| BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC MỞ TP.HCM**  (*Đề thi có 2 trang*) | **ĐỀ THI CUỐI KỲ HK3 Năm Học 21-22**  MÔN: KHAI PHÁ DỮ LIỆU  THỜI GIAN: 90 PHÚT  (*Không sử dụng tài liệu*) |
| --- | --- |

*Sinh viên được sử dụng máy tính tay để tính toán*

**Câu 1 (2 điểm)**

Cho dãy số gồm 12 số sau:

**4, 7, 7, 10, 0, 4, 7, 2, 3, 6, 5, 4**

Anh chị hãy thêm vào dãy số trên 1 số bất kỳ tùy ý, như vậy dãy số cuối cùng để thực hiện các yêu cầu dưới đây có tất cả là **13 số.**

Tính toán và ghi ra các kết quả sau:

* Mean, Min, Max, Median và Mod của dãy số này.
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Width với K=3.
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Frequency với K=3.

Yệu cầu trả lời: Ghi vắn tắt nhưng cần rõ ràng cách thức thực hiện và kết quả của các thuật toán rời rạc hóa được thực hiện.

**Câu 2 (2 điểm)**

Bảng dưới đây là một phần kết quả cuối năm của học sinh trường An Bình. Văn hóa được xếp loại theo 4 mức: Giỏi/Khá/Trung bình/Kém, Đạo đức (2 mức): Tốt/Kém và Thể chất (2 mức): Đạt/Không đạt

| MS | Văn Hóa={Giỏi/Khá/TB/Kém} | Đạo đức={Tốt/Kém} | Thể chất ={Đạt,Không đạt} |
| --- | --- | --- | --- |
| 1 | Giỏi | Tốt | Đạt |
| 2 | Khá | Tốt | Không đạt |
| 3 | Trung bình | Kém | Không đạt |
| 4 | Kém | Tốt | Đạt |
| 5 | Khá | Kém | Đạt |

Ta dự định sử dụng Luật Kết Hợp để phát hiện mối liên hệ kết quả đánh giá giữa các lĩnh vực văn hóa, đạo đức và thể chất của học sinh trường An Bình.

Ví dụ:

Văn Hóa Kém 🡪 Đạo Đức Kém, Thể Chất Không Đạt.

Văn Hóa Khá, Thể Chất Đạt 🡪 Đạo Đức Tốt.

Anh chị hãy:

* Mô tả vắn tắt phương án chuyển đổi dữ liệu để có thể sử dụng cho thuật toán Luật Kết Hợp..
* Viết lại bảng dữ liệu trên theo phương án anh chị đã đề xuất để có thể áp dụng vào thuật toán Luật Kết Hợp.

**Câu 3 (2 điểm)**

Khi giải quyết bài toán phân lớp, ta thường đánh giá mô hình dựa trên kết quả của ma trận sai lầm (confusion matrix) để tính toán các đại lượng cần thiết.

Bảng dưới đây là số liệu trong ma trận sai lầm của một mô hình :

| A | B | <- Được phân lớp là |
| --- | --- | --- |
| 40 | 10 | A |
| 15 | 35 | B |

Anh chị hãy tính toán và ghi ra các kết quả của những đại lượng theo anh chị là cần thiết để có thể đánh giá thuật toán này (làm tròn tới 2 số thập phân). Lưu ý: Cần ghi ra các số liệu cụ thể trong từng công thức tính toán và kết quả tương ứng.

Vì sao không chỉ tính toán kết quả cho cả bộ dữ liệu (bao gồm tất cả các lớp – class) mà còn phải tính toán các kết quả cho từng lớp (class)? Anh chị thử nêu một ví dụ minh họa và giải thích lý do.

**Câu 4 (2 điểm)**

Bảng dữ liệu dưới đây lưu lại thông tin của các khách hàng vay tín dụng của ngân hàng Lạc Việt. Các thông tin lưu giữ gồm: tuổi, tình trạng gia đình, thu nhập 1 tháng (triệu đồng), số tiền vay (triệu đồng), kết quả đánh giá (Yes: Cho vay, No: Không)

| ID | Tuổi | Gia đình | Thu nhập | Số tiền vay | Kết quả |
| --- | --- | --- | --- | --- | --- |
| 1 | 25 | Có | 20 | 120 | Yes |
| 2 | 22 | Không | 15 | 200 | No |
| 3 | 24 | Có | 18 | 180 | Yes |
| 4 | 30 | Không | 17 | 130 | Yes |
| 5 | 31 | Có | 22 | 250 | No |
| 6 | 35 | Có | 22 | 150 | Yes |
| 7 | 22 | Không | 23 | 210 | No |
| 8 | 28 | Có | 19 | 280 | No |

Anh chị hãy dùng thuật toán KNN để cho biết xem có quyết định cho khách hàng có thông tin dưới đây vay tiền hay không? Anh chị toàn quyền chọn công thức tính khoảng cách và giá trị K.

| Tuổi | Gia đình | Thu nhập | Số tiền vay | Kết quả |
| --- | --- | --- | --- | --- |
| 23 | Có | 32 | 250 | Yes/No? |

Yêu cầu: Cần trình bày vắn tắt nhưng rõ ràng các kết quả mà anh chị đã tính toán để qua đó có kết quả phân loại cuối cùng.

