Фонд оценочных средств ООП «Искусственный интеллект и инженерия данных» по направлению 09.03.04 – Программная инженерия
Дисциплина «Технологии аналитической обработки информации»

Контрольный опрос 4
"Классификация"
(Пример вопросов теста)

1) Укажите верное определение задачи классификации.

	Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее известно.
	Нахождение часто встречающихся зависимостей между классами объектов.
	Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее не известно.
	Разделение заданного множества объектов на два класса: те, что могут быть использованы для принятия стратегических решений, и остальные.

2) Укажите верную последовательность выполнения этапов процесса классификации.

	Построение модели на основе обучающей выборки➔Оценка точности модели на основе тестовой выборки➔Классификация ранее неизвестных данных
	Оценка точности модели на основе обучающей выборки➔Построение модели на основе тестовой выборки➔Классификация ранее неизвестных данных
	Оценка точности модели на основе тестовой выборки➔Построение модели на основе обучающей выборки➔Классификация ранее неизвестных данных
	Оценка точности модели на основе обучающей выборки➔Оценка точности модели на основе тестовой выборки➔Классификация ранее неизвестных данных
	Построение модели на основе тестовой выборки➔Оценка точности модели на основе обучающей выборки➔Классификация ранее неизвестных данных

3) Вычислите Gini index для атрибута **Shirt Size**:

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

<table>
<tr>
<td>1.</td>
<td>Текущий контроль</td>
<td>Контрольный опрос 4 "Классификация"</td>
<td>
Пример вопросов теста:
1) Укажите верное определение задачи классификации.
<table>
<tr><td></td><td>Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее известно.</td></tr>
<tr><td></td><td>Нахождение часто встречающихся зависимостей между классами объектов.</td></tr>
<tr><td></td><td>Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее не известно.</td></tr>
<tr><td></td><td>Разделение заданного множества объектов на два класса: те, что могут быть использованы для принятия стратегических решений, и остальные.</td></tr>
</table>
2) Укажите верную последовательность выполнения этапов процесса классификации.
<table>
<tr><td></td><td>Построение модели на основе обучающей выборки->Оценка точности модели на основе тестовой выборки->Классификация ранее неизвестных данных</td></tr>
<tr><td></td><td>Оценка точности модели на основе обучающей выборки-> Построение модели на основе тестовой выборки->Классификация ранее неизвестных данных</td></tr>
<tr><td></td><td>Оценка точности модели на основе тестовой выборки-> Построение модели на основе обучающей выборки-> Классификация ранее неизвестных данных</td></tr>
<tr><td></td><td>Оценка точности модели на основе обучающей выборки->Оценка точности модели на основе тестовой выборки->Классификация ранее неизвестных данных</td></tr>
<tr><td></td><td>Построение модели на основе тестовой выборки->Оценка точности модели на основе обучающей выборки->Классификация ранее неизвестных данных</td></tr>
</table>
3) Вычислите Gini index для атрибута Shirt Size:
<table>
<tr><th>Customer ID</th><th>Gender</th><th>Car Type</th><th>Shirt Size</th><th>Class</th></tr>
<tr><td>1</td><td>M</td><td>Family</td><td>Small</td><td>C0</td></tr>
<tr><td>2</td><td>M</td><td>Sports</td><td>Medium</td><td>C0</td></tr>
<tr><td>3</td><td>M</td><td>Sports</td><td>Medium</td><td>C0</td></tr>
<tr><td>4</td><td>M</td><td>Sports</td><td>Large</td><td>C0</td></tr>
<tr><td>5</td><td>M</td><td>Sports</td><td>Extra Large</td><td>C0</td></tr>
<tr><td>6</td><td>M</td><td>Sports</td><td>Extra Large</td><td>C0</td></tr>
<tr><td>7</td><td>F</td><td>Sports</td><td>Small</td><td>C0</td></tr>
<tr><td>8</td><td>F</td><td>Sports</td><td>Small</td><td>C0</td></tr>
<tr><td>9</td><td>F</td><td>Sports</td><td>Medium</td><td>C0</td></tr>
<tr><td>10</td><td>F</td><td>Luxury</td><td>Large</td><td>C0</td></tr>
<tr><td>11</td><td>M</td><td>Family</td><td>Large</td><td>C1</td></tr>
<tr><td>12</td><td>M</td><td>Family</td><td>Extra Large</td><td>C1</td></tr>
<tr><td>13</td><td>M</td><td>Family</td><td>Medium</td><td>C1</td></tr>
<tr><td>14</td><td>M</td><td>Luxury</td><td>Extra Large</td><td>C1</td></tr>
<tr><td>15</td><td>F</td><td>Luxury</td><td>Small</td><td>C1</td></tr>
<tr><td>16</td><td>F</td><td>Luxury</td><td>Small</td><td>C1</td></tr>
<tr><td>17</td><td>F</td><td>Luxury</td><td>Medium</td><td>C1</td></tr>
<tr><td>18</td><td>F</td><td>Luxury</td><td>Medium</td><td>C1</td></tr>
<tr><td>19</td><td>F</td><td>Luxury</td><td>Medium</td><td>C1</td></tr>
<tr><td>20</td><td>F</td><td>Luxury</td><td>Large</td><td>C1</td></tr>
</table>
</td>
</tr>
</table>

