Задача 1

Пусть имеется информация о покупках, совершаемых 100 миллионами людей, каждый из которых ходит в магазин в среднем 100 раз в году и покупает 10 из 1000 предложенных товаров. Предположим, что два человека попадают под подозрение, если они купили в течение года в точности один и тот же набор товаров (возможно, для изготовления бомбы?). С помощью принципа Бонферрони определите, будет ли эффективным метод выявления террористов, основанный на поиске таких пар людей.

Задача 2

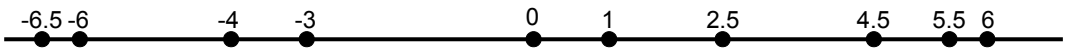
Рассмотреть смесь из D-мерных распределений Бернулли. В такой смеси $\mathbf{x}$ – D-мерный бинарный вектор, каждый компонент x_i которого подчиняется распределению бернулли с параметром μ_{ki} при заданном векторе μ_k:

$$p(\mathbf{x}|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}$$

Вероятность k-го вектора μ_k равна π_k. Выписать выражения для E и M шагов при исользовании EM алгоритма для нахождения неизвестных параметров μ_k и π_k.

Задача 3

На рисунке показан набор из 10 точек, расположенных на прямой. Примените алгоритм иерархической кластеризации с single-link расстоянием между кластерами. Постройте дендрограмму.



Какова алгоритмическая сложность этого алгоритма?

Задача 4

В таблице даны попарные расстояния между объектами из обучающей выборки. Проведите иерархическую кластеризацию с использованием complete-link расстояния между кластерами.

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0.0	1.5	5.0	4.0	2.5	0.5
x_2		0.0	4.0	0.5	3.5	2.0
x_3			0.0	6.0	2.0	1.0
x_4				0.0	5.5	4.5
x_5					0.0	1.0
x_6						0.0

Задача 5

Пусть алгоритм, кластеризуюзщий точки в многомерном Евклидовом пространстве, оптимизирует критерий (k задано)

$$J = \frac{1}{2} \sum_{k} \sum_{x_i \in C_k} \sum_{x_j \in C_k} \|x_i - x_j\|^2.$$

Покажите, что такой алгоритм эквивалентен стандартному алгоритму k-means.

Задача 6

Пусть даны 2 кластеризации C и Ω одного и того же набора данных. Покажите, что

$$MI(C, \Omega) \leq \frac{1}{2}(H(C) + H(\Omega)),$$

где $MI(C, \Omega)$ – mutual information, а $H(C)$ и $H(\Omega)$ – соответствующие энтропии.

Задача 7

Пусть дана обучающая выборка X_N, которая сгенерирована из распределения Стьюдента с неизвестными параметрами μ и σ и известным количеством степеней свободы ν. Используя принцип максимального правдоподобия, получите оценки для неизвестных параметров μ и σ.