A. Последовательность

1.5 секунд, 256 мегабайт

Дана последовательность $y(t)$, полученная по формуле:

$$y(t) = a_0 + \sum_{i=1}^{4} a_i \cdot \sin\left(\frac{2\pi t}{m_i} + b_i\right),$$

где $m = (12, 24, 168, 672)$. Известны округлённые к ближайшему целому значения $y(t)$ для начала ряда. Требуется предсказать его продолжение.

Входные данные

Вам даны 168 строк с округлёнными значениями $y(t)$ для $t = 1, \dots, 168$. t-я строка содержит одно целое число $[y(t)]$. Все числа по модулю не превышают 10^4.

Выходные данные

Выведите 168 строк. t-я строка должна содержать значение $y(t + 168)$.

Система оценки

Целевая функция ошибки RMSE. Пусть $Score = 50 \cdot \frac{N-S}{N-J}$, где S — RMSE вашего решения, а N — наивного решения, а J — эталонного решения с 5% запасом.

Тогда $\text{Verdict} = \begin{cases} \text{Ok} & \text{Score} = 50 \\ \text{PartiallyCorrect} & 0 \leq \text{Score} < 50 \\ \text{WrongAnswer} & \text{Score} < 0 \end{cases}$

Для локального тестирования вы можете использовать следующий набор тестов: https://disk.yandex.ru/d/poSCFyDzIHKynw

Каждый тест устроен следующем образом:

- Первые 168 строк — входные данные.
- Следующие 168 строк — реальное продолжение последовательности.
- Последние две строки — ошибка эталонного решения с 5% запасом и наивного решения.

B. Наивный байесовский классификатор

1 секунда, 256 мегабайт

Реализуйте наивный байесовский классификатор.

Априорные вероятности классов оцениваются обыкновенным частотным методом.

Для оценки вероятности встречи слов в каждом классе используется модель Бернулли с аддитивным сглаживанием (сглаживание Лапласа) $p(x) = \frac{count(x)+\alpha}{\sum_{y \in Q} count(y)+\alpha \cdot |Q|}$, где x — рассматриваемое событие, а Q — множество всех событий.

Каждое слово — это отдельный категориальный признак с двумя возможными событиями встретилось / не встретилось.

Входные данные

В первой строке содержится целое положительное число K ($1 \leq K \leq 10$) — число классов.

Во второй строке содержится K целых положительных чисел λ_C ($1 \leq \lambda_C \leq 10$) — штрафы за ошибки классификации сообщений соответствующих классов.

В третьей строке содержится целое положительное число α ($1 \leq \alpha \leq 10$) — интенсивность аддитивного сглаживания.

Следующая строка содержит целое положительное число N ($1 \leq N \leq 200$) — число сообщений в обучающей выборке.

Следующие N строк содержат описания соответствующих сообщений из обучающей выборки. Каждое сообщение в ней начинается с целого положительного числа C_i ($1 \leq C_i \leq K$) — класса, к которому относится i-е сообщение. Далее следует целое положительное число L_i ($1 \leq L_i \leq 10^4$) — число слов в i-м сообщении. Затем следует содержание сообщения — L_i слов состоящих из маленьких латинских букв.

Далее в отдельной строке содержится целое положительное число M ($1 \leq M \leq 200$) — число сообщений в проверочной выборке.

Следующие M строк содержат описания соответствующих сообщений из проверочной выборки. Каждое сообщение в ней начинается с целого положительного числа L_j ($1 \leq L_j \leq 10^4$) — число слов в j-м сообщении. Затем следует содержание сообщения — L_j слов состоящих из маленьких латинских букв.

Гарантируется, что сумма длин всех сообщений в обучающей и проверочной выборках меньше чем $2 \cdot 10^6$.

Выходные данные

Выведите M строк — результаты мягкой классификации оптимального наивного байесовского классификатора соответствующих сообщений из проверочной выборки. Допустимая абсолютная и относительная погрешность 10^{-4}.

