BÀI TẬP THỰC HÀNH 3

PHÂN LỚP DỮ LIỆU

1. Trả lời ngắn gọn các câu hỏi sau:

a. Tại sao phân lớp Bayes (Bayesian classification) được gọi là “naive” (“ngây thơ”)?

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

c. 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.

2. Bảng sau thể hiện điểm cuối kỳ và giữa kỳ của sinh viên

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

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

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

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

3. Cài đặt thuật toán KNN theo yêu cầu sau:

- **Input**:
  - U là mẫu cần phân lớp.
  - T là tập huấn luyện: $T_1 = (t_{1,1}, t_{1,2}, ..., t_{1,n})$, …, $T_m = (t_{m,1}, t_{m,2}, ..., t_{m,n})$
  - Thuộc tính $t_{i,n}$ là nhãn (label) của T_i
  - m là số lượng mẫu trong tập huấn luyện
  - n là số lượng thuộc tính trong mỗi mẫu.
  - k là số lượng láng giềng gần nhất ta cần tìm

- **Output**: Lớp của mẫu U

Qui định:

- Làm bài theo nhóm. Mỗi nhóm tối đa 2 sinh viên.
- Hạn nộp: xem trên Moodle
- Bài nộp gồm file pdf/doc/docx trả lời câu hỏi lý thuyết, có đánh giá công việc từng cá nhân trong nhóm + thư mục source code.
- Đặt tất cả các nội dung được yêu cầu nộp