ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Khai thác dữ liệu và ứng dụng
BT03: Phân lớp dữ liệu

GVLT: Thầy Lê Hoài Bắc

GVTH: Thầy Nguyễn Tiến Huy

Sv: Nguyễn Phan Mạnh Hùng 1312727
La Ngọc Thùy An 1312716

1 Câu 1

1.1 Câu a

Câu hỏi: Tại sao phân lớp Bayes (Bayesian Classification) được gọi là "naive" (ngây thơ)?
Trả lời: Phân lớp Bayes được gọi là "naive" do việc giả định về giá trị của một thuộc tính là độc lập so với giá trị của các thuộc tính khác.

1.2 Câu b

Câu hỏi: Tại sao cần có bước tỉa nhánh (tree pruning) trong cây quyết định?
Trả lời: Trước hết ta cần biết rằng dữ liệu thực tế luôn tồn tại những dữ liệu nhiễu (bất thường), do đó nếu như cây càng chi tiết và cố gắng fit mọi bộ dữ liệu thì sẽ dễ dẫn tới trường hợp overfitting. Do đó, bước tỉa nhánh nhằm loại bỏ những nhánh nhỏ lẻ (mang ít giá trị trong việc phân lớp dữ liệu) mà vẫn giữ được tính tổng quát của cây và không làm tăng sai số trên dữ liệu thực tế.

1.3 Câu c

Câu hỏi: Các phương pháp như Decision tree, Bayesian, neural network được gọi là eager classification; ngược lại các phương pháp như kNN được gọi là lazy classification. So sánh ưu nhược điểm của hai nhóm phương pháp này.
Trả lời:

Eager classification	Lazy classification
Ưu điểm	
Không gian lưu trữ nhỏ. Mang tính tổng quát. Ít bị ảnh hưởng bởi nhiễu. Truy vấn nhanh.	Không bị mất thông tin do lưu trữ toàn bộ dữ liệu. Giai đoạn train nhanh (hoặc gần như không có như trong kNN). Tính toán được các hàm phức tạp.
Nhược điểm	
Gặp khó khăn trong việc xây dựng các hàm phức tạp, và khó đưa ra các xấp xỉ tốt cục bộ của hàm mục tiêu. Thông tin không bảo toàn do thường chỉ lưu trữ hàm hypothesis.	Không gian lưu trữ lớn, do phải lưu toàn bộ dữ liệu. Dễ bị các điểm nhiễu ảnh hưởng. Truy vấn chậm, do thường phải duyệt toàn bộ tập dữ liệu.

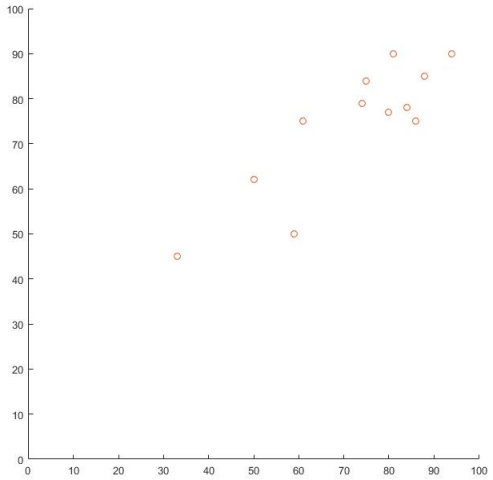
2 Câu 2

Dữ liệu

X Giữa 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

2.1 Câu a

Câu hỏi: Giữa điểm giữa kỳ (x) và điểm cuối kỳ (y) có mối quan hệ tuyến tính không?
Trả lời:



Hình 1: Biểu đồ biểu diễn các điểm dữ liệu

Dựa vào biểu đồ trên, ta nhận thấy điểm dữ liệu biểu diễn điểm giữa kì và cuối kì dường như dao động xung quanh 1 đường thẳng.
Dựa vào đó, ta đưa ra dự đoán rằng giữa điểm giữa kỳ (x) và điểm cuối kỳ (y) có mối quan hệ tuyến tính.

2.2 Câu b

Câu hỏi: Dùng phương pháp method of least squares để tìm phương trình dự đoán điểm cuối kỳ dựa vào điểm giữa kỳ.
Trả lời:

Xét phương trình:	$h(x) = \theta_0 + \theta_1 x$
Hàm chi phi dựa vào method of least squares:	$J(X) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$
n	số bộ dữ liệu, hay số cặp điểm giữa kì và cuối kì.
x_i	điểm giữa kì thứ i
y_i	điểm cuối kì tương ứng thứ i

Ta cần tìm $\theta = \begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix}$, sao cho $J(X)$ là nhỏ nhất.

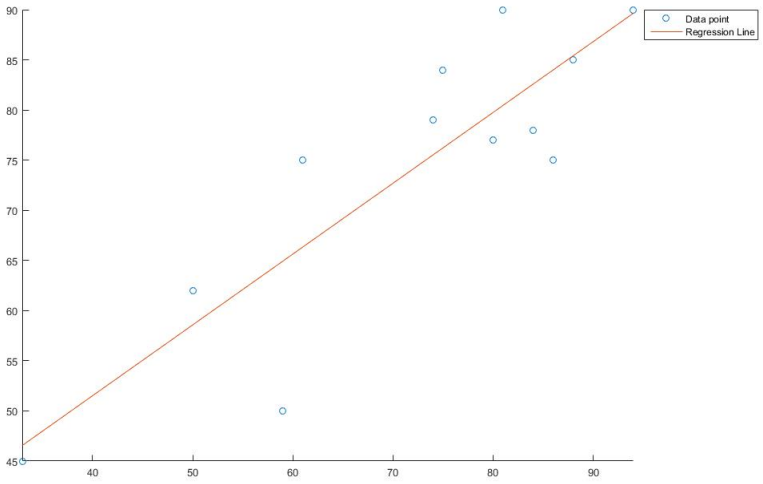
Sử dụng Normal equation, ta tìm được θ theo công thức sau:

$$\theta = (A^T A)^{-1} A^T Y \tag{1}$$

Với A = [*alpha* X], trong đó $\alpha^T = (1, 1, ..., 1) \in R^n$, và $X^T = (x_1, x_2, ..., x_n)$

Ta giải được $\theta = \begin{bmatrix}23.2841 \\ 0.7059\end{bmatrix}$.

Vậy $h(x) = 23.2841 + 0.7059x$ Dựa vào kết quả tìm được, ta có được đường thẳng sau:



Hình 2: Biểu đồ biểu diễn phương trình dự đoán điểm cuối kì theo điểm giữa kì

2.3 Câu c

Với phương trình tìm được, ta dễ dàng tính $h(79) = 79.0490$

2.4 Source

Nhóm em cũng đã viết chương trình câu 2 bằng matlab.
Source được lưu trong thư mục MethodOfLeastSquare trong Program nếu Thầy cần kiểm tra.

3 Câu 3

Sources được lưu trong thư mục KNN trong thư mục Program.
Trước khi chạy chương trình, xin Thầy đọc qua file README.txt để nắm được input's format.

4 Đánh giá công việc

Bảng phân công nhiệm vụ

Nhiệm vụ	Thành viên
Câu 1	Mạnh Hùng
Câu 2	2 thành viên cùng thảo luận, làm cá nhân và tổng hợp, so khớp kết quả.
Câu 3	Thùy An