Каждый j-й результат мягкой классификации должен содержать K чисел p_C — вероятности того, что j-е сообщение относится к классу C.

входные данные

```
3
1 1 1
1
4
1 2 ant emu
2 3 dog fish dog
3 3 bird emu ant
1 3 ant dog bird
5
2 emu emu
5 emu dog fish dog fish
5 fish emu ant cat cat
2 emu cat
1 cat
```

выходные данные

```
0.4869739479 0.1710086840 0.3420173681
0.1741935484 0.7340501792 0.0917562724
0.4869739479 0.1710086840 0.3420173681
0.4869739479 0.1710086840 0.3420173681
0.4869739479 0.3420173681 0.1710086840
```

В примере условные вероятности выглядят следующим образом:

$p(w_x\|c_y)$	ant	bird	dog	emu	fish
c_1	$3/4$	$1/2$	$1/2$	$1/2$	$1/4$
c_2	$1/3$	$1/3$	$2/3$	$1/3$	$2/3$
c_3	$2/3$	$2/3$	$1/3$	$2/3$	$1/3$

Слово cat не рассматривается, так как оно ни разу не встретилось в обучающей выборке.

Для первого запроса X:
$p(c_1) \cdot p(X|c_1) = \frac{2}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \left(1 - \frac{1}{2}\right) \cdot \left(1 - \frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) \cdot \left(1 - \frac{1}{4}\right)$
и $p(c_1|X) = \frac{3/256}{3/256+1/243+2/243}$

C. Категориальная корреляция

1 секунда, 256 мегабайт

Вычислите коэффициент корреляции Пирсона между категориальным и числовым признаком. Так как первый признак категориальный сперва требуется применить one-hot преобразование к нему, а затем вычислить среднее взвешенное значение корреляций между новыми признаками и b.

Входные данные

Первая строка содержит два натуральных числа N и K, разделённых пробелами: N $(1 \leq N \leq 10^5)$ — число объектов, K ($1 \leq K \leq 10^5$) — число значений категории первого признака. Вторая строка содержит N натуральных чисел, разделённых пробелами: i-е из них a_i $(1 \leq a_i \leq K)$ — значение первого признака i-го объекта. Третья строка содержит N целых чисел, разделённых пробелами: i-е из них b_i $(|b_i| \leq 10^9)$ — значение второго признака i -го объекта.

Выходные данные

Выведите одно вещественное число с плавающей точкой — коэффициент корреляции Пирсона между a и b. Абсолютная или относительная погрешность ответа не должна превышать 10^{-9}

входные данные

```
6 3
1 2 2 3 3 3
1 2 3 4 5 6
```

выходные данные

```
0.19203297584037293
```

В примере значение корреляции между первым новым признаком $(1, 0, 0, 0, 0, 0)$ и b равно -0.654653671, а его вес равен единице, так как соответствующие значение встретилось только один раз. Значение корреляции между вторым новым признаком $(0, 1, 1, 0, 0, 0)$ и b равно -0.414039336, а его вес равен двум. Значение корреляции между третьим новым признаком $(0, 0, 0, 1, 1, 1)$ и b равно 0.878310066, а его вес равен трём.

D. Условная дисперсия

1 секунда, 256 мегабайт

Вычислите критерий связи двух признаков категориального X и числового Y на основе математического ожидания условной дисперсии $D(Y|X)$. Вероятности для X оцениваются обыкновенным частотным методом.

Входные данные

Первая строка содержит одно целое положительное число K ($1 \leq K \leq 10^5$) — максимальное число различных значений признака X.

Следующая строка содержит целое положительное число N ($1 \leq N \leq 10^5$) — число объектов.

Следующие N строк содержат описания соответствующих объектов. Каждая из этих N строк содержит описание одного объекта: два целых числа x и y $(1 \leq x \leq K, |y| \leq 10^9)$ — значения признаков X и Y.

Выходные данные

Выведите одно вещественное число с плавающей точкой — математическое ожидание условной дисперсии. Допустимая абсолютная и относительная погрешность 10^{-6}.

