begin{pmatrix} p & q \\ r & s \end{pmatrix}$ его значение равно $ps - qr$.

Сколько существует способов расставить числа 1, 2, 3, 6 по клеткам матрицы 2×2 (каждое число используется ровно один раз) так, чтобы $ps - qr > 0$?

Задание 3. Свой среди чужих

На плоскости даны восемь обучающих точек — четыре синие и четыре оранжевых, а также три тестовые точки T_1 , T_2 , T_3 , обозначенные фиолетовыми ромбами.



Цвет тестовой точки будем предсказывать по цветам её соседей.

Иногда разные координаты имеют очень разный масштаб: одна может меняться от 0 до 10000, а другая — от 0 до 0.00001. Тогда при вычислении расстояний первая координата начинает полностью «перебивать» вторую. Чтобы обе координаты влияли честно, первую иногда растягивают или сжимают.

В этой задаче первая координата умножается на положительное число $\lambda > 0$:

$(x, y) \mapsto (\lambda x, y)$ Расстояние между точками считаем евклидовым в новых координатах $(\lambda x, y)$. Цвет тестовой точки предсказываем так:

- находим 3 ближайших к ней обучающих точки;
- если среди них больше синих, считаем тестовую точку синей, если больше оранжевых — оранжевой;
- если расстояния совпадают, то более близкой считается точка с меньшим λx , а при равных λx — с меньшим y .

Зафиксируем $\lambda = 1$ и посмотрим, какого цвета при этом каждая тестовая точка T_i .

Назовём значение $\lambda > 0$ *интересным* для точки T_i , если при таком λ предсказанный цвет T_i отличается от её цвета при $\lambda = 1$.

Рассмотрим для каждой точки все её интересные значения λ .

Оказывается, что для каждой T_i возможны только два случая:

- интересных значений нет вообще — тогда цвет точки не меняется ни при каком $\lambda > 0$;
- интересные значения существуют и устроены так: есть число $L_i > 0$, что при всех $0 < \lambda < L_i$ цвет точки T_i другой, чем при $\lambda = 1$, а при всех $\lambda > L_i$ цвет уже такой же, как при $\lambda = 1$. При $\lambda = L_i$ цвет может совпадать или не совпадать с цветом при $\lambda = 1$ — это надо аккуратно проверить.

Число L_i будем называть *границей* интересных значений для точки T_i . Для каждой тестовой точки T_i :

- выведите -1 , если интересных значений λ нет;
- найдите её границу L_i и выведите число $(L_i)^2$ в ином случае.

Задание 4. Мониторинг энергопотребления магазина

В течение нескольких дней подряд в магазине измеряли энергопотребление каждые 15 минут. Таким образом, за одни сутки получается 96 измерений. Результаты вы можете скачать в форматах [XLSX](#), [ODS](#) или [CSV](#).

В таблице содержится непрерывный ряд измерений без пропусков; в каждой строке указаны t (номер измерения, начиная с 0) и **power** (потребление в ваттах). Обозначим $x_t = \text{power}[t]$.

Чтобы проанализировать динамику потребления, рассмотрим несколько производных величин.

Во-первых, будем отслеживать изменение мощности от шага к шагу. Для каждого $t \geq 1$ определим разность: $d_t = x_t - x_{t-1}$. Во-вторых, нас интересуют сглаженные характеристики. Скользящие средние вычисляются по «хвостовым» окнам, то есть берут текущий момент и несколько предыдущих:

- среднее потребление за последние 8 шагов: $\text{ma_8}[t] = \frac{1}{8} \sum_{j=0}^7 x_{t-j} \quad (t \geq 7)$,
- среднее изменение потребления за последние 4 шага: $\text{ma_diff_4}[t] = \frac{1}{4} \sum_{j=0}^3 d_{t-j} \quad (t \geq 4)$.

Кроме того, по всему ряду значений x_t необходимо найти медиану **med**. Медиана — это число, которое стоит посередине после сортировки всех значений (или среднее двух центральных при чётном числе точек).

Поскольку измерения выполнялись много дней подряд, выражение $t \bmod 96$ указывает, на каком 15-минутном интервале суток находится момент t . Значения от 36 до 84 соответствуют промежутку с 09:00 до 21:00 включительно.

Сколько моментов времени t одновременно удовлетворяют трём условиям:

$\text{ma_diff_4}[t] \geq 3.0$, $\text{ma_8}[t] \geq \text{med}$, $36 \leq (t \bmod 96) \leq 84$?

