УТВЕРЖДАЮ

Зав. кафедрой СП ЮУрГУ

\_\_\_\_\_\_\_\_\_Л.Б. Соколинский

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_2022 г.

Фонд оценочных средств ООП «Искусственный интеллект и инженерия данных» по направлению 09.03.04 – Программная инженерия

Дисциплина «Технологии аналитической обработки информации»

| **№ КМ** | **Вид КМ** | **Наименование КМ** | **Оценочные средства** |
| --- | --- | --- | --- |
|  | Текущий контроль | Контрольный опрос 1  "Введение в дисциплину" | Пример вопросов теста:  1) 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. |   2) Укажите задачу, которая относится к сфере Data Mining.   |  |  | | --- | --- | |  | Нахождение ожидаемой прибыли от нового клиента компании на основе данных о существующих клиентах | |  | Нахождение разбиения клиентов компании на две группы: мужчины и женщины | |  | Нахождение списка клиентов компании, фактическая прибыль от которых меньше ожидаемой | |  | Нахождение общей прибыли компании, полученной ей от клиентов | |  | Нахождение разбиения клиентов компании на группы в зависимости величины полученной от них прибыли |   3) Укажите верное определение задачи кластеризации.   |  |  | | --- | --- | |  | Определение, какие из имеющихся данных могут быть использованы для принятия стратегических решений, а какие – нет. | |  | 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 объектами. | |  | Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее не известно. | |
|  | Текущий контроль | Контрольный опрос 2  "Поиск шаблонов" | Пример вопросов теста:  1) Укажите верное определение задачи поиска шаблонов.   |  |  | | --- | --- | |  | 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 объектами. | |  | Определение, какие из имеющихся данных могут быть использованы для принятия стратегических решений, а какие – нет. | |  | Определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. Множество классов, к которым может быть отнесен объект, заранее не известно. |   2) Укажите верную формулировку принципа Априори.   |  |  | | --- | --- | |  | Если набор товаров является частым, то все его подмножества также являются частыми. | |  | Если набор товаров является частым, то поддержка любого его подмножества не больше заданного порогового значения. | |  | Если набор товаров является частым, то его поддержка не меньше заданного порогового значения. | |  | Если набор товаров не является частым, то все его подмножества также не являются частыми. | |  | Если набор товаров является частым, то поддержка любого его подмножества не меньше заданного порогового значения. | |  | Если набор товаров является частым, то его поддержка не больше заданного порогового значения. |   3) Вычислите достоверность ассоциативного правила чипсы->кола для множества корзин (вписать значение, округленное до сотых).  /var/folders/71/j5hjxk7x3fg2h9tl8qp12bdr0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/baskets1.png |
|  | Текущий контроль | Контрольный опрос 3  "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 основную идею алгоритма классификации СART.   |  |  | | --- | --- | |  | Для классификации множества объектов выбирается атрибут, разбиение заданного множества по которому дает наименьшее значение индекса Джини (Gini Index). | |  | Для классификации множества объектов используется значение меток классов у заданного количества кортежей обучающей выборки, которые в смысле заданной метрики находятся близко от кортежей тестовой выборки. | |  | Для классификации множества объектов выбирается атрибут, разбиение заданного множества по которому дает наибольшее соотношение прироста информации и информации, необходимой для разбиения (Gain Ratio). | |  | Для классификации множества объектов выбирается атрибут, разбиение заданного множества по которому дает наибольший прирост информации (Information Gain). |   3) Вычислите Gini index для атрибута **Shirt Size**:  /var/folders/71/j5hjxk7x3fg2h9tl8qp12bdr0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/wFPpVZcE50I3QAAAABJRU5ErkJggg== |