входные данные

```
2
4
1 1
2 2
2 3
1 4
```

выходные данные

```
1.25
```

E. Расстояния

1 секунда, 256 мегабайт

Посчитайте зависимость категориального признака Y от числового X по внутриклассовому и межклассовому расстоянию:

- Внутриклассовое расстояние $= \sum_{i,j:y_i=y_j} |x_i - x_j|$
- Межклассовое расстояние $= \sum_{i,j:y_i \neq y_j} |x_i - x_j|$

Входные данные

Первая строка содержит одно целое положительное число K ($1 \leq K \leq 10^5$) — максимальное число различных значений Y второго признака.

Следующая строка содержит одно целое положительное число N ($1 \leq N \leq 10^5$) — число объектов.

Следующие N строк содержат описания соответствующих объектов. Каждая из этих N строк содержит описание одного объекта: два целых числа x и y ($|x| \leq 10^7$,$1 \leq y \leq K$) — значения первого и второго признака описываемого объекта.

Выходные данные

В первой строке выведите одно целое число — внутриклассовое расстояние.

Во второй строке выведите одно целое число — межклассовое расстояние.

входные данные

```
2
4
1 1
2 2
3 2
4 1
```

выходные данные

```
8
12
```

F. F-мера

1 секунда, 256 мегабайт

В результате эксперимента по классификации на K классов была получена матрица неточностей (Confusion matrix) CM, где $CM[c,t]$ — число объектов класса c, которые были классифицированы как t. Посчитайте по данной матрице неточностей средневзвешенную по классам микро, макро и обычную F-меру.

Входные данные

Первая строка содержит целое число K — число классов ($1 \leq K \leq 20$). Далее идёт K строк — описание матрицы неточностей. Каждая строка c содержит K целых чисел — c-я строка матрицы неточностей. $\forall c,t : 0 \leq CM[c,t] \leq 100$ и $\exists c,t : CM[c,t] \geq 1$.

Выходные данные

Выведите три вещественных числа с плавающей точкой — взвешенно усреднённую по классам микро, макро и обычную F-меру. Абсолютная погрешность ответа не должна превышать 10^{-6}.

входные данные

```
2
0 1
1 3
```

выходные данные

```
0.705882353
0.600000000
0.600000000
```

входные данные
3 3 1 1 3 1 1 1 3 1

выходные данные
0.333333333 0.326860841 0.316666667

В первом примере классы распределены как 1:4. Точность (precision), полнота (recall) и F-мера первого класса равны 0, а второго 0.75. При этом средняя точность, полнота и F-мера равны 0.6.

G. Индекс Джини

1 секунда, 256 мегабайт

Требуется оценить хаотичность разбиения упорядоченных объектов на два множества всеми возможными способами при помощи Индекса Джини.

Входные данные

Первая строка содержит два разделённых пробелом натуральных числа N и K $(2 \le N, K \le 10^5)$ — число объектов и классов.

Вторая строка содержит N разделённых пробелом натуральных чисел c_i $(1 \le c_i \le K)$ — классы соответствующих объектов.

Выходные данные

Выведите $N - 1$ вещественное число с плавающей точкой — оценку хаотичности разбиений.

Ответ считается верным, если его относительная или абсолютная погрешность не превышает 10^{-9}.

входные данные
5 3 1 2 2 3 3

выходные данные
0.4 0.46666666666666656 0.2666666666666666 0.5

входные данные
5 3 1 2 3 2 1

выходные данные
0.5 0.6 0.6 0.5

В первом примере оценка третьего разбиения вычисляется следующим образом:

$$\frac{3}{5}\left(1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2+\left(\frac{0}{3}\right)^2\right)\right)+\frac{2}{5}\left(1-\left(\left(\frac{0}{2}\right)^2+\left(\frac{0}{2}\right)^2+\left(\frac{\;}{2}\right)^2\right)\right)=\frac{\;}{15}.$$

H. Условная энтропия

1 секунда, 256 мегабайт

Вычислите критерий связи двух категориальных признаков X и Y на основе математического ожидания условной энтропии $H(Y|X)$. Вероятности оцениваются обыкновенным частотным методом. При расчётах используйте натуральный логарифм $\ln(x)$ либо логарифм идентичный натуральному $\log_e(x)$.

