**ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KHOA HỌC MÁY TÍNH**

**Bài tập thực hành:**

**Phân lớp dữ liệu**

**Thông tin nhóm:**

**Đàm Thiệu Quang 1241393**

**Nguyễn Thị Yến 1241444**

Mục lục

[Bài 1: Trả lời ngắn gọn các câu hỏi sau: 3](#_Toc368581648)

[Bài 2: Bảng sau thể hiện điểm cuối kỳ và giữa kỳ của sinh viên: 3](#_Toc368581649)

[Bài 3: Cài đặt thuật toán kNN theo yêu cầu: 4](#_Toc368581650)

# Bài 1: Trả lời ngắn gọn các câu hỏi sau:

1. Tại sao phân lớp Bayes được gọi là “naive”?

Vì Naïve Bayes xem các thuộc tính xảy ra độc lập với nhau.

1. Tại sao cần có bước tỉa nhánh (tree pruning) trong cây quyết định:

Cần có bước tiả nhánh để đơn giản tập luật, bằng cách loại bỏ những tiền đề thừa và các luật thừa.

1. So sánh:

* Ưu điểm:

|  |  |
| --- | --- |
| Eager classification | Lazy clasification |
| Dễ xây dựng | Dễ sử dụng và cài đặt |
| Phân lớp mẫu mới nhanh | Xử lý tốt với dữ liệu nhiễu |
| Dễ dàng diễn giải cho các cây kích thước nhỏ | Thích hợp hơn trong không gian tìm kiếm phức tạp |

* Nhược điểm:

|  |  |
| --- | --- |
| Eager classification | Lazy clasification |
| Không thể học tăng cường | Cần lưu tất cả các mẫu |
| Tìm kiếm trong một không gian tìm kiếm lớn nên một số lớp có thể quan trọng để phân loại các trường hợp cụ thể có thể bị bỏ qua | Cần nhiều thời gian để xác định lớp cho một mẫu mới |
|  | Phụ thuộc vào giá trị k do người dùng lựa chọn |
|  | Thuộc tính phi số |

# Bài 2: Bảng sau thể hiện điểm cuối kỳ và giữa kỳ của sinh viên:

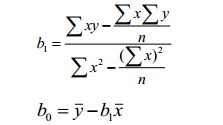
|  |  |
| --- | --- |
| x  Giữ kỳ | Y  Cuối kỳ |
| 75 | 84 |
| 50 | 62 |
| 80 | 77 |
| 74 | 79 |
| 94 | 90 |
| 86 | 75 |
| 59 | 50 |
| 84 | 78 |
| 61 | 75 |
| 33 | 45 |
| 88 | 85 |
| 81 | 90 |

1. Giữa điểm giữa kỳ (x) và điểm cuối kỳ (y) có mối quan hệ tuyến tính không?

Điểm giữa kỳ và điểm cuối kỳ không có quan hệ tuyến tính với nhau

1. Dùng phương pháp *method of least squares* để tìm phương trình dự đoán điểm cuối kỳ dựa vào điểm giữa kỳ:

Mô hình xấp xỉ: = β1\*x + β0



Ta có:

Suy ra:

β1 = 0.705885

β0 = 23.28411

= 0.705885\*x + 23.28411

1. Dự đoán điểm cuối kỳ của sinh viên có điểm giữa kỳ là : 79

Ta có: x=79

Suy ra: y = 0.705885\*79 + 23.28411= 79.049

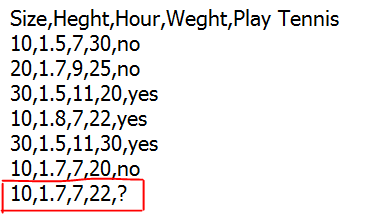
Vậy Điểm cuối kỳ của sinh viên có điểm giữa kỳ bằng 79 là 79.049

# Bài 3: Cài đặt thuật toán kNN theo yêu cầu:

1. Input:

Nhóm em tạo file input theo định dạng của file .csv thông thường, với:

* Dòng đầu tiên là danh sách tên các thuộc tính trong tập mẫu
* m dòng kế tiếp là m mẫu trong tập huấn luyện
* Dòng tiếp theo là mẫu cần phân lớp
* ví dụ: với số dòng trong tập huấn luyện là 6.



1. Thông số K:

Yêu cầu người dùng nhập vào thông số K

1. Tính khoảng cách:

Nhóm em sử dụng khoảng cách Euclide để tính khoảng các giữa các mẫu trong tập huấn luyện và mẫu cần phân lớp