|  | Текущий контроль | Контрольный опрос 4  "Кластеризация" | Пример вопросов теста:  1) Укажите задачи реальной предметной области, которые являются задачами кластеризации.   |  |  | | --- | --- | |  | Определение смысловых групп клиентов банка, взявших кредит, на основе персональных данных этих клиентов. | |  | Определение адресов электронной почты, которые часто фигурируют совместно в списке адресатов писем. | |  | Определением смысловых групп писем электронной почты на основе данных о ключевых словах этих писем. | |  | Определение возможности выдачи кредита клиенту банка на основе персональных данных этого клиента. | |  | Определение списка клиентов банка, имеющих задолженности по выплатам кредита. | |  | Определение списка корреспондентов, отправивших наибольшее количество электронных писем. | |  | Определение пакета услуг, которые часто выбираются клиентами банка совместно с взятием кредита. | |  | Определение категории письма электронной почты: "спам" или "обычная почта" – на основе данных о ключевых словах этого письма. |   2) Укажите верное определение медоида в алгоритме кластеризации k-medoids.   |  |  | | --- | --- | |  | Медоид – объект со случайными координатами, выбираемый в качестве центра кластера в процессе кластеризации. | |  | Медоид – объект исходного множества, которому в процессе кластеризации присваивается метка кластера. | |  | Медоид – объект исходного множества, выбираемый в качестве центра кластера в процессе кластеризации. | |  | Медоид – объект, имеющий средние по всем объектам исходного множества координаты, выбираемый в качестве центра кластера в процессе кластеризации. |   3) Укажите вид алгоритмов кластеризации, к которому относится алгоритм DBSCAN.   |  |  | | --- | --- | |  | Плотностные алгоритмы | |  | Разделительные алгоритмы | |  | Агломеративные алгоритмы | |  | Дивизимные алгоритмы | |  | Нечеткие алгоритмы | |
|  | Текущий контроль | Контрольный опрос 5  "Поиск аномалий" | Пример вопросов теста:  1) Укажите статистические методы и алгоритмы поиска аномалий.   |  |  | | --- | --- | |  | Метод вложенных циклов | |  | Метод решеток | |  | Правило трех сигм | |  | Изолирующий лес | |  | HOTSAX | |  | OneClass SVM | |  | Z-значения |   2) Пусть к непустому множеству объектов применен некий алгоритм поиска аномалий. Затем к полученному непустому подмножеству аномальных объектов вновь применен первоначальный алгоритм. Укажите верное утверждение.   |  |  | | --- | --- | |  | Безусловно, ни один из полученных аномальных объектов не будет вновь классифицирован как аномалия | |  | Все полученные аномальные объекты будут вновь классифицированы как аномалии, если алгоритм использует понятие плотности объектов и те же пороговые значения, что на первом шаге | |  | Безусловно, все полученные аномальные объекты будут вновь классифицированы как аномалии |   3) Укажите верное определение диссонанса (discord) временного ряда.   |  |  | | --- | --- | |  | Диссонанс – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа максимально. Ближайший сосед подпоследовательности – подпоследовательность ряда, которая имеет минимальное расстояние до нее. | |  | Диссонанс – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа минимально. Ближайший сосед подпоследовательности – подпоследовательность ряда, которая не пересекается с данной и имеет максимальное расстояние до нее. | |  | Диссонанс – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа максимально. Ближайший сосед подпоследовательности – подпоследовательность ряда, которая не пересекается с данной и имеет минимальное расстояние до нее. | |  | Диссонанс – подпоследовательность ряда, каждая точка которой является выбросом в соответствии с правилом трех сигм. | |  | Диссонанс – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа минимально. Ближайший сосед подпоследовательности – подпоследовательность ряда, которая имеет максимальное расстояние до нее. | |
|  | Текущий контроль | Практическое задание 1  «Поиск шаблонов. 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 одного из следующих алгоритмов поиска частых наборов: Apriori, FP Growth, ECLAT.  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. |
|  | Текущий контроль | Практическое задание 2  «Поиск шаблонов. Поиск ассоциативных правил» | Вопросы для подготовки к устному опросу:  1. Дайте определения следующих понятий: шаблон (ассоциативное правило), достоверность, устойчивый шаблон.  2. Объясните алгоритм поиска устойчивых шаблонов с помощью поиска частых наборов.  3. Приведите пример устойчивого, но практически бесполезного шаблона.  4. Дайте определение меры lift полезности шаблонов.  5. Дайте определения понятий максимально частого и замкнутого набора, иерархии наборов. |