Входные данные

Первая строка содержит два целых положительных числа K_x и K_y ($1 \le K_x, K_y \le 10^5$) — максимальное число различных значений признаков X и Y.

Следующая строка содержит целое положительное число N ($1 \le N \le 10^5$) — число объектов.

Следующие N строк содержат описания соответствующих объектов. Каждая из этих N строк содержит описание одного объекта: два целых положительных числа x и y $(1 \le x \le K_x, 1 \le y \le K_y)$ — значения признаков X и Y.

Выходные данные

Выведите одно вещественное число с плавающей точкой — математическое ожидание условной энтропии. Допустимая абсолютная и относительная погрешность 10^{-6}.

входные данные
2 3 5 1 2 2 1 1 1 2 2 1 3

выходные данные
0.9364262454248438

I. Марковская цепь

1 секунда, 256 мегабайт

Даны несколько строк. Известно, что почти все они были получены сэмплированием из одной марковской цепи, но одна строка получена из простого случайного распределения, в котором каждая буква выбирается независимо от остальных.

Найдите эту строку.

Входные данные

Первая строка содержит натуральное число N $(3 \le N \le 10)$ — число строк.

Далее следует N строк, которые состоят только из маленьких латинских букв и пробелов. Сумма длин всех строк не превышает 10^4.

Выходные данные

Выведите одно натуральное число — номер строки, которая была получена из простого случайного распределения. Строки нумеруются с единицы.

входные данные
3 sus asus sus xsus uss axss

выходные данные
3

J. Коэффициент ранговой корреляции Спирмена

1 секунда, 256 мегабайт

Посчитайте ранговую корреляцию Спирмена двух численных признаков.

Входные данные

Первая строка содержит целое положительное число N ($1 \le N \le 10^5$) — число объектов.

Следующие N строк содержат описания соответствующих объектов. Каждая из этих N строк содержит описание одного объекта: два целых числа x_1 и x_2 $(-10^9 \le x_1, x_2 \le 10^9)$ — значения первого и второго признака описываемого объекта. Гарантируется, что все значения каждого признака различны.

Выходные данные

Выведите одно вещественное число с плавающей точкой — коэффициент ранговой корреляции Спирмена двух признаков у заданных объектов. Допустимая абсолютная и относительная погрешность 10^{-6}.

входные данные

```
5
1 16
2 25
3 1
4 4
5 9
```

выходные данные

```
-0.500000000
```

K. k-ближайших соседей

2 секунды❓, 256 мегабайт

Требуется ответить на несколько запросов вычисления среднего среди k-ближайших объектов к запросу. Все объекты одномерные, если не считать целевой признак.

Входные данные

Первая строка содержит одно целое положительное число n ($1 \le n \le 10^5$) — число объектов.

Следующие n строк содержат описание объектов. Каждая из этих строк содержит два разделённых пробелом целых числа: x_i ($|x_i| \le 10^9$) и y_i $(1 \le y_i \le 10^9)$ — значения обычного и целевого признака i-го объекта. Гарантируется, что все x_i различны.

Далее следует строка с одним целым положительным числом m ($1 \le m \le 2 \cdot 10^4$) — число запросов.

Следующие m строк содержат описание запросов. Каждая из этих строк содержит два разделённых пробелом целых числа: x_q ($|x_q| \le 10^9$) и k_q $(1 \le k_q \le n)$ — положение запроса и интересующее число ближайших объектов к нему.

Выходные данные

Для каждого запроса выведите одно число — среднее значение целевого признака k-ближайших объектов. Если нельзя однозначно выбрать k-ближайших объектов, то выведите -1.

входные данные

```
5
1 4
5 3
3 4
7 2
9 8
4
2 1
6 2
5 3
8 4
```

выходные данные

```
-1.0
2.5
3.0
4.25
```