

Задание 1

Предположим X и Y две случайные переменные отражающие возраст и вес, соответственно. Рассмотрим случайную выборку из 20 наблюдений

$X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$

$Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220)$

Необходимо:

1. Найти среднее, медиану и моду величины X
2. Найти дисперсию Y
3. Построить график нормального распределения для X
4. Найти вероятность того, что возраст больше 80
5. Найти двумерное мат. ожидания и ковариационную матрицу для этих двух величин
6. Определять корреляцию между X и Y
7. Построить диаграмму рассеяния, отображающая зависимость между возрастом и весом

Задание 2

Для следующего набора данных

	X_1	X_2	X_3
a	17	17	12
b	11	9	13
c	11	8	19

Рассчитайте ковариационную матрицу и обобщенную дисперсию

Задание 3

Даны два одномерных нормальных распределения N_a и N_b с мат. ожиданиями 4, 8 и СКО 1, 2 соответственно.

1. Для каждого из значения $\{5, 6, 7\}$ определите какое из распределений сгенерировало значение с большей вероятностью.
2. Найди значение, которой могло быть сгенерировано обоими распределениями с равной вероятностью

Задание 4

Рассмотрим данные

i	a_1	a_1
x_1	4	2.9
x_2	2.5	1
x_3	3.5	4

x_4	2	2.1
-------	---	-----

Строки соответствуют наблюдениям, столбцы признакам. x_{i1} - Означает 1-й признак i-го наблюдения.

Есть ядро (функция сходства):

$$K(x_i, x_j) = \|x_i - x_j\|^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$$

Рассчитайте ядерную матрицу. Ядерная матрица должна получиться размером 4x4, на главной диагонали которой стоят 0. По смыслу, ядерная матрица отражает близость между двумя наблюдениями согласно метрике, задаваемой функцией сходства.

Для справки смотрите материалы лекции №2.

Задание 5

Рассмотрим данные в виде матрицы D:

	a_1	a_2
x_1	8	-20
x_2	0	-1
x_3	10	-19
x_4	10	-20
x_5	2	0

Строки соответствуют наблюдениям, столбцы признакам.

1. Рассчитайте среднее μ и ковариационную матрицу Σ для матрицы D ([справка](#))
2. Рассчитайте собственные числа для матрицы Σ ([справка](#))
3. **Рассчитайте матрицу корреляций**

Задание 6

Для данных и полученных результатов из второго задания:

1. Рассчитай первые две главные компоненты ([справка](#))
2. На основании собственных чисел рассчитайте остаточную дисперсию для каждой компоненты
3. **В декартовых координатах исходного пространства, отобразите вектора соответствующие главным компонентам. (таким образом будет видно, на какие вектора проводилась проекция). В качестве точки, откуда начинаются вектора, используйте μ**

Задание №7

Дан набор данных:

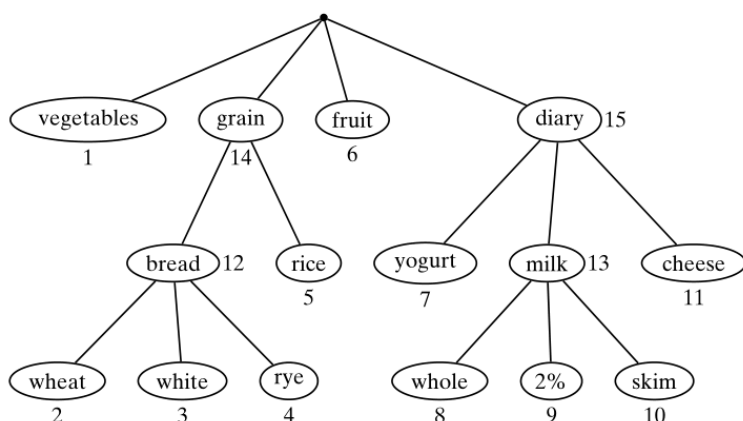
tid	itemset
t_1	<i>ABCD</i>
t_2	<i>ACDF</i>
t_3	<i>ACDEG</i>
t_4	<i>ABDF</i>
t_5	<i>BCG</i>
t_6	<i>DFG</i>
t_7	<i>ABG</i>
t_8	<i>CDFG</i>

Предположив, что минимальный уровень поддержки равен 3 / 8. Продемонстрируйте, как алгоритм Apriori перебирает данный набор данных.