Задание 5. Оптимальная конфигурация

Ограничение по времени: 1 секунда

Ограничение по памяти: 256 мегабайт

Обучение больших нейронных сетей происходит на кластере — вычислительной инфраструктуре, которая позволяет использовать одновременно большое количество видеокарт. Видеокарты на кластере объединяются в ноды (блоки по 8 карт).

Для обучения больших языковых моделей методом обучения с подкреплением видеокарты делятся на два типа: одни генерируют данные, другие обновляют веса модели на основе генерированных данных. Вторая задача гораздо сложнее, поэтому при наилучшем распределении ресурсов на неё должно выделяться в k раз больше видеокарт. Обучение с оптимальным распределением ресурсов мы будем называть оптимальным.

Вы работаете в очень большой компании, поэтому ваши ресурсы не ограничены. Вы хотите выделить минимальное количество нод так, чтобы их видеокарты можно было разбить на два вышеописанных типа так, чтобы обучение при этом было оптимальным.

Формат входных данных

В первой и единственной строке входных данных вводится одно целое число k ($1 \leq k \leq 1000000$).

Формат выходных данных

Выведите одно число: минимальное число нод, которые нужно выделить так, чтобы обучение было оптимальным.

Примеры

стандартный ввод	стандартный вывод
5	3

Задание 6. Распознавание объектов

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

Одна из задач машинного обучения — распознавание объектов. Она заключается в том, чтобы выделить на изображении зоны, в которых находятся объекты, и сказать, что это за объекты.

Гоша запустил умную нейронную сеть для распознавания объектов. Она вернула таблицу $n \times m$, где в каждой клетке указан номер класса, к которому принадлежит находящийся в клетке объект по мнению модели, или 0, если модель считает, что в этой клетке нет никаких интересных её объектов.

К сожалению, этого недостаточно: Гоша хочет выделить в таблице прямоугольники-рамки со сторонами, параллельными её краям, так, чтобы внутри этих рамок были распознанные объекты. Объектами Гоша считает все клетки, которые модель распознала как клетки одного конкретного класса.

Помогите Гоше: выделите в таблице все распознанные умной нейронной сетью объекты. Так как объектов в таблице может быть очень много, посчитайте и выведите всего одно число: сумму площадей всех прямоугольников-рамок объектов.

Формат входных данных

В первой строке вводятся два целых числа n, m ($1 \leq n, m \leq 1000$).
В следующих n строках вводится по m чисел a_{ij} — номера классов объектов, которые модель нашла в конкретной клетке, или 0, если в этой клетке не был распознан ни один объект.

Формат выходных данных

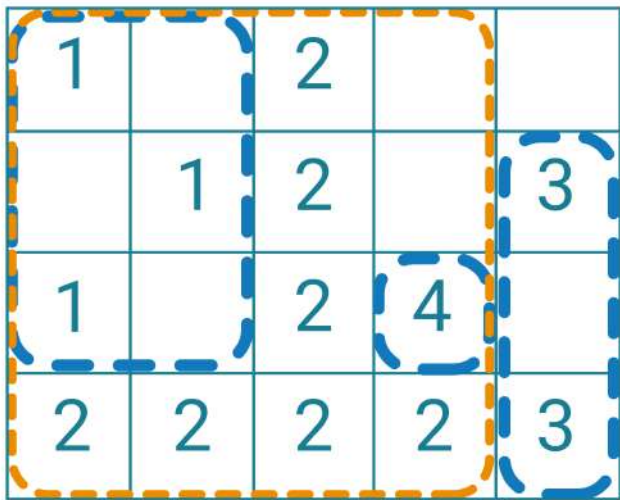
Выведите одно целое число: сумму площадей прямоугольников-рамок для всех распознанных объектов.

Примеры

стандартный ввод	стандартный вывод
1 0 2 0 0 0 1 2 0 3 1 0 2 4 0 2 2 2 2 3	26

Замечание

Рассмотрим первый пример из условия.



Максимальное количество баллов за олимпиаду — 600

Задание 1. Математика в чат-боте

Дима выбрал два натуральных числа a и b . Затем он отправил модели ИИ запрос — вычислить a^b . Он записал это выражение на бумажке, сфотографировал и загрузил фотографию. Из-за неаккуратного почерка модель распознала выражение как $a \cdot b$ и посчитала именно его. В итоге её ответ оказался меньше правильного на 110.

Какой ответ выдала модель?

Задание 2. Максимальный след по всем перестановкам

Матрицей $n \times m$ будем называть таблицу из чисел, состоящую из n строк и m столбцов. Умножение матриц выполняют по правилу «строка на столбец». Если

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

то

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

Суммой диагональных элементов (следом) матрицы называют число $tr \begin{pmatrix} p & q \\ r & s \end{pmatrix} = p + s$.