<table>
<tr>
<td>2.</td>
<td>Текущий контроль</td>
<td>Контрольный опрос 5 "Кластеризация"</td>
<td>Пример вопросов теста:
1) Укажите задачи реальной предметной области, которые являются задачами кластеризации.
<table>
<tr><td></td><td>Определение смысловых групп клиентов банка, взявших кредит, на основе персональных данных этих клиентов.</td></tr>
<tr><td></td><td>Определение адресов электронной почты, которые часто фигурируют совместно в списке адресатов писем.</td></tr>
<tr><td></td><td>Определением смысловых групп писем электронной почты на основе данных о ключевых словах этих писем.</td></tr>
<tr><td></td><td>Определение возможности выдачи кредита клиенту банка на основе персональных данных этого клиента.</td></tr>
<tr><td></td><td>Определение списка клиентов банка, имеющих задолженности по выплатам кредита.</td></tr>
<tr><td></td><td>Определение списка корреспондентов, отправивших наибольшее количество электронных писем.</td></tr>
<tr><td></td><td>Определение пакета услуг, которые часто выбираются клиентами банка совместно с взятием кредита.</td></tr>
<tr><td></td><td>Определение категории письма электронной почты: "спам" или "обычная почта" – на основе данных о ключевых словах этого письма.</td></tr>
</table>
2) Укажите верное определение медоида в алгоритме кластеризации k-medoids.
<table>
<tr><td></td><td>Медоид – объект со случайными координатами, выбираемый в качестве центра кластера в процессе кластеризации.</td></tr>
<tr><td></td><td>Медоид – объект исходного множества, которому в процессе кластеризации присваивается метка кластера.</td></tr>
<tr><td></td><td>Медоид – объект исходного множества, выбираемый в качестве центра кластера в процессе кластеризации.</td></tr>
<tr><td></td><td>Медоид – объект, имеющий средние по всем объектам исходного множества координаты, выбираемый в качестве центра кластера в процессе кластеризации.</td></tr>
</table>
3) Укажите вид алгоритмов кластеризации, к которому относится алгоритм DBSCAN.
<table>
<tr><td></td><td>Плотностные алгоритмы</td></tr>
<tr><td></td><td>Разделительные алгоритмы</td></tr>
<tr><td></td><td>Агломеративные алгоритмы</td></tr>
<tr><td></td><td>Дивизимные алгоритмы</td></tr>
<tr><td></td><td>Нечеткие алгоритмы</td></tr>
</table>
</td>
</tr>
<tr>
<td></td>
<td>Текущий контроль</td>
<td>Практическое задание 1 «Построение хранилища данных»</td>
<td>Вопросы для подготовки к устному опросу:
1. Дайте определение понятия хранилища данных.
2. Опишите этапы технологического цикла аналитической обработки данных.
3. Перечислите методы очистки, трансформации и редукции данных.
4. Кратко опишите сходства и различия схем "звезда" и "снежинка" и проанализируйте их преимущества и недостатки по отношению друг к другу.
5. Объясните принцип ETL (Extract-Transform-Load) построения хранилища данных.</td>
</tr>
</table>