Предположив, что минимальный уровень поддержки равен 2 / 8. Продемонстрируйте, как алгоритм FPGrowth перебирает данный набор данных.

Задание №8

На рисунке представлена классификация различных продуктов. Каждый лист дерева это конкретный продукт, внутренний узел дерева представляет категорию продукта более верхнего уровня.



Был получен следующий набор данных:

tid	itemset
1	2 3 6 7
2	1 3 4 8 11
3	3 9 11
4	1 5 6 7
5	1 3 8 10 11
6	3 5 7 9 11
7	4 6 8 10 11
8	1 3 5 8 11

Выполните следующее задание:

1. Каков размер области поиска наборов элементов, если ограничиваться только наборами, состоящими из простых элементов?
2. Предположив, что минимальный уровень поддержки = 7/8. Найдите все часто встречающиеся наборы элементов, состоящие только из элементов высокого уровня в таксономии. Имейте в виду, что если в транзакции появляется простой элемент, предполагается, что все его предки высокого уровня также присутствуют в транзакции.

Задание №9

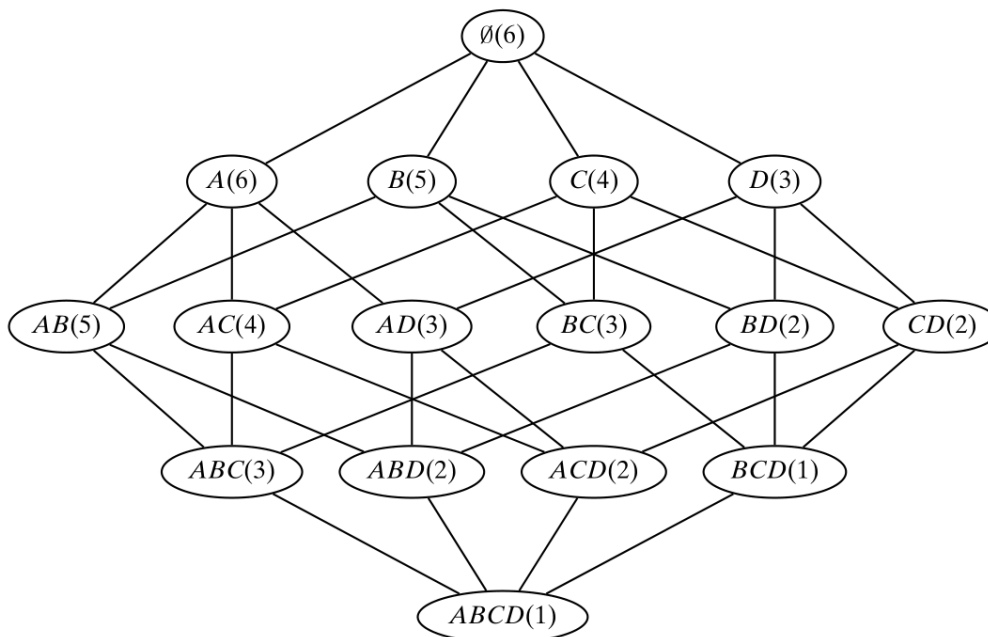
Дан набор данных:

Tid	Itemset
1	<i>ACD</i>
2	<i>BCD</i>
3	<i>AC</i>
4	<i>ABD</i>
5	<i>ABCD</i>
6	<i>BCD</i>

Найдите все минимальные генераторы для минимального уровня поддержки = 1. Множество является минимальным генератором, когда оно не имеет подмножеств с тем же уровнем поддержки.

Задание №10

Дана решетка наборов и их частоты.