|  | Текущий контроль | Практическое задание 3  «Классификация. Байесовская классификация» | Вопросы для подготовки к устному опросу:  1. Объясните постановку задачи классификации.  2. Объясните типовой процесс классификации и его этапы: обучение модели, оценка модели, применение модели.  3. Сформулируйте теорему Байеса.  4. Объясните, почему на практике, как правило, Байесовская классификация дает невысокую точность классификации.  5. Сформулируйте проблему нулевой вероятности в Байесовской классификации и ее решение. |
|  | Текущий контроль | Практическое задание 4  «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. |
|  | Текущий контроль | Практическое задание 5  «Классификация. Бэггинг» | 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. |
|  | Текущий контроль | Практическое задание 6  «Классификация. Случайный лес» | Вопросы для подготовки к устному опросу:  1. Объясните цель и идею ансамблевой классификации.  2. Объясните работу метода случайного леса.  3. Назовите преимущества и недостатки метода случайного леса.  4. Объясните работу одной из разновидностей метода случайного леса, Forest-RI.  5. Объясните работу одной из разновидностей метода случайного леса, Forest-RC. |
|  | Текущий контроль | Практическое задание 7  «Классификация. Бустинг» | Вопросы для подготовки к устному опросу:  1. Объясните цель и идею ансамблевой классификации.  2. Объясните работу метода бустинга (на примере алгоритма AdaBoost).  3. Назовите преимущества и недостатки метода бустинга.  4. Укажите ошибку обучения ансамбля классификаторов в методе бустинга.  5. Объясните, как убывает ошибка обучения ансамбля в бустинге при увеличении количества классификаторов в ансамбле. |
|  | Текущий контроль | Практическое задание 8  «Кластеризация. Разделительная 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. |
|  | Текущий контроль | Практическое задание 9  «Кластеризация. Плотностная кластеризация» | Вопросы для подготовки к устному опросу:  1. Объясните идею плотностной кластеризации и работу алгоритма DBSCAN.  2. Дайте определения основных понятий, используемых в алгоритме DBSCAN: окрестность точки, корневая точка, непосредственная достижимость, достижимость.  3. Назовите преимущества и недостатки алгоритма DBSCAN.  4. Объясните, почему алгоритм DBSCAN является нечувствительным выбросам и шумам в исходных данных.  5. Объясните, каким можно подбирать параметры и алгоритма DBSCAN. |
|  | Текущий контроль | Практическое задание 10  «Кластеризация. Иерархическая кластеризация» | Вопросы для подготовки к устному опросу:  1. Объясните идею иерархической кластеризации и работу алгоритма агломеративной кластеризации.  2. Объясните идею иерархической кластеризации и работу алгоритма дивизимной кластеризации.  3. Объясните способ построения дендрограмм.  4. Назовите преимущества и недостатки иерархической кластеризации.  5. Дайте определения следующих мер схожести, используемых в иерархической кластеризации: Single linkage, Complete linkage, Average linkage. |
|  | Текущий контроль | Практическое задание 11  «Кластеризация. 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. |
|  | Текущий контроль | Практическое задание 12  «Поиск аномалий. Поиск точечных аномалий» | Вопросы для подготовки к устному опросу:  1. Дайте определение понятия аномалии и разновидностей аномалий (точечная, коллективная, контекстная).  2. Объясните работу метода максимального правдоподобия для поиска аномалий.  3. Объясните работу метода построения гистограмм для поиска аномалий.  4. Объясните работу метода z-значимости для поиска аномалий.  5. Объясните работу правила трех сигм для поиска аномалий. |
|  | Текущий контроль | Практическое задание 13  «Поиск аномалий. Поиск коллективных аномалий» | Вопросы для подготовки к устному опросу:  1. Дайте определение понятия аномалии и разновидностей аномалий (точечная, коллективная, контекстная).  2. Объясните работу метода вложенных циклов для поиска аномалий.  3. Объясните работу метода решеток для поиска аномалий.  4. Дайте определение понятия диссонанса временного ряда.  5. Объясните принцип поиска аномалий на основе кластеризации данных. |
|  | 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 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. |

