Задание 1

Дано
Данные хранятся в блоках памяти. Пользователь отправляет команды чтения данных из блоков в некотором порядке. Некоторые из этих команд можно объединить в кластеры, т.е. в группы команд в которых блоки часто читаются последовательно (друг за другом). Пусть уже известны 2 алгоритма кластеризации команд, но неизвестно как сравнить какой из алгоритмов для множества последовательностей команд.

Пример
Последовательность команд: 2 5 7 2 5 3 8 7 2 5 7 2 3 5 7 2 5 7
Алгоритм A1 может выделить кластер 2 5 7
Алгоритм A2 может выделить кластер 5 7 2 5

Найти
предложить метрику для оценки качества алгоритмов кластеризации в двух случаях:
a) ограничения на память и вычислительные ресурсы отсутствуют,
b) есть ограничения на память и вычислительные ресурсы.
(написать явно формулу и/или алгоритм вычисления метрики)

Решение
Среди метрик, используемых в кластерном анализе, можно выделить «вариации на тему» евклидова расстояния, т.е. метрики, которые так или иначе используют разность между координатами многомерных векторов.
В данном случае мы можем применить два подхода:
1) Применить классическую метрику Евклида, т.е. расстояние между двумя *n*-мерными векторами (кластером и фрагментом потока команд) определяется нормой разности этих векторов:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=0}^{n-1} (x_i - y_i)^2}$$

Тогда алгоритм кластеризации пытается минимизировать суммарное квадратичное отклонение точек кластера от их текущего центра **c** (m – число элементов в кластере):

$$E = \sum_{j=0}^{m-1} d(\boldsymbol{c}, \boldsymbol{x}_j)^2 = \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} (c_i - x_{ji})^2$$

Сравнение двух кластеров как раз можно осуществить, сравнив отклонения E, однако такое сравнение необходимо привести к общему масштабу (в условии задачи кластеры определены над векторными пространствами разной размерности). С этой целью введем понятие относительного евклидового расстояния:

$$d'(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{\sum_{i=0}^{n-1} (x_i - y_i)^2}{n}}$$

Тогда расстояния в пространствах разной размерности можно будет сравнивать. Также от размерности зависит число элементов в кластере. Для начала рассмотрим правило формирования элементов кластера из последовательности команд. Вектор заданной длины должен формироваться из последовательности команд на каждом дискретном такте поступления новой команды – поскольку совпадение с кластером может начаться в любой момент. Т.о. если всего команд L, то для заданной размерности векторного

пространства n число элементов в кластере будет равно $m = L - n$. Тогда логичным было бы усреднить ошибку по отношению к числу элементов кластера (за исключением текущего центра, т.е. $m - 1$):

$$E = \sum_{j=0}^{m-1} \frac{d'(\boldsymbol{c}, \boldsymbol{x}_j)^2}{m-1} = \frac{1}{n(m-1)} \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} (c_i - x_{ji})^2$$

Для примера из условий задачи:

Элементы кластера A1: [**[2, 5, 7]**, [5, 7, 2], [7, 2, 5], [2, 5, 3], [5, 3, 8], [3, 8, 7], [8, 7, 2], [7, 2, 5], **[2, 5, 7]**, [5, 7, 2], [7, 2, 3], [2, 3, 5], [3, 5, 7], [5, 7, 2], [7, 2, 5], **[2, 5, 7]**]

$$\boldsymbol{c} = [2,5,7]$$

Элементы кластера A2: [[2, 5, 7, 2], **[5, 7, 2, 5]**, [7, 2, 5, 3], [2, 5, 3, 8], [5, 3, 8, 7], [3, 8, 7, 2], [8, 7, 2, 5], [7, 2, 5, 7], [2, 5, 7, 2], [5, 7, 2, 3], [7, 2, 3, 5], [2, 3, 5, 7], [3, 5, 7, 2], **[5, 7, 2, 5]**, [7, 2, 5, 7]]

$$\boldsymbol{c} = [5,7,2,5]$$

Жирным выделены текущие центры кластеров.

