Nội dung ôn tập

Chương 1: Ma trận, Định thức, công thức Bayes, Kiểm định, ước lượng.
Chương 2: Thuật toán thu thập dữ liệu cơ bản, xử lý giá trị thiếu, gá trị ngoại lai, co giãn, chuẩn hoá dữ liệu, giảm chiều dữ liệu
Chương 3: Cho viết code vẽ đồ thị.
Chương 4: Đánh giá mô hình (bài tập), Độ lệch, phương sai, quá vừa, dưới vừa, học máy và phân loại (lý thuyết).

Bài tập:

Chương 1:

1. Một công ty có một hệ thống máy tính có thể xử lý 1600 hoá đơn trong 1 giờ. Công ty mới nhập một hệ thống máy tính mới, hệ thống này chạy kiểm tra trong 40 giờ cho thấy số hoá đơn xử lý trung bình trong 1 giờ là 1708 với độ lệch tiêu chuẩn 215. Với mức ý nghĩa 2,5% hãy nhận định xem hệ thống mới có tốt hơn hệ thống cũ hay không?

2. Một xí nghiệp hỏi 310 khách hàng về một loại sản phẩm; kết quả có 93 người đánh giá cao sản phẩm. Với độ tin cậy $0,95$ hãy:

a) Ước lượng xác suất khách hàng ưa chuộng sản phẩmbằng khoảng tin cậy.

b) Cần hỏi bao nhiêu khách hàng để sai số của xác suất trên không vượt quá $0,03$.

3. Trọng lượng *X* (gram) của mỳ chính được đóng trên máy tự động là một biến ngẫu nhiên có phân bố chuẩn với trọng lượng theo quy định là 450(gram)/gói. Có dư luận của người tiêu dùng là các gói mỳ chính của hãng không đủ trọng lượng quy định. Để thẩm định ý kiến đó người ta lấy ngẫu nhiên 28 gói và được bảng số liệu sau

Trọng lượng(gram)	430	436	448	452	454	460
Số gói tương ứng	3	8	10	4	2	1

Với mức ý nghĩa 5%, hãy cho biết ý kiến của người tiêu dùng là đúng hay sai?

4. Định mức thời gian hoàn thành sản phẩm là 34 phút. Có ý kiến cho rằng cần giảm định mức. Để thẩm định ý kiến đó người ta theo dõi thời gian hoàn thành sản phẩm ở 62 công nhân ta thu được kết quả như sau:

Thời gian lao động X phút	32 -32,5	32,5 - 33	33 - 33,5	33,5 - 34	34 - 34,5	34,5 - 35
Số công nhân tương ứng	2	7	15	25	9	4

Với mức ý nghĩa 2,5% hãy kết luận về ý kiến trên.

5. Điều tra mức lương của 100 công nhân ở một số công ty trong năm nay thu được số liệu sau:

Mức lương (triệu đồng/6 tháng)	45,6	46,0	46,4	46,8	47,2	47,6	48,0
Số công nhân tương ứng	4	11	24	31	19	9	2

Giả thiết mức lương nói trên có phân bố chuẩn

a) Với độ tin cậy 90%, hãy ước lượng tỷ lệ công nhân có mức lương dưới 46,5 triệu đồng trong 6 tháng.

b) Năm trước mức lương trung bình mỗi công nhân là 46,5 triệu đồng/ 6 tháng. Với mức ý nghĩa 2,5% có thể cho rằng mức lương trung bình của mỗi công nhân năm nay cao hơn năm trước không?

c) Với độ tin cậy 95%, hãy ước lượng mức lương trung bình của mỗi công nhân trong 6 tháng bằng khoảng tin cậy.

6. Một nhà máy sản xuất một chi tiết của điện thoại di động có tỷ lệ sản phẩm đạt tiêu chuẩn chất lượng là 87%. Trước khi xuất xưởng người ta dùng một thiết bị kiểm tra để kết luận sản phẩm có đạt yêu cầu chất lượng hay không. Thiết bị có khả năng phát hiện đúng sản phẩm đạt tiêu chuẩn với xác suất là 0,92 và phát hiện đúng sản phẩm không đạt tiêu chuẩn với xác suất là 0,96. Tìm xác suất để 1 sản phẩm được chọn ngẫu nhiên sau khi kiểm tra:

a) Được kết luận là đạt tiêu chuẩn.

b) Được kết luận là đạt tiêu chuẩn thì lại không đạt tiêu chuẩn.

c) Được kết luận đúng với thực chất của nó.