Выполните следующие задание:

1. Выпишите список всех замкнутых наборов (closed itemsets). Множество замкнутое когда нет надмножеств с тем же уровнем поддержки.

2. Является ли набор BCD выводимым? Является ли набор ABCD выводимым? Какие границы их поддержки?. (стр. 250 в учебнике)

Задание №11

Для алфавита {A,C,G,T} посчитайте, сколько всего может быть разных последовательностей длины k

Даны последовательности

Id	Sequence
s_1	AATACAAGAAC
s_2	GTATGGTGAT
s_3	AACATGGCCAA
s_4	AAGCGTGGTCAA

Найдите все подпоследовательности с минимальным уровнем поддержки = 4

Задание №12

Дан набор значений 2,4,10,12,3,20,30,11,25. Предположим количество кластеров $k = 3$, и выбраны начальные средние значения $m_1 = 2$, $m_2 = 4$, $m_3 = 6$. Покажите, какие кластеры будут сформированы после первой итерации алгоритма k-средних, и рассчитайте новые значения центров кластеров для следующей итерации.

Задание №13

Дан набор точек x и вероятности из принадлежности к кластерам C_1 и C_2 .

x	$P(C_1 x)$	$P(C_2 x)$
2	0.9	0.1
3	0.8	0.1
7	0.3	0.7
9	0.1	0.9
2	0.9	0.1
1	0.8	0.2

Выполните следующие задание:

1. Найдите оценку максимального правдоподобия для средних μ_1 и μ_2

2. Предположим, что $\mu_1 = 2$, $\mu_2 = 7$ и $\sigma_1 = \sigma_2 = 1$. Найдите вероятности принадлежности точки $x = 5$ к кластерам C_1 и C_2 . Априорные вероятности каждого кластера $P(C_1) = P(C_2) = 0.5$ и $P(x = 5) = 0.029$

Задание №14

Даны категориальные данные размерности 5

Point	X_1	X_2	X_3	X_4	X_5
$\mathbf{x}_1$	1	0	1	1	0
$\mathbf{x}_2$	1	1	0	1	0
$\mathbf{x}_3$	0	0	1	1	0
$\mathbf{x}_4$	0	1	0	1	0
$\mathbf{x}_5$	1	0	1	0	1
$\mathbf{x}_6$	0	1	1	0	0

Близость двух наблюдений определяется через количество совпадений и несовпадений значений признаков. Допустим, что n_{11} количество признаков одновременной равных 1 для наблюдений $\mathbf{x}_i$ и $\mathbf{x}_j$, и n_{10} количество признаков равных 1 для наблюдения $\mathbf{x}_i$ и в то же время равных 0 для наблюдения $\mathbf{x}_j$. По аналогии определяются значения n_{01} and n_{00} :

		$\mathbf{x}_j$	
		1	0
$\mathbf{x}_i$	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

Определим следующие метрики:

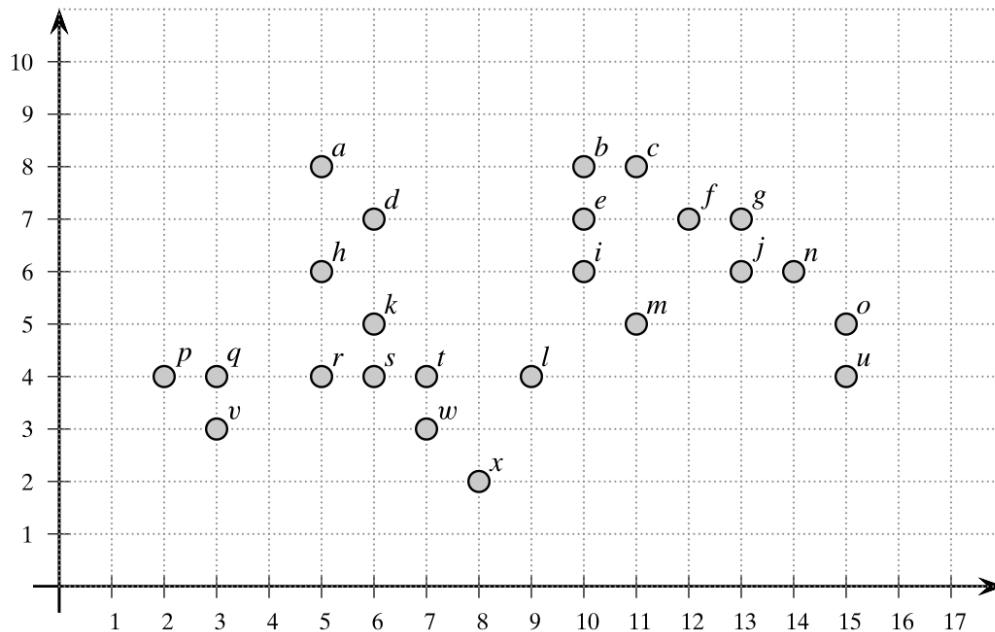
- Коэффициент простого совпадения $SMC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$
- Коэффициент Жаккара $JC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$
- Коэффициент Рассела и Рао $RC(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11} + n_{10} + n_{01} + n_{00}}$