Тогда

$$E1 = \frac{1}{42} \sum_{j=0}^{14} \sum_{i=0}^{2} (c_i - x_{ji})^2 = 8.711111111$$

$$E2 = \frac{1}{52} \sum_{j=0}^{13} \sum_{i=0}^{3} (c_i - x_{ji})^2 = 8.232142857$$

Видно, что отклонение у кластера во втором случае меньше. Также можем заключить что сложность такого метода сравнения алгоритмов кластеризации $O(mn)$ возведений в квадрат.

Другим способом сравнения в евклидовой метрике является сравнение текущего центра **c** с центром масс кластера **s**. Каждая координата центра масс кластера вычисляется по формуле:

$$s_i = \frac{1}{m} \sum_{j=0}^{m-1} x_{ji}, \;\; i = 0 \ldots n-1$$

Затем идет сравнение центра масс и текущего центра, для упрощения вычислений можно сравнивать квадраты относительных расстояний

$$D = d'(\boldsymbol{s}, \boldsymbol{c})^2 = \frac{1}{n} \sum_{i=0}^{n-1} (s_i - c_i)^2$$

Для условий задачи

$$\boldsymbol{s}1 = [4.500000000, 4.687500000, 4.812500000]$$

$$D1 = \frac{1}{3}\sum_{i=0}^{2}(s_i - c_i)^2 = 3.710937500$$

$$\boldsymbol{s}2 = [4.666666667, 4.666666667, 4.666666667, 4.666666667]$$

$$D2 = \frac{1}{4}\sum_{i=0}^{3}(s_i - c_i)^2 = 3.194444445$$

Расстояние от центра масс во втором случае меньше (т.е. второй алгоритм, как и в случае со среднеквадратичным отклонением, показал лучший результат). Сложность вычислений в этот раз $O(n)$ возведений в квадрат плюс затраты на вычисление центра масс, т.е. вычислительная сложность ниже.

2) Второй подход использует другую метрику. Поскольку в данном случае важно полное совпадение координат (речь идет о координатах в отдельности, а не векторе целиком). Если координаты не совпадают, то неважно на сколько – это уже просто другая команда. Тогда можно использовать метрику Хемминга, однако в кластерном анализе используют «Процент несогласия» - меру инвариантную к длине размерности векторного пространства кластера (аналогично относительной евклидовой метрике):

$$d'(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\begin{cases}0, x_i = y_i\\1, x_i \neq y_i\end{cases}$$

Используя такую метрику, можем произвести сравнение лишь на основе среднего отклонения от текущего центра. Поскольку сравнение с центром масс в метрике Хэмминга скорее всего даст максимальное отклонение в обоих случаях. Среднее отклонение определим по аналогии с евклидовой метрикой выражением:

$$E = \sum_{j=0}^{m-1}\frac{d'(\boldsymbol{c}, \boldsymbol{x}_j)}{m-1} = \frac{1}{n(m-1)}\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\begin{cases}0, c_i = x_{ji}\\1, c_i \neq x_{ji}\end{cases}$$

Тогда для условий задачи:

$$E1 = \frac{1}{42}\sum_{j=0}^{14}\sum_{i=0}^{2}\begin{cases}0, c_i = x_{ji}\\1, c_i \neq x_{ji}\end{cases} = 0.7333333333$$

$$E2 = \frac{1}{52}\sum_{j=0}^{13}\sum_{i=0}^{3}\begin{cases}0, c_i = x_{ji}\\1, c_i \neq x_{ji}\end{cases} = 0.7857142857$$

Отклонение от текущего центра в кластере по алгоритму А1 меньше, т.е. результат противоположный полученному в метрике Евклида. Однако текущий результат можно считать более достоверным, поскольку важным является не расстояние от вектора команд до текущего центра кластера, а совпадение либо не совпадение координат. Сложность вычисления отклонения $O(mn)$ операций сравнения.