Chương 2

Bayes và độ chính xác

2.1 Bài tập

2.1.1 Độ chính xác và Tỉ lệ lỗi

BÀI 1: Tính độ chính xác của bảng kết quả sau:

STT	Dự đoán	Thực tế
1	Đúng	Đúng
2	Đúng	Sai
3	Đúng	Sai
4	Sai	Đúng
5	Sai	Sai
6	Đúng	Đúng
8	Sai	Sai
9	Sai	Đúng
10	Sai	Đúng
11	Đúng	Đúng

Bảng 2.1: Bảng kết quả

BÀI 2: Đánh giá độ chính xác của các giả thuyết:

- Giả thuyết 1: Thời tiết-Đẹp thì không nên hoãn trận đấu
- Giả thuyết 2: Trời mà gió to cùng với thời tiết không phải nắng chúng ta sẽ không thi đấu
- Giả thuyết 3: Dù trời có mưa nhưng nếu gió yếu sẽ không ảnh hưởng đến trận đấu. Vì vậy ta sẽ không hoãn.

Thời tiết	Sức gió	Kết quả
Mưa	Mạnh	Không
Mưa	Yếu	Không
Nắng	Bình thường	Không
Âm u	Bình thường	Có
Nắng	Mạnh	Không
Mưa	Mạnh	Không
Âm u	Yếu	Có
Nắng	Bình thường	Không
Nắng	Yếu	Có
Mưa	Yếu	Có

Bảng 2.2: Bảng dữ liệu

Giải thuyết nào tốt nhất?

2.1.2 Bayes và Naive Bayes

BÀI 1:

Một bệnh nhân bị nổi các nốt mụn trên cơ thể và đi khám, bác sĩ đang có hai hướng chẩn đoán cho bệnh nhân này: Nấm da, thủy đậu hoặc viêm da tiếp xúc . Theo dữ liệu của bệnh viện, xác suất mắc bệnh nấm da thường là 69.25%, xác suất mắc thủy đậu là 0.2% và xác suất mắc viêm da tiếp xúc là 87%. Theo kinh nghiệm bác sỹ, biểu hiện các nốt mụn này xuất hiện khi mắc nấm da thường là 92%, khi bị thủy đậu là 96.7% và viêm da tiếp xúc là 98.5%. Vậy khả năng bệnh nhân đó sẽ bị mắc bệnh gì dựa trên các thông tin dữ liệu này? (Tính cụ thể)

BÀI 2:

Cho bảng dữ liệu sau:

Thời tiết	**Sức gió**	**Độ ẩm**	**Kết quả**
Mưa	Mạnh	Cao	Không
Mưa	Yếu	Trung bình	Không
Nắng	Bình thường	Trung bình	Không
Âm u	Bình thường	Thấp	Có
Nắng	Mạnh	Thấp	Không
Mưa	Mạnh	Cao	Không
Âm u	Yếu	Cao	Có
Nắng	Bình thường	Trung bình	Không
Nắng	Yếu	Thấp	Có
Mưa	Yếu	Trung bình	Có
Âm u	Bình thường	Trung bình	Có
Nắng	Yếu	Thấp	Không
Mưa	Yếu	Thấp	Có

Bảng 2.3: Bảng dữ liệu

a. Tính xác suất lớp.

b. Tính xác suất từng thuộc tính theo lớp cụ thể. Trong trường hợp có xác suất thuộc tính bằng 0. Áp dụng phương pháp Laplace để giải quyết.

c. Cho véc-tơ $x = \{$***Thời tiết*** : Nắng, ***Sức gió*** : Bình thường, ***Độ ẩm*** : Yếu$\}$, xác suất x rơi vào lớp nào?

2.2 Thực hành

2.2.1 Đọc và xử lý dữ liệu

BÀI 1 (Học hỏi và nâng cao trình độ)

Tìm hiểu một số hàm trong package pandas:

- Đọc file excel
- Ghi file csv
- Lấy ra theo cột
- Lấy ra theo dòng

Lưu ý: Dữ liệu các bạn tự tạo và thao tác tìm hiểu.

BÀI 2

Đọc tập dữ liệu : Số lượng dòng và cột (biết, mỗi một phần tử của véc-tơ ngăn cách nhau bởi dấu ",").

BÀI 3

Đọc tập dữ liệu và xử lý tập dữ liệu, thực hiện một số yêu cầu sau:

a. Số lượng thuộc tính và dòng trong bản ghi dữ liệu.
b. Có bao nhiêu thuộc tính là kiểu định tính, định lượng?
c. Kiểm tra các vấn đề có ở trong tập dữ liệu, phương pháp xử lý đối với từng vấn đề.
d. Thực hiện phân tách đầu ra và đầu vào để chuẩn bị cho mô hình học máy. (Nhãn là cột **Test Results**)

2.2.2 Xây dựng và huấn luyện mô hình

BÀI 1

Sử dụng dữ liệu Breast Cancer Wisconsin từ thư viện **sklearn** để xây dựng mô hình Naive Bayes. Hướng dẫn lấy dữ liệu:

```
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()
X = data['data']
y = data['target']
```

Yêu cầu:

- Xây dựng bằng nhiều mô hình Bayes các phân phối khác nhau
- Đánh giá kết quả

BÀI 2

Sử dụng dữ liệu từ bài 3 (Đọc và xử lý dữ liệu), thực hiện yêu cầu:

- Xây dựng mô hình học máy bằng thuật toán Bayes ngây thơ
- Đánh giá kết quả

Gợi ý: Để có thể áp dụng phân phối Gaussian cho mô hình, cần thực hiện mã hóa các ký tự được đại diện bởi số. Ví dụ: 1 - Nam, 0 - Nữ.