Дана матрица

$$A = \begin{pmatrix} -1 & 4 \\ 5 & 10 \end{pmatrix}.$$

Также даны числа $-4, -5, 20, 25$. Рассматриваются все 24 матрицы B вида

$$B = \begin{pmatrix} x & y \\ z & w \end{pmatrix},$$

в которых x, y, z, w — некоторая перестановка чисел $-4, -5, 20, 25$. Для каждой такой B рассмотрите произведения AB и BA .

а) Найдите наибольшее возможное значение $tr(AB)$.

Задание 3. Минимизация L_1 и L_2

Вы настраиваете простейшую регрессионную модель, которая всегда предсказывает одно и то же число c (константная модель). Дан набор истинных значений:

$$y = \{1, 2, 3, 9, 10, 10\}.$$

Рассмотрим две функции качества: $L_1(c) = \sum_i |y_i - c|$ и $L_2(c) = \sum_i (y_i - c)^2$.

а) Найдите значение c_1 , минимизирующее $L_1(c)$. Если оптимальных значений несколько, в ответ запишите наименьшее.

Ответ: 3

Критерий оценивания: точное совпадение ответа — 50 баллов

б) Найдите значение c_2 , минимизирующее $L_2(c)$. Если оптимальных значений несколько, в ответ запишите наименьшее.

Задание 4. Group By

Система электронного тестирования фиксирует результаты проверочных работ школьников, которые вы можете скачать в форматах XLSX, ODS или CSV. Файл содержит пять столбцов:

- **student_id** — идентификатор ученика (целое число);
- **subject** — предмет, по которому выполнен тест (строка);
- **score** — полученный балл (вещественное число);
- **cheat_flag** — подозрение на списывание (True/False);
- **attempt_no** — номер попытки (целое число, 1 означает первую попытку).

Выполните следующие действия с данными:

1. Очистите столбец **score**: пустые значения замените на 0, значения меньше 0 также замените на 0, значения больше 100 замените на 100.
2. Удалите строки, где **cheat_flag** = True.
3. Оставьте только строки, соответствующие первой попытке, то есть те, где **attempt_no** = 1.
4. Для каждого ученика вычислите его средний балл по предметам — это среднее всех значений **score**, которые остались после действий выше.

Сколько учеников имеют средний балл в диапазоне $60 \leq \text{avg_score} < 80$?

Задание 5. Дерево принятия решений

Ограничение по времени: 1 секунда

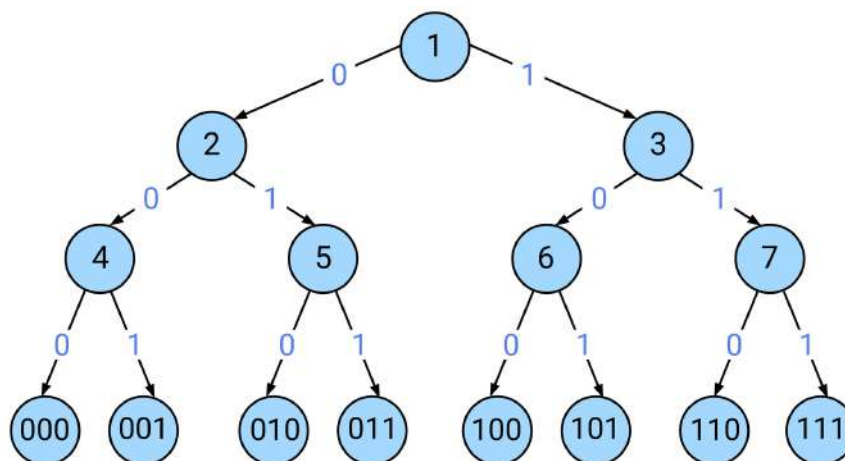
Ограничение по памяти: 256 мегабайт

В машинном обучении часто используют деревья решений. Каждая внутренняя вершина такого дерева соответствует некоторому вопросу, а каждое ребро — выбору ответа (да/нет). Таким образом, за несколько вопросов исходные данные могут быть разбиты на достаточно большое количество классов. Очередным проектом для Димы стало дерево принятия решений. Его структура похожа на структуру решающего дерева. Это полное бинарное дерево глубины n . В каждой вершине, кроме вершин последнего уровня, хранится число p ($0 \leq p \leq 100$). Это число обозначает вероятность выбора: с вероятностью p процентов алгоритм выберет пойти влево и, соответственно, с вероятностью $100 - p$ процентов — вправо.