Постройте дендограммы полученные после иерархической кластеризации при следующих параметрах:

- Метод одиночной связи с метрикой RC
- Метод полной связи с метрикой SMC
- Невзвешенный центроидный метод с метрикой JC

Задание №15

Дан рисунок



Допустим, что используется Евклидово расстояние, $\epsilon = 2$ и $\text{minPts} = 3$. Выполните следующие задачи:

1. Выпишите список всех основных точек
2. Покажите, является ли точка a прямо достижимой из точки d
3. Покажите, является ли точка o достижимой по плотности из точки i . Если нет, то покажите на какой точке цепочка построения пути оборвалась.
4. Покажите кластеры полученные алгоритмом DBSCAN и выпавшие точки.

Задание №16

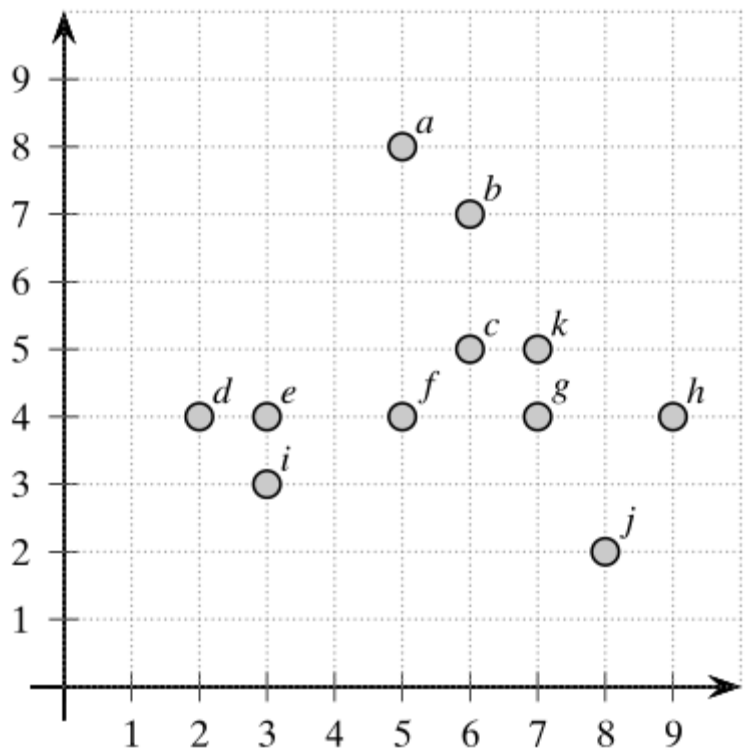
Даны следующие метрики:

1.
$$L_{\infty}(\mathbf{x}, \mathbf{y}) = \max_{i=1}^d \{|x_i - y_i|\}$$
2.
$$L_{\frac{1}{2}}(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^d |x_i - y_i|^{\frac{1}{2}} \right)^2$$
3.
$$L_{\min}(\mathbf{x}, \mathbf{y}) = \min_{i=1}^d \{|x_i - y_i|\}$$

$$L_{pow}(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^d 2^{i-1} (x_i - y_i)^2 \right)^{1/2}$$

4.