2) 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. |

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. |

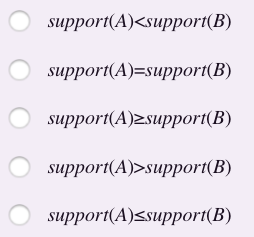
5) Установите соответствие между базовыми задачами интеллектуального анализа данных и приведенными задачами реальной предметной области.

|  |  |  |
| --- | --- | --- |
| 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 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. |

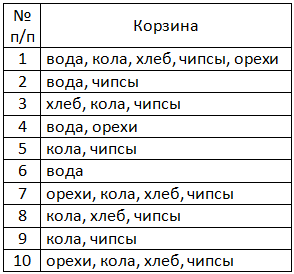
7) Пусть имеются наборы товаров A и B, причем A⊆B. Eкажите верное утверждение о поддержке наборов A и B.



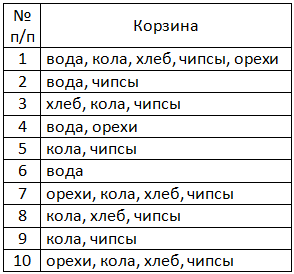
8) Пусть имеется множество частых наборов ={{a,b,c},{a,b,d},{a,c,d},{a,c,e},{b,c,d}}. Укажите множество кандидатов в частые наборы , которое будет сформировано алгоритмом Apriori.

|  |  |
| --- | --- |
|  | {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}} |

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



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



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

11) 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. |

12) Пусть с помощью метода *k* ближайших соседей выполняется бинарная классификация одномерной выборки данных (первая строка – объекты выборки, вторая строка – классы объектов):

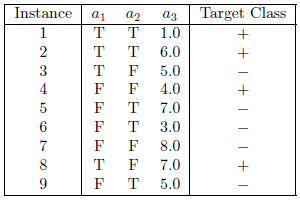
/var/folders/71/j5hjxk7x3fg2h9tl8qp12bdr0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/GBJUF8MO0XAAAAAElFTkSuQmCC

При равенстве голосов соседей больший вес имеют **+** – соседи.

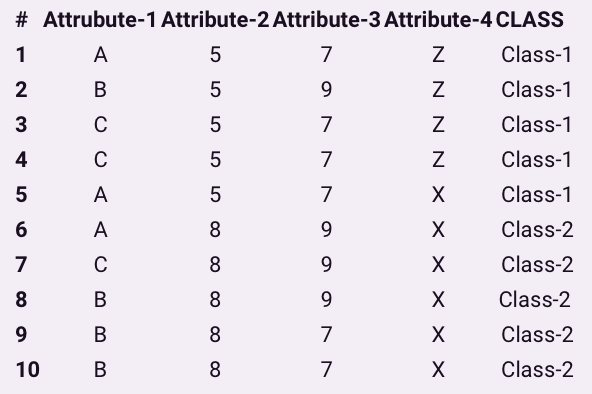
Укажите все значения параметра **k**, при которых объект **5.0** принадлежит классу **–** **(минус)**.

|  |  |
| --- | --- |
|  | 5 |
|  | 9 |
|  | 4 |
|  | 1 |
|  | 6 |
|  | 7 |
|  | 2 |
|  | 3 |

13) Используя энтропию, вычислите Info для атрибута по следующей выборке:

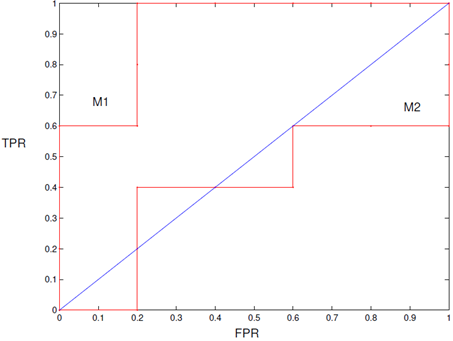


14) Какой из перечисленных атрибутов является наилучшим атрибутом разбиения, если осуществляется построение классификационной модели с двумя классами и следующей обучающей выборкой?



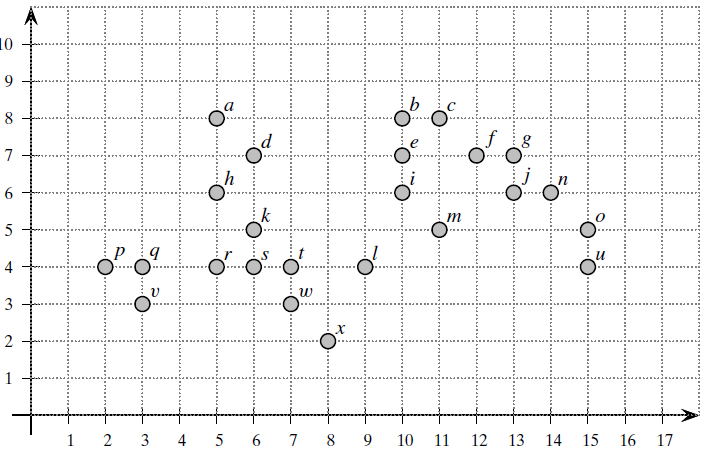
|  |  |
| --- | --- |
|  | Attribute-3 |
|  | Attribute-2 |
|  | Attribute-4 |
|  | Attribute-1 |

15) 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 |

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

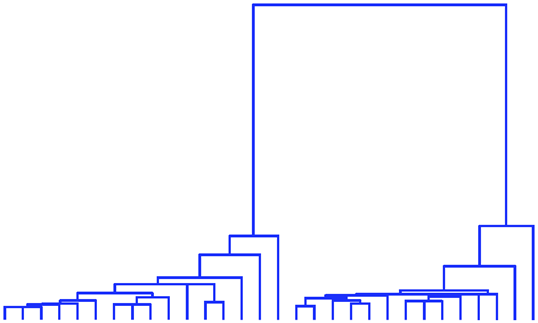


|  |  |
| --- | --- |
|  | C1={a, d, h, k, p, q, r, s, t, l, v, w, x} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума отсутствуют |
|  | C1={a, d, h, k, p, q, r, s, t, v, w} C2={b, c, e, f, g, i, j, n, m, o, u} Точки шума: {l, x} |
|  | C1={p, q, v} C2={a, d, h, k, r, s, t, w, x} C3={b, c, e, f, g, i,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

17) Укажите основную идею дивизимных алгоритмов кластеризации.

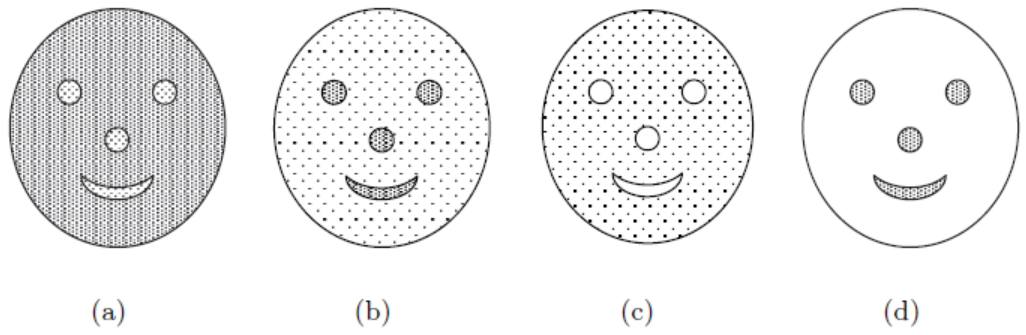
|  |  |
| --- | --- |
|  | Предполагается, что каждый исходный объект образует отдельный кластер, и затем выполняется слияние близких друг к другу объектов или кластеров до тех пор, пока не будет получен единственный кластер или не будет выполнено условие завершения слияния. |
|  | Предполагается, что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

18) 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 |

19) 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 |

20) 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 |

21) Укажите методы и алгоритмы поиска аномалий, не требующие обучения модели.

|  |  |
| --- | --- |
|  | OneClass SVM |
|  | HOTSAX |
|  | Изолирующий лес |
|  | Z-значения |
|  | Метод вложенных циклов |
|  | Метод решеток |
|  | Правило трех сигм |

22) Укажите статистическую функцию, которая является наименее чувствительной (среди остальных) к аномалиям (выбросам) в данных.

|  |  |
| --- | --- |
|  | среднее |
|  | стандартное отклонение |
|  | медиана |
|  | мода |

23) Укажите утверждение, верно отражающее правило трех сигм для поиска аномалий в одномерных данных, имеющих нормальное распределение.

|  |  |
| --- | --- |
|  | Аномалией является точка, для которой отношение модуля разности этой точки и среднего значения выборки к среднеквадратическому отклонению не больше трех |
|  | Аномалией является точка, для которой модуль разности этой точки и утроенного среднего значения выборки не больше наперед заданного аналитиком параметра |
|  | Аномалией является точка, для которой отношение модуля разности этой точки и среднего значения выборки к среднеквадратическому отклонению не меньше трех |
|  | Аномалией является точка, для которой модуль разности этой точки и утроенного среднего значения выборки не меньше наперед заданного аналитиком параметра |

24) Укажите статистические методы и алгоритмы поиска аномалий.

|  |  |
| --- | --- |
|  | Правило трех сигм |
|  | Изолирующий лес |
|  | OneClass SVM |
|  | Метод вложенных циклов |
|  | Z-значения |
|  | Метод решеток |
|  | HOTSAX |

25) Укажите алгоритм кластеризации, который в силу чувствительности к шумам и выбросам в данных даст худшие (среди перечисленных алгоритмов) результаты при его применении для поиска аномалий.

|  |  |
| --- | --- |
|  | иерархическая кластеризация |
|  | k-means |
|  | DBSCAN |
|  | Fuzzy c-Means |