	Текущий контроль	Практическое задание 2 «OLAP-запросы»	Вопросы для подготовки к устному опросу: 1. Дайте определение понятия OLAP-куба. 2. Приведите пример измерений многомерного куба. 3. Дайте определения основных типов OLAP-кубов: полный куб, куб-айсберг, замкнутый куба, оболочка куба. 4. Объясните идеи следующих алгоритмов вычисления OLAP-куба: метод многомерной агрегации, метод нисходящего вычисления подкубов. 5. Объясните отличия между конструкциями ROLLUP BY и CUBE BY.
3.	Текущий контроль	Практическое задание 3 «Поиск частых наборов»	Вопросы для подготовки к устному опросу: 1. Дайте определения следующих понятий: база транзакций, поддержка, частый набор. 2. Объясните принцип антимонотонности поддержки. 3. Объясните работу алгоритма Apriori поиска частых наборов. 4. Объясните идею и схему использования фрагментации базы транзакций для поиска частых наборов. 5. Объясните идею и схему использования сэмплинга базы транзакций для поиска частых наборов.
4.	Текущий контроль	Практическое задание 4 «Поиск ассоциативных правил»	Вопросы для подготовки к устному опросу: 1. Дайте определения следующих понятий: ассоциативное правило, достоверность, устойчивое правило. 2. Объясните алгоритм поиска устойчивых правил с помощью поиска частых наборов. 3. Приведите пример устойчивого, но практически бесполезного правила. 4. Дайте определение меры lift полезности шаблонов. 5. Дайте определения понятий максимально частого и замкнутого набора, иерархии наборов.
5.	Текущий контроль	Практическое задание 5 «Классификация. Деревья решений»	Вопросы для подготовки к устному опросу: 1. Объясните принцип построения дерева решений. 2. Дайте определения критерия выбора атрибута разбиения Information Gain. 3. Дайте определение критерия выбора атрибута разбиения Gain Ratio. 4. Дайте определение критерия выбора атрибута разбиения Gini Index. 5. Дайте определения показателей качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера.
6.	Текущий контроль	Практическое задание 6 «Ансамблевая классификация»	Вопросы для подготовки к устному опросу: 1. Объясните цель и идею ансамблевой классификации. 2. Объясните работу метода бэггинга. 3. Объясните, допустимо ли вхождение в ансамбль, выполняющий бэггинг, разнородных классификаторов. 4. Назовите преимущества и недостатки метода бэггинга. 5. Объясните, почему бэггинг предполагает примерную вероятность 0.632 включения элемента исходной обучающей выборки в выборку участника ансамбля. 6. Объясните работу метода случайного леса. 7. Назовите преимущества и недостатки метода случайного леса. 8. Объясните работу одной из разновидностей метода случайного леса, Forest-RI. 9. Объясните работу одной из разновидностей метода случайного леса, Forest-RC. 10. Объясните работу метода бустинга (на примере алгоритма AdaBoost). 11. Назовите преимущества и недостатки метода бустинга. 12. Укажите ошибку обучения ансамбля классификаторов в методе бустинга. 13. Объясните, как убывает ошибка обучения ансамбля в бустинге при увеличении количества классификаторов в ансамбле.

7.	Текущий контроль	Практическое задание 7 «Кластеризация»	Вопросы для подготовки к устному опросу: 1. Объясните идею разделительной кластеризации и работу алгоритма k Means. 2. Напишите формулу меры для выявления кластеров в k Means (Sum of Squared Errors) 3. Назовите преимущества и недостатки алгоритма k Means. 4. Объясните один из способов (на выбор) подбора начальных центроидов в алгоритме k Means. 5. Объясните работу алгоритма k Medoids. 6. Объясните идею плотностной кластеризации и работу алгоритма DBSCAN. 7. Дайте определения основных понятий, используемых в алгоритме DBSCAN: окрестность точки, корневая точка, непосредственная достижимость, достижимость. 8. Назовите преимущества и недостатки алгоритма DBSCAN. 9. Объясните, почему алгоритм DBSCAN является нечувствительным выбросам и шумам в исходных данных. 10. Объясните, каким можно подбирать параметры $MinPts$ и Pts алгоритма DBSCAN. 11. Объясните идею иерархической кластеризации и работу алгоритма агломеративной кластеризации. 12. Объясните идею иерархической кластеризации и работу алгоритма дивизимной кластеризации. 13. Объясните способ построения дендрограмм. 14. Назовите преимущества и недостатки иерархической кластеризации. 15. Дайте определения следующих мер схожести, используемых в иерархической кластеризации: Single linkage, Complete linkage, Average linkage.
8.	Текущий контроль	Практическое задание 8 «Качество кластеризации»	Вопросы для подготовки к устному опросу: 1. Объясните работу метода локтя определения оптимального количества кластеров. 2. Объясните работу метода кросс-валидации определения оптимального количества кластеров. 3. Дайте определение силуэтного коэффициента и объясните его применение для определения оптимального количества кластеров. 4. Объясните способ оценки неслучайности кластеризуемых данных на основе числа Хопкинса. 5. Объясните способ оценки качества кластеризации на основе предварительной классификации.
9.	Промежуточный контроль	Итоговый тест	Примеры вопросов итогового теста (см. далее).

