**Câu trắc nghiệm: nội dung toàn bộ bài giảng, đánh dấu sai bị trừ điểm**

**Câu viết.** (***Viết các nội dung cơ bản nhất càng đầy đủ càng tốt về hai trong các câu sau đây)***

1. Trình bày quy trình mô hình chuẩn công nghiệp khai phá dữ liệu CRISP-DM. Tại sao bước hiểu kinh doanh và hiểu dữ liệu là quan trọng. Nêu điểm giống và khác nhau (càng nhiều càng tốt) giữa bài toán khai phá dữ liệu và bài toán CSDL.
2. Khái niệm luật kết hợp và các độ đo liên quan. Khái niệm tập phố biến và tiếp cận chung hai bước tìm luật kết hợp trong một CSDL giao dịch.
3. Phát biểu bàitoán tìm tập phổ biến và luật mạnh. Trình bày các nội dung cơ bản của thuật toán Apriori tìm tập mục phổ biến (input-output-nội dung với các bước chính). Giải thích sơ bộ về tính đúng đắn của thuật toán.
4. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp Bayes (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm).
5. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp cây quyết định (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).
6. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp k-người láng giềng gần nhất (căn cứ, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).
7. Phát biểu bàitoán phân lớp. Trình bày thuật toán phân lớp SVM (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, nhận xét).
8. Trình bày ba bộ độ đo phổ biến (hồi tưởng-chính xác, chính xác-hệ số lỗi và hệ số tương quan Matthews) được sử dụng trong đánh giá các bộ phân lớp dữ liệu nhị phân. So sánh hai bộ độ đo đầu tiên.
9. Phát biểu bàitoán phân cụm. Trình bày thuật toán phân cụm K-mean gán cứng (nội dung cơ bản, ưu điểm- nhược điểm).
10. Phát biểu bàitoán phân cụm. Trình bày thuật toán phân cụm phân cấp HAC (độ đo tương tự cụm, nội dung cơ bản, ưu điểm- nhược điểm).
11. Phát biểu bàitoán phân cụm. Trình bày các phương pháp đánh giá thuật toán phân cụm.
12. Định nghĩa hệ thống tư vấn, các tính chất và so sánh bài toán lọc cộng tác trong hệ tư vấn với bài toán phân lớp.
13. Trình bày sơ đồ khối thủ tục lọc cộng tác hướng người dùng/hướng mục và giải thích sơ bộ.
14. Các thành phần của bộ lọc dựa trên nội dung trong hệ tư vấn.
15. Các phương thức đánh giá hiệu năng hệ thống tư vấn và các độ đo liên quan.
16. Hai chiến lược hàm mất mát điển hình (tối thiểu “bình phương lỗi” và “entropy chép”) trên ba phương án điều chỉnh tham số (điểm dữ liệu, lô nhỏ, tập dữ liệu)
17. Thuật toán quy hoạch động lan truyền ngược theo biến kích hoạt/biến tiền kích hoạt
18. Cho tập ví dụ có nhãn nhị phân trên mặt phẳng trong đó là tọa độ của điểm dữ liệu, hoặc là nhãn của điểm dữ liệu tương ứng. Hãy xác định nhãn của điểm dữ liệu mới theo thuật toán 3-NN.
19. Cho một cơ sở dữ liệu giao dịch gồm 9 giao dịch (giao dịch ::=”(”<mục>[,<mục]>]”)”

(a,b,e), (b,d), (b,c), (a,b,d), (a,c), (b,c), (a,c), (a,b,c,e), (a,b,c). Cho độ hỗ trợ tối thiểu so=0.5.

Hãy mô tả hoạt động thuật toán Apriori theo F1, C2, F2, C3, F3, … với Fi là tập các tập mục phổ biến có độ dài i, Ci là tập ứng viên độ dài i.

1. Giả sử cần phân lớp nhị phân có 2 lớp với nhãn lớp tương ứng là 1, 0. Bảng sau đây trình bày một tập dữ liệu test gồm 42 ví dụ:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

Sau khi xây dựng xong một bộ phân lớp nhị phân, người ta cho kiểm tra với dữ liệu test trên đây và nhận được kết quả tổng hợp tương ứng như sau (thứ tự vị trí dữ liệu test giữ nguyên):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

Hãy đánh giá bộ phân lớp nhị phân nói trên qua tính toán (cần viết công thức tính)

* Cặp độ đo (độ chính xác, độ hồi tưởng) và độ đo F1.
* Cặp độ đo (độ chính xác, hệ số lỗi).
* Độ đo tương quan Matthews

*Ghi chú: Số liệu trong trong bài thi đối với hai câu hỏi 19, 20 sẽ khác đi.*MỘT BÀI THI MẪU NĂM HỌC HỌC KỲ 1 NĂM HỌC 2023-2024 (Thi 08h00 ngày 27/12/2023, Thứ Tư, Phòng 309 GĐ2)

**Đề 1.** *Làm bài trong 25 phút lên tờ giấy thi gốm hai trang này*

**Câu 1. Mệnh đề đúng: điền ✓vào 🞎 để nhận được 🗹** (**đánh dấu vào mệnh đề sai sẽ bị trừ điểm**)

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

🞎

**Câu 2.** Giả sử cần phân lớp dữ liệu với miền dữ liệu có hai lớp với nhãn lớp tương ứng là 1, 0. Bảng sau đây trình bày một tập dữ liệu test gồm 10 ví dụ

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |

Một bộ phân lớp nhị phân cho kết quả phân lớp đối với 10 dữ liệu test trên đây như sau

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |

Tính giá trị các độ đo: độ chính xác, độ hồi tưởng và độ đo F1.

**Câu 3.** Trình bày cô đọng (ngắn gọn - đủ ý) về thuật toán cây quyết định: mô hình phân lớp, cây quyết định, tư tưởng chung của thuật toán cây quyết định, tối ưu hóa trong thuật toán cây quyét định.