Для данных представленных следующим рисунком



Используя метод DBSCAN проведите кластеризацию при следующих параметрах:

- $\epsilon = 2$ и $\text{minPts} = 5$ и метрика (1)
- $\epsilon = 4$ и $\text{minPts} = 3$ и метрика (2)
- $\epsilon = 1$ и $\text{minPts} = 6$ и метрика (3)
- $\epsilon = 4$ и $\text{minPts} = 6$ и метрика (3)

Для всех случаев построить кластеры и отобразить основные точки, достижимые по плотности точки и выпавшие точки

Задание №17

Даны следующие данные

$\mathbf{x}_i$	a_1	a_2	a_3	Class
$\mathbf{x}_1$	T	T	5.0	Y
$\mathbf{x}_2$	T	T	7.0	Y
$\mathbf{x}_3$	T	F	8.0	N
$\mathbf{x}_4$	F	F	3.0	Y
$\mathbf{x}_5$	F	T	7.0	N
$\mathbf{x}_6$	F	T	4.0	N
$\mathbf{x}_7$	F	F	5.0	N
$\mathbf{x}_8$	T	F	6.0	Y
$\mathbf{x}_9$	F	T	1.0	N

Используя наивный байесовский классификатор определите класс точки (T,F,1.0)

Задание №18

Даны два класса c_1 and c_2 со следующими мат. ожиданиями и матрицами ковариации:

$$\mu_1 = (1, 3)$$

$$\mu_2 = (5, 5)$$

$$\Sigma_1 = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Классифицируйте точку (3,4) используя Байесовский вывод, предположив, что классы распределены по нормальному закону, и $P(c_1) = P(c_2) = 0.5$

Задание №19

Даны следующие данные

Point	Age	Car	Risk
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Постройте решающее дерево используя порог для чистоты (purity threshold) равным 100%.

В качестве критерия для разделения используйте энтропию. Классифицируйте наблюдение (Age=27, Car=Vintage)

Задание №20

Даны следующие данные

i	$\mathbf{x}_i$	y_i
$\mathbf{x}_1$	(4,2.9)	1
$\mathbf{x}_2$	(3.5,4)	1
$\mathbf{x}_3$	(2.5,1)	-1
$\mathbf{x}_4$	(2,2.1)	-1

1. Рассчитайте μ_{+1} и μ_{-1} , а так матрицу B - матрица межклассового разброса
2. Рассчитайте S_{+1} и S_{-1} , а также матрица S - матрица внутриклассового разброса
3. Найдите направление w , которое лучше всего дискриминирует классы
4. На направлении w , найдите точку, которая лучше всего делит классы

Задание №21

Даны следующие данные, принадлежащие к двум классам (y) и множители Лагранжа (α)

i	x_{i1}	x_{i2}	y_i	α_i
$\mathbf{x}_1$	4	2.9	1	0.414
$\mathbf{x}_2$	4	4	1	0
$\mathbf{x}_3$	1	2.5	-1	0
$\mathbf{x}_4$	2.5	1	-1	0.018
$\mathbf{x}_5$	4.9	4.5	1	0
$\mathbf{x}_6$	1.9	1.9	-1	0
$\mathbf{x}_7$	3.5	4	1	0.018
$\mathbf{x}_8$	0.5	1.5	-1	0
$\mathbf{x}_9$	2	2.1	-1	0.414
$\mathbf{x}_{10}$	4.5	2.5	1	0

- Рассчитайте уравнение гиперплоскости $h(x)$ для метода опорных векторов

- Какое расстояние от точки x_6 до гиперплоскости $h(x)$. Находится ли эта точка в границах опорных векторов?
- Классифицируйте точку $(3, 3)$ используя гиперплоскость $h(x)$