Договоримся: движение влево обозначим цифрой 0, а движение вправо — цифрой 1. Таким образом, каждая вершина нижнего уровня соответствует двоичной строке длины n (последовательности решений от корня до листа).

Вероятность получения этой строки равна произведению вероятностей всех выборов, сделанных на пути от корня до листа.

Дима уже написал структуру для такого дерева и хочет протестировать её. Для этого он создал дерево глубины 3. Значит, всего в нём 7 внутренних вершин. Каждая вершина имеет своё число p . Ниже показана схема расположения этих вершин:



Найдите вероятности всех двоичных строк длины 3 и выведите их в порядке неубывания вероятности. Если вероятности совпадают, строки должны выводиться в лексикографическом порядке.

Формат входных данных

В первой строке заданы 7 целых чисел p ($0 \leq p \leq 100$) — вероятности для вершин, как показано на рисунке.

Формат выходных данных

Выведите 8 строк. Каждая строка должна содержать двоичную строку длины 3. Строки должны идти в порядке неубывания вероятности. При равенстве вероятностей строки сравниваются лексикографически.

Примеры

стандартный ввод	стандартный вывод
40 90 20 90 100 70 0	011 110 001 101 010 100 000 111

Замечание

В первом тестовом примере двоичные строки имеют следующие вероятности:

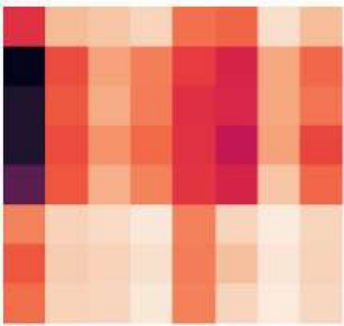
- 011 — 0.0
- 110 — 0.0
- 001 — 0.036
- 101 — 0.036
- 010 — 0.04
- 100 — 0.084
- 000 — 0.324
- 111 — 0.48

Задание 6. Тепловая карта

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

Слава готовит постер на конференцию. К сожалению, сейчас его тепловая карта не влезает на постер: она слишком большая. Поэтому Слава решил выделить на текущей карте некоторый прямоугольный фрагмент и использовать его для презентации. Славина тепловая карта выглядит как таблица из n строк и m столбцов. Клетка на пересечении i -й строки и j -го столбца имеет цвет c_{ij} . Используются k цветов, которые пронумерованы от 1 до k . Пример аналогичной карты приведён справа. Слава хочет продемонстрировать весь размах значений, поэтому на выбранном фрагменте должна быть хотя бы одна клетка каждого цвета. При этом юный докладчик хочет минимизировать площадь карты, ведь ему нужно уместить её на постер.

Помогите ему: найдите прямоугольник, который можно будет вырезать из его карты так, чтобы на нём были клетки всех k цветов. Гарантируется, что на исходной тепловой карте присутствуют клетки всех k цветов.



Формат входных данных

В первой строке вводятся три натуральных числа n, m, k ($1 \leq n, m \leq 250, 1 \leq k \leq 20$).
В следующих n строках задаётся по m натуральных чисел c_{ij} ($1 \leq c_{ij} \leq k$).

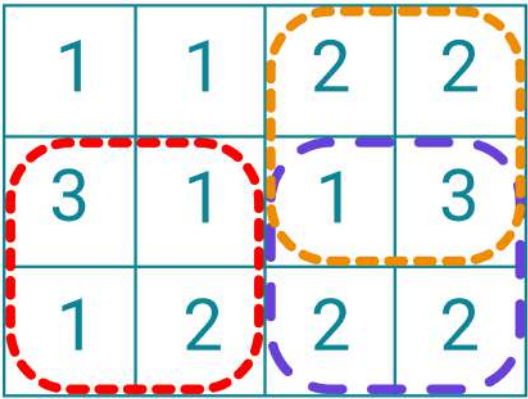
Формат выходных данных

Выведите 4 числа x_1, y_1, x_2, y_2 , задающие прямоугольник, который нужно вырезать. Прямоугольник задаётся своими верхней и нижней строками (x_1, x_2) и левым и правым столбцами (y_1, y_2) . Если есть несколько способов вырезать график наименьшей площади, выведите любой.

Примеры

стандартный ввод	стандартный вывод
3 4 3 1 1 2 2 3 1 1 3 1 2 2 2	1 3 2 4

Замечание



Все возможные варианты вырезать график наименьшей площади для первого примера.