Паспорт фонда оценочных средств приведен в п. 6.3 РПД.

Разработчик М.Л. Цымблер

ФГАОУ ВО «Южно-Уральский государственный университет
(национальный исследовательский университет)»
Кафедра системного программирования

Дисциплина «Технологии аналитической обработки информации»

ИТОГОВЫЙ ТЕСТ

1) Укажите верное определение термина Data Warehouse (хранилище данных).

	Методы и технологии ввода, структурированного хранения и обработки баз данных в режиме реального времени.
	Методы и технологии поддержки базы данных, которая интегрирует копии фрагментов данных из различных источников и обновляется на регулярной основе.
	Методы и технологии, направленные на обеспечение быстрой подготовки бизнес-отчетов о данных, хранящихся в базе данных.
	Методы и технологии обнаружения скрытых закономерностей (трендов и аномалий) в данных, хранящихся в базе данных.

2) Укажите верное определение термина Data Cleaning (очистка данных).

	Перемещение данных, подвергаемых интеллектуальному анализу, из источников данных в хранилище данных.
	Выявление и исправление ошибок, несоответствий в данных с целью улучшения их качества.
	Отбор результатов интеллектуального анализа данных, полезных в предметной области.
	Отбор данных предметной области, подвергаемых интеллектуальному анализу.

3) Укажите задачу, которая относится к сфере Data Mining.

	Предсказание суммы выигрыша, полученного игроком в игре, предполагающей выбрасывание пары игральных костей, на основе полученных от других игроков данных о вероятностях выигрыша при указанной сумме ставки
	Предсказание суммы выигрыша, полученного игроком в игре, предполагающей выбрасывание пары игральных костей, на основе полученных от других игроков данных об истории ставок и выигрышей других игроков
	Предсказание количества игр, выигранных игроком в игре, предполагающей выбрасывание пары игральных костей, на основе полученных от других игроков данных о вероятностях выбросить определенное количество очков на одной кости за определенное количество игр
	Предсказание количества очков, выброшенных игроком на паре игральных костей, на основе полученных от других игроков данных о вероятностях выбросить определенное количество очков на одной кости

4) Укажите верное определение задачи шаблонов и ассоциативных правил.

	Определение, какие из имеющихся данных могут быть использованы для принятия стратегических решений, а какие – нет.
	Нахождение часто встречающихся зависимостей между объектами.
	Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее известно.
	Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее не известно.

5) Установите соответствие между базовыми задачами интеллектуального анализа данных и приведенными задачами реальной предметной области.

Определение списка корреспондентов, отправивших наибольшее количество электронных писем.		Задача поиска ассоциативных правил
Определение категории письма электронной почты: "спам" или "обычная почта" – на основе данных о ключевых словах этого письма.		Задача НЕ из области интеллектуального анализа данных

Определение адресов электронной почты, которые часто фигурируют совместно в списке адресатов писем.		Задача кластеризации
Определением смысловых групп писем электронной почты на основе данных о ключевых словах этих писем.		Задача классификации

6) Укажите верное определение термина OLAP (OnLine Analitycal Processing, оперативный анализ данных).

	Методы и технологии ввода, структурированного хранения и обработки баз данных в режиме реального времени.
	Методы и технологии, направленные на обеспечение подготовки в режиме реального времени бизнес-отчетов о данных, хранящихся в базе данных.
	Методы и технологии поддержки базы данных, которая интегрирует копии фрагментов данных из различных источников и обновляется на регулярной основе.
	Методы и технологии обнаружения скрытых закономерностей (трендов и аномалий) в данных, хранящихся в базе данных.