7. Ba phân xưởng I, II, III cùng sản xuất ra một loại sản phẩm. Tỉ lệ phế phẩm do ba phân xưởng sản xuất ra tương ứng là $3\%, 1\%, 2\%$. Lấy ngẫu nhiên một sản phẩm từ một lô hàng gồm 1000 sản phẩm trong đó có 500 sản phẩm do phân xưởng I, 350 sản phẩm do phân xưởng II và 150 sản phẩm do phân xưởng III sản xuất.

a) Tìm xác suất để sản phẩm lấy được là phế phẩm.

b) Tính xác suất để phế phẩm đó là do phân xưởng I, II, III sản xuất.

Chương 2.

8 . Thu thập dữ liệu là gì? Trình bày thuật toán thu thập dữ liệu cơ bản?

9. Nêu các xử lý giá trị thiếu?

10.

a) Thế nào là giá trị ngoại lai? Nêu cách sử dụng biểu đồ hộp phát hiện giá trị ngoại lai?

b) Cho day dữ liệu: 10, 22, 25, 33, 35, 37, 40, 41, 42, 45, 50, 60. Tính Q1, Q2, Q3, IQR và cho biết dữ liệu trên có giá trị ngoại lai hay không? Vẽ minh hoạ bằng biểu đồ hộp.

11.
a) Thế nào là giá trị ngoại lai? Nêu cách sử dụng giá trị trung bình và phương sai để phát hiện giá trị ngoại lai
b) Cho dãy dữ liệu: 10, 22, 25, 33, 35, 37, 40, 41, 42, 45, 50, 60.
Tính tính giá trị trung bình và phương sai của mẫu dữ liệu, sử dụng khoảng 4s xác đinh xem dữ liệu trên có giá trị ngoại lai hay không?

12. a) Nêu phân phối chuẩn và phân phối chuẩn tắc? Công thức chuẩn hoá dữ liệu?
b) Giả sử dữ liệu sau được trích ra từ một biến ngẫu nhiên có phân phối chuẩn? Hãy xác định giá trị trung bình và độ lệch tiêu chuẩn của dãy dữ liệu và chuẩn hoá dữ liệu đó.
11, 12, 13, 15, 20, 21, 22, 24, 25, 35, 32, 37, 39.

13. Nêu công thức của phép co giãn cực đại– cực tiểu? Áp dụng để ánh xạ dữ liệu sau vào đoạn [0, 1]. Loại bỏ giá trị ngoại lai (nếu có) trước khi thực hiện pháp co giãn.
10, 22, 25, 33, 35, 37, 40, 41, 42, 45, 50, 60.

14. Giảm chiều dữ liệu là gì? Tại sao cần giảm chiều dữ liệu? Nêu thuật toán phân tích thành phần chính (PCA) để giảm chiều dữ liệu?

Chương 4.

15. a) Mục tiêu của việc đánh giá mô hình là gì? Nêu các độ đo đánh giá cho mô hình phân loại: Accuracy, Precision, Recall, F1-score?
b) Giả sử có 2000 ảnh cần phân loại bệnh nhân có bị bệnh ung thư da hay không, trong dữ liệu có 1800 ảnh là bệnh nhân bị mắc bệnh, 200 ảnh không phải là ảnh bệnh nhân bị mắc bệnh. Mộ mô hình X cho kết quả dự đoán như sau:
- Trong 1800 ảnh là bệnh nhân bị mắc bệnh, 1600 ảnh được dự đoán đúng, 200 ảnh được dự đoán sai.
- Trong 200 ảnh không phải là ảnh bệnh nhân bị mắc bệnh thì có 170 ảnh dự đoán đúng và 30 ảnh dự đoán sai.
Hãy tính các độ đo Accuracy, Precision, Recall, F1-score và nhận xét về độ chính xác của mô hình?

16. a) Nêu các độ đo đánh giá mô hình hồi quy: MSE và MAE?
B) Cho dữ liệu sau:

STT	1	2	3	4	5	6	7	8	9	10	11
Giới tính	1	2	1	2	2	1	2	2	1	1	1
Giờ ôn tập	5	10	5	3	8	5	8	5	6	11	2
Điểm thicuối kỳ	70	93	69	74	88	69	79	80	78	97	70

Điểm thi cuối kỳ được dự đoán theo giới tính (1 = nữa, 2 = nam), và giờ ôn tập theo phương trình sau: Diemthi =42.5+7.942* (gender) }+3.235*(ontap). Hãy tính các độ đo MSE và MAE cho mô hình và nhận xét?

17. Xác định từ điển và các vecto đặc trưng của các câu văn trong đoạn văn sau bằng phương pháo “bag-of-word”:

“An và Hồng chơi thân với nhau. Nhưng họ có sở thích khác nhau. An thích đi du lịch nhưng ghét mua sắm. Hồng thích đi mua sắm nhưng ghét đi du lịch.”