**Câu 5 (2 điểm)**

Bài toán phân lớp đa nhãn (Multi Label Classify): là một bài toán trong Khai Phá Dữ Liệu. Trong bài toán này, dữ liệu có thể cùng lúc được gán nhiều nhãn. Ví dụ khi phân loại các bài báo, một bài báo có thể cùng lúc liên quan nhiều chủ đề: bài báo A vừa được gán nhãn liên quan đến chủ đề Thể thao vừa gán nhãn chủ đề Thời sự, bài báo B thì chỉ được gán 1 nhãn duy nhất là chủ đề Công nghệ, bài báo C cùng lúc được gán các chủ đề Văn hóa, Thời sự và Công Nghệ,…

Với các hiểu biết về bài toán phân lớp cổ điển (Binary và Multi Class Classify tức là bài toán phân lớp nhị phân và phân lớp với dữ liệu đa lớp) anh chị thử đề xuất một vài cách để có thể áp dụng giải quyết cho bài tóan nêu trên (phân lớp đa nhãn). Lưu ý: Trình bày vắn tắt nhưng rõ ràng, có thể minh họa bằng sơ đồ.

------- HẾT--------

| BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC MỞ TP.HCM** | **ĐÁP ÁN THI CUỐI KỲ HK3 Năm Học 21-22**  MÔN: KHAI PHÁ DỮ LIỆU  THỜI GIAN: 90 PHÚT  (*Không sử dụng tài liệu*) |
| --- | --- |

**Câu 1**. (***2.0 điểm***)

**Phụ thuộc vào số mà SV thêm vào (ý đồ tránh SV sao chép bài làm)**

* Mean, Min, Max, Median và Mod của dãy số (1 điểm)
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Width với K=3. (0.5 điểm)
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Frequency với K=3. (0.5 điểm)

**Câu 2**. (***2.0 điểm***)

Xây dựng bộ dữ liệu chứa các giá trị nhị phân (1/0 T/F Y/N): VHGiỏi, VHKhá, VHTrung bình, VHKém, ĐĐTốt, ĐĐKém, TCĐạt và TCKhông đạt (1 điểm)

| MS | Văn Hóa | Đạo đức | Thể chất ={Đạt,Không đạt} |
| --- | --- | --- | --- |
| 1 | Giỏi | Tốt | Đạt |
| 2 | Khá | Tốt | Không đạt |
| 3 | Trung bình | Kém | Không đạt |
| 4 | Kém | Tốt | Đạt |
| 5 | Khá | Kém | Đạt |

Kết quả chuyển đổi (1 điểm)

| MS | VHGioi | VHKha | VHTrungBinh | VHKem | ĐĐTot | ĐĐKem | TCDat | TCKgDat |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |

**Câu 3**. (***2.0 điểm***)

| A | B | <- Được phân lớp là |
| --- | --- | --- |
| 40 | 10 | A |
| 15 | 35 | B |

1 điểm: Nếu tính toán các giá trị sau

|  | TP rate=  TP/(TP+FN) | FP rate=  FP/(FP+TN) | Precision=  TP/(TP+FP) | Recall=TP rate |
| --- | --- | --- | --- | --- |
| Class A | 40/50=0.8 | 15/(15+35)=0.3 | 40/(40+15)=0.73 | 0.8 |
| Class B | 35/50=0.7 | 10/(10+40)=0.2 | 35/(35+10)=0.78 | 0.7 |
| Toàn bộ Data | (40+35)/100=0.75 | (10+15)/(40+35)=0.33 | (40+35)/(100)=0.75 | 0.75 |

1 điểm: Nếu minh họa và giải thích như sau.

Bảng 1

| A | B | <- Được phân lớp là |
| --- | --- | --- |
| 25 | 25 | A |
| 25 | 25 | B |

Bảng 2

| A | B | <- Được phân lớp là |
| --- | --- | --- |
| 0 | 50 | A |
| 0 | 50 | B |

Nếu chỉ tính trên toàn bộ dữ liệu thì TP ở cả 2 bảng đều như nhau

Bảng 1 (25+25)/100 = 50% = Bảng 2 (0+50)/100

Nhưng khi tính toán TP cho từng lớp thì mới thấy rõ hiệu quả của từng mô hình để sử dụng khi phân lớp.

| Mô hình | TP của lớp A | TP của lớp B |
| --- | --- | --- |
| Bảng 1 | 50% | 50% |
| Bảng 2 | 0 | 100% |

**Câu 4**. (***2.0 điểm***)

* Cần thực hiện chuẩn hóa các vùng dữ liệu trước khi tính toán khoảng cách (1 điểm).
* Sử dụng giá trị K (tùy ý sinh viên) và xác định chính xác, giải thích đúng để trả lời (1 điểm).

**Câu 5**. (***2.0 điểm***)

* Xây dựng một loạt các mô hình phân lớp nhị phân (Binary Classify), mỗi mô hình cho phép nhận dạng 1 nhãn. Khi nhận dạng: Sẽ lần lượt thực hiện tuần tự các mô hình phân lớp nhị phân này và tổng hợp kết quả. (1 điểm)
* Chuyển bài toán phân lớp đa nhãn thành bài toán phân lớp đa lớp: mỗi lớp mới sẽ tương ứng với các tổ hợp nhãn xuất hiện trong dữ liệu. (1 điểm)

**LƯU Ý VỀ ĐỀ THI KẾT THÚC MÔN HỌC**

* Môn học: **Khai Phá Dữ Liệu**.
* Mã môn học: **ITEC3417**
* Số tín chỉ: 3
* Hệ đào tạo: Chính quy
* Hình thức đề thi: Tự luận
* Thời gian thi: 90 phút
* Sử dụng tài liệu: Không sử dụng tài liệu