7) Укажите верную последовательность этапов создания хранилища данных.

	Извлечение сырых данных→Загрузка данных→Очистка и агрегация данных
	Загрузка данных→Очистка и агрегация данных→Извлечение сырых данных
	Загрузка данных→Извлечение сырых данных→Очистка и агрегация данных
	Извлечение сырых данных→Очистка и агрегация данных→Загрузка данных

8) Пусть имеется хранилище данных с двумя измерениями и одной мерой. Измерения: Год={2015, 2016}, Город={Челябинск, Москва}. Мера - Сумма продаж.
Таблица фактов имеет следующий вид:

Продажи

Год	Город	Сумма
2015	Челябинск	100
2015	Москва	50
2016	Челябинск	200
2016	Москва	80

Укажите верный результат запроса
select Год, Город, sum(Сумма)
from Продажи
CUBE BY (Год, Город);

Выберите один ответ:

Год	Город	Сумма
2015, 2016	Челябинск, Москва	430

Год	Город	Сумма
2015		150
2016		280
		430

Год	Город	Сумма
2015	Челябинск	100
2015	Москва	50
2016	Челябинск	200
2016	Москва	80
		430

Год	Город	Сумма
2015	Челябинск	100
2015	Москва	50
2015		150
2016	Челябинск	200
2016	Москва	80
2016		280
		430

9) Пусть в предметной области имеется 3 измерения, которые могут принимать соответственно 3, 4 и 5 значений, и определены 2 меры. Вычислите объем куба данных.

10) Укажите разновидность хранилища, имеющего приведенную схему:

```
create table Поставщики (
    ИД int primary key,
    Имя char(20),
    Рейтинг int);

create table Продажи (
    Поставщик int foreign key Поставщики (ИД),
    Деталь int foreign key Детали (ИД),
    Место int foreign key Места (ИД),
    СуммаПродажи int,
    КоличПродажи int);

create table Детали (
    ИД int primary key,
    Имя char(20),
    Цена int,
    Производитель int foreign key Производители (ИД));

create table Производители (
    ИД int primary key,
    Имя char(20),
    Рейтинг int);

create table Места (
    ИД int primary key,
    Адрес char(40),
    Город char (15),
    Страна char(3));
```

Выберите один ответ:

	Таблица фактов
	Снежинка
	Звезда
	Созвездие

11) Укажите верное определение поддержки набора.

	Доля транзакций в базе транзакций, которые содержат данный набор.
	Доля транзакций в базе транзакций, которые содержат данный набор БЕЗ других наборов.
	Доля транзакций в базе транзакций, которые НЕ содержат данный набор.
	Доля транзакций в базе транзакций, которые содержат данный набор совместно с другими наборами.

12) Пусть имеются наборы товаров A и B, причем A⊆B. Екажите верное утверждение о поддержке наборов A и B.

- $support(A)<support(B)$
- $support(A)=support(B)$
- $support(A)\geq support(B)$
- $support(A)>support(B)$
- $support(A)\leq support(B)$

13) Пусть имеется множество частых наборов L_3={{a,b,c},{a,b,d},{a,c,d},{a,c,e},{b,c,d}}. Укажите множество кандидатов в частые наборы C_4, которое будет сформировано алгоритмом Apriori.

	{a,c,d,e}
	{{a,b,c,d}, {a,c,d,e}, {a,b,c,e}, {a,b,d,e}}
	{{a,b,c,d}, {a,c,d,e}, {a,b,c,e}}
	{{a,b,c,d}, {a,c,d,e}, {a,b,d,e}}
	{a,b,c,d}
	{{a,b,c,d},{a,c,d,e}}

14) Вычислите значение поддержки ассоциативного правила кола→(орехи,чипсы) для множества корзин.

№ п/п	Корзина
1	вода, кола, хлеб, чипсы, орехи
2	вода, чипсы
3	хлеб, кола, чипсы
4	вода, орехи
5	кола, чипсы
6	вода
7	орехи, кола, хлеб, чипсы
8	кола, хлеб, чипсы
9	кола, чипсы
10	орехи, кола, хлеб, чипсы

15) Укажите один частый 3-элементный набор при minsup=5.

№ п/п	Корзина
1	вода, кола, хлеб, чипсы, орехи
2	вода, чипсы
3	хлеб, кола, чипсы
4	вода, орехи
5	кола, чипсы
6	вода
7	орехи, кола, хлеб, чипсы
8	кола, хлеб, чипсы
9	кола, чипсы
10	орехи, кола, хлеб, чипсы

	(хлеб, чипсы, вода)
	(кола, чипсы, орехи)
	(кола, хлеб, чипсы)
	(орехи, кола, хлеб)
	(орехи, кола, вода)
	(вода, чипсы, орехи)

16) Укажите верное определение тестовой выборки для задачи классификации.

	Пересечение множеств, используемых для построения и проверки модели классификации.
	Множество классифицированных объектов, используемых для построения модели классификации.
	Множество классифицированных объектов, классификация которых должна быть выполнена на основе построенной модели для ее проверки.
	Множество не классифицированных кортежей, классификация которых должна быть выполнена на основе построенной модели.

17) Укажите верное определение термина "подгонка" (overfitting) для задачи классификации.

	Множество не классифицированных кортежей, классификация которых должна быть выполнена на основе построенной модели.
	Множество классифицированных объектов, используемых для построения модели классификации.
	Множество классифицированных объектов, классификация которых должна быть выполнена на основе построенной модели для ее проверки.
	Пересечение множеств, используемых для построения и проверки модели классификации.

18) Используя энтропию, вычислите Info для атрибута $a1$ по следующей выборке:

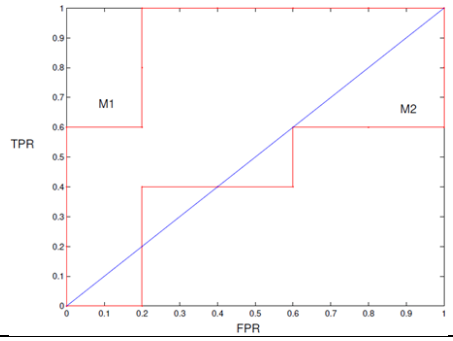
Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

19) Какой из перечисленных атрибутов является наилучшим атрибутом разбиения, если осуществляется построение классификационной модели с двумя классами и следующей обучающей выборкой?

#	Attrubute-1	Attribute-2	Attribute-3	Attribute-4	CLASS
1	A	5	7	Z	Class-1
2	B	5	9	Z	Class-1
3	C	5	7	Z	Class-1
4	C	5	7	Z	Class-1
5	A	5	7	X	Class-1
6	A	8	9	X	Class-2
7	C	8	9	X	Class-2
8	B	8	9	X	Class-2
9	B	8	7	X	Class-2
10	B	8	7	X	Class-2

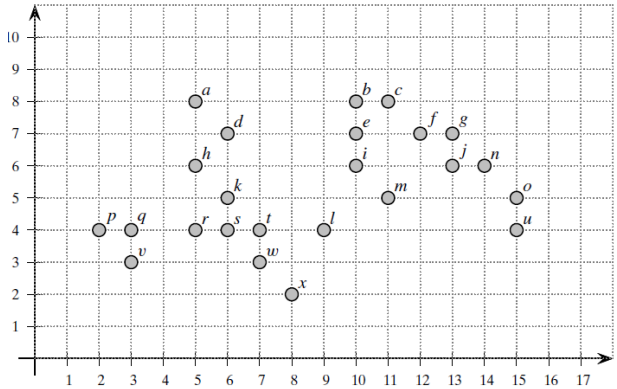
	Attribute-3
	Attribute-2
	Attribute-4
	Attribute-1

20) Выберите верное утверждение о классификаторах M1 и M2 на основе следующего графика их ROC кривых.



	M1 лучше (точнее), чем M2
	График не дает возможности однозначно указать, какой из классификаторов лучше (точнее)
	M2 лучше (точнее), чем M1
	Ценность (точность) M1 и M2 одинакова

21) Пусть выполняется кластеризация следующего множества точек алгоритмом DBSCAN с параметрами MinPts=3, Eps=2. Укажите результирующие кластеры.



