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

Итоговый тест
(Примеры вопросов итогового теста)

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 Mining (интеллектуальный анализ данных).

	Методы оптимизации запросов к сверхбольшим базам данных, в т.ч. в режиме реального времени.
	Методы и технологии ввода, структурированного хранения и обработки баз данных в режиме реального времени.
	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.

3) Укажите **три** позиции, которые **НЕ** являются непосредственными задачами предварительной обработки данных.

	Интеллектуальный анализ данных
	Очистка данных
	Интеграция данных
	Редукция данных
	Оперативный анализ данных
	Нормализация таблиц хранилища данных

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

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

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 схему метода бустинга ансамблевой классификации.

	1. Применить сэмплинг без повторений к исходной обучающей выборке и сформировать выборки для классификаторов-участников ансамбля. Обучить каждого участника на своей выборке. 2. Получить от каждого участника ансамбля класс ранее не известного объекта. Получить итоговый класс ранее неизвестного объекта на основе мажоритарного голосования участников ансамбля.
	1. Применить сэмплинг с повторением к исходной обучающей выборке и сформировать выборки для классификаторов-участников ансамбля. Обучить каждого участника на своей выборке. 2. Получить от каждого участника ансамбля класс ранее не известного объекта. Получить итоговый класс ранее неизвестного объекта на основе мажоритарного голосования участников ансамбля.
	1. Присвоить одинаковые веса объектам исходной обучающей выборки. 2. Применить сэмплинг с повторением на основе весов объектов к исходной обучающей выборке и сформировать выборку для классификатора-участника ансамбля. Обучить участника на своей выборке. 3. Оценить точность участника ансамбля на объектах исходной обучающей выборки. Повысить вес неверно классифицированных объектов исходной обучающей выборки. Вычислить вес участника ансамбля на основе показанной им точности классификации. 4. Повторить пп. 2, 3 последовательно для каждого участника ансамбля. 5. Получить от каждого участника ансамбля класс ранее не известного объекта. Получить итоговый класс ранее неизвестного объекта на основе взвешенного голосования участников ансамбля.
	1. В качестве классификаторов-участников ансамбля взять деревья решений. Сформировать выборки для классификаторов-участников ансамбля

	путем случайного отбора подмножества атрибутов исходной выборки. Обучить каждого участника на своей выборке. 2. Получить от каждого участника ансамбля класс ранее не известного объекта. Получить итоговый класс ранее неизвестного объекта на основе мажоритарного голосования участников ансамбля.
	1. Присвоить одинаковые веса объектам исходной обучающей выборки. 2. Применить сэмплинг без повторений на основе весов объектов к исходной обучающей выборке и сформировать выборку для классификатора-участника ансамбля. Обучить участника на своей выборке. 3. Оценить точность участника ансамбля на объектах исходной обучающей выборки. Повысить вес неверно классифицированных объектов исходной обучающей выборки. Вычислить вес участника ансамбля на основе показанной им точности классификации. 4. Повторить пп. 2, 3 последовательно для каждого участника ансамбля. 5. Получить от каждого участника ансамбля класс ранее не известного объекта. Получить итоговый класс ранее неизвестного объекта на основе взвешенного голосования участников ансамбля.

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

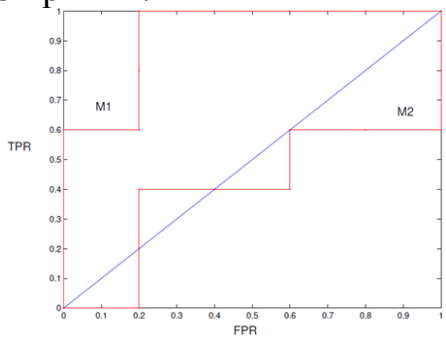
Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

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

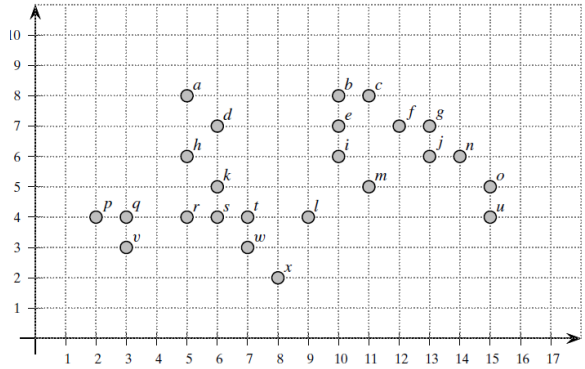
	Attribute-3
	Attribute-2
	Attribute-4
	Attribute-1

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.



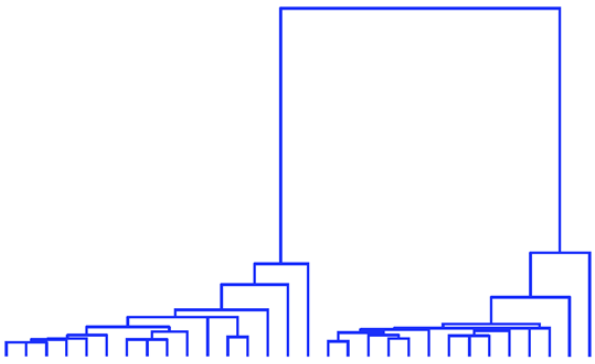
	C1={a, d, h, k, p, q, r, s, t, l, v, w, x} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума отсутствуют
	C1={a, d, h, k, p, q, r, s, t, v, w} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l, x}
	C1={p, q, v} C2={a, d, h, k, r, s, t, w, x} C3={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l}
	C1={a, 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) Укажите верный результат кластеризации объектов множества {4, 5, 8, 9, 10} посредством алгоритма k-means при k=2.

	C1={4, 5, 8, 9} C2={10}
	C1={4, 5} C2={8, 9, 10}
	C1={4, 5, 8} C2={9, 10}
	C1={4, 5, 10} C2={8, 9}

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