**Контрольная работа по дисциплине «Методы машинного обучения»**

1. Рассматривается задача классификации на два класса: положительный и отрицательный. В ходе тестирования классификатора получены следующие результаты: полнота составляет \_\_\_\_%, общая точность составляет \_\_\_\_%. Какие значения может принимать точность?

Примечание: в ответе будет множество значений.

Решение в отдельном файле. Если и , то . Многие студенты склонны считать, что TP, FP, TN и FN – это доли, т. е. . Практически все забыли о неотрицательности всех этих величин. Никто явно не сказал о том, что частоты – это целые числа.

2. Рассматриваются объекты типа «\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_». Требуется для объектов этого типа привести следующие примеры измерений.

2а) Измерение вписывается в схему звезда.

2б) Измерение вписывается в схему снежинка, но не схему звезда.

2в) Измерение не вписывается в схему снежинка.

Примечание: в каждом пункте требуется чётко указать ответ.

При проверке надо смотреть, что измерения представимы связями со словарями. При использовании схем «звезда» и «снежинка» атрибуты объектов представляются связями со словарями. Нужны и словари, и связи, причём имя атрибута переходит на связь или композицию связей, а не словарь.

Пример измерения в схеме «звезда» содержит: одна таблица фактов (объектов), один словарь (категории), одна связь (измерение).

Пример измерения в схеме «снежинка» содержит: одна таблица фактов (объектов), два словаря (категории), две связи (от фактов к словарю и от словарю к другому словарю). При этом искомое измерение является композицией этих связей.

3. Нужен файл. Магазин собрал сведения о покупках (транзакции в файле). Были построены ассоциативные правила. Какое правило, содержащее в условии 2 элемента, имеет наибольшую поддержку?

Примечание: если наибольшую поддержку имеют несколько правил, то в ответе достаточно указать одно правило.

//Сколько транзакций всего? Сколько транзакций поддерживают правило? Поддержка в процентах? Какова достоверность правила? Одинакова ли достоверность у выбранных правил?

Продукты – варианты 1, 5. 44 транзакции, 7 элементов. Максимальная поддержка для набора из двух элементов составляет 20 транзакций (из 44, это 45,45%), имеется 1 такой набор: макароны + соусы. Максимальная поддержка для набора из трёх элементов составляет 10 транзакций (из 44, это 22,73%), имеется 1 такой набор: макароны + соусы + чай. По математическому определению годится только одно условие, по техническому – три.

1. макароны + соусы => чай (достоверность 10/20=50,00%)
2. соусы + чай => макароны (достоверность 10/13=76,92%)
3. макароны + чай => соусы (достоверность 10/13=76,92%)

Компьютеры – варианты 2, 3, 4. 42 транзакции, 7 элементов. Максимальная поддержка для набора из двух элементов составляет 19 транзакций (из 42, это 45,24%), имеется 2 таких набора: пленка + планшет; мышь + ноутбук. Максимальная поддержка для набора из трёх элементов составляет 10 транзакций (из 42, это 23,81%), имеется 1 такой набор: пленка + планшет + ноутбук. Набор мышь + ноутбук с наибольшей поддержкой 9 транзакций (из 42, это 21,43%) расширяется до одного набора мышь + ноутбук + стилус. По математическому определению годятся два условия, по техническому – три.

1. пленка + планшет => ноутбук (достоверность 10/19=52,63%)
2. ноутбук + планшет => пленка (достоверность 10/13=76,92%)
3. ноутбук + пленка => планшет (достоверность 10/13=76,92%)
4. мышь + ноутбук => стилус (достоверность 9/19=47,37%)

У студентов разные операционные системы. Сложности с установкой ПО. Сложности с чтением cp1251. Сложности осознания и обработки, когда название элементов содержит запятые. Сложности с определением поддержки правила: или поддержка условия (математический подход), или поддержка объединения условия и следствия (технический подход). Обсудить lift. Сложности с не очень чистым представлением данных – не подряд транзакция записана, повтор записей и т. п.

4. Нужен файл. Государственная избирательная комиссия зафиксировала результаты выборов по партиям и по регионам (таблица в файле). Требуется кластеризовать регионы по правилу k-средних для числа кластеров K от 1 до 12. Для каждого числа кластеров K найти максимальный радиус кластера. Построить график этой величины от K. На основании графика предположить, сколько групп регионов разумно выделить по итогам выборов.

Регионы – вариант 1, 3.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 |
| Max Rad | 21,2 | 21,2 | 26,2 | 21,2 | 21,2 | 26,2 | 26,2 | 67,7 | 64,1 | 70,1 | 69,4 |
| Av Rad | 6,36 | 6,76 | 7,1 | 7,2 | 7,48 | 8,09 | 8,49 | 9,75 | 10,4 | 11,1 | 12 |

Территориальные избирательные комиссии – вариант 2, 4, 5.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 |
| Max Rad | 39,2 | 39,2 | 41,4 | 42 | 51,5 | 53,4 | 51,6 | 51,6 | 54 | 58,6 | 62,2 |
| Av Rad | 8,54 | 8,96 | 8,96 | 9,27 | 9,88 | 10,3 | 10,8 | 11,5 | 12,8 | 13,5 | 15,7 |

У студентов разные операционные системы. Сложности с установкой ПО. Сложности с чтением cp1251. Сложности со знаком процентов после чисел. Сложности с присутствием кавычек внутри текста.

Сложности с однозначностью результатов запуска k-средних. Многие программы имеют случайную инициализацию.

Сложности с разными «улучшателями», которые в готовом ПО включены по умолчанию. Нормализация и т. п.