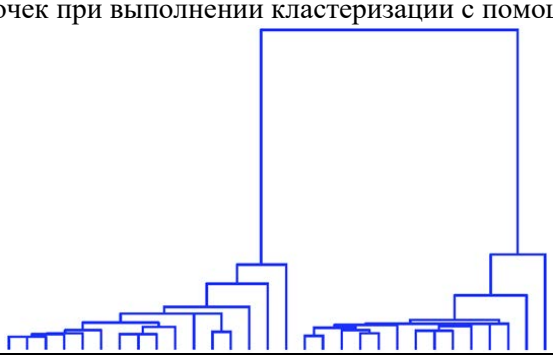
	C1={a, d, h, k, p, q, r, s, t, l, v, w, x} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума отсутствуют
	C1={a, d, h, k, p, q, r, s, t, v, w} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l, x}
	C1={p, q, v} C2={a, d, h, k, r, s, t, w, x} C3={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l}
	C1={a, d, h, k, p, q, r, s, t, v, w, x} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l}
	C1={p, q, v} C2={a, d, h, k, r, s, t, w} C3={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l, x}
	C1={p, q, v} C2={a, d, h, k, l, r, s, t, w, x} C3={b, c, e, f, g, i, j, n, m, o, u} Точки шума отсутствуют

22) Укажите основную идею дивизимных алгоритмов кластеризации.

	Предполагается, что каждый исходный объект образует отдельный кластер, и затем выполняется слияние близких друг к другу объектов или кластеров до тех пор, пока не будет получен единственный кластер или не будет выполнено условие завершения слияния.

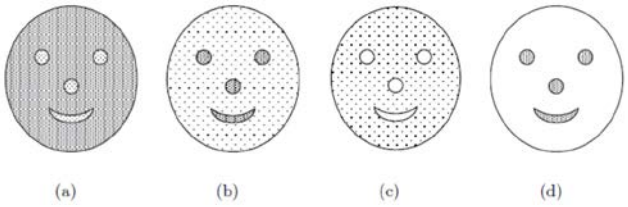
	Предполагается, что все исходные объекты входят в один кластер, и затем итеративно выполняется его разбиение на менее мощные кластеры до тех пор, пока не будут получены кластеры-синглтоны или не будет выполнено условие завершения разбиения.
	Кластеризация выполняется в два этапа: 1) разбиение исходного множества объектов на кластеры (в каждом кластере имеется, по крайней мере, один объект и каждый объект принадлежит в точности одному кластеру); 2) итеративное перемещение объектов между кластерами с целью улучшить начальное разбиение (чтобы объекты из одного кластера были более "близкими", а из разных кластеров – более "далекими").
	Добавление объектов в кластер до тех пор, пока количество соседних объектов не превысит некоторого заданного порога; при этом в окрестности каждого объекта кластера должно находиться некоторое минимальное количество других объектов.

23) Используя следующую дендрограмму, укажите оптимальное количество кластеров для соответствующего множества точек при выполнении кластеризации с помощью алгоритма k-means.



	10
	5
	Имеющиеся данные не позволяют дать однозначный ответ на вопрос
	2
	3

24) Даны четыре множества точек – "лица" на рисунках a, b, c, d. Интенсивность цвета и количество точек показывают плотность. Линии отделяют области точек, при этом НЕ являясь точками. Укажите лица, для которых ни один алгоритм кластеризации НЕ способен отыскать кластеры, соответствующие глазам, носу и рту.



	a
	b
	c
	d

25) Пусть имеется множество, состоящее из четного количества точек метрического пространства. Эти точки разбиты на четное количество кластеров. При выполнении кластеризации используется мера SSE (Sum of Squared Errors): сумма квадратов расстояний от точки кластера до центроида этого кластера, когда суммирование выполняется по всем кластерам. Половина указанных кластеров являются более плотными, другая половина – менее плотными, и соответствующие области хорошо отделимы друг от друга. Укажите свойство, при котором кластеризация данного множества точек дает минимальное значение SSE.

	Центроиды должны быть случайно расположены в более плотной и в менее плотной областях
	Более половины центроидов должны быть расположены в менее плотной области
	Центроиды должны быть поровну расположены в более плотной и в менее плотной областях
	Более половины центроидов должны быть расположены в более плотной области
	Все центроиды кластеров должны быть расположены в более плотной области
	Все центроиды кластеров должны быть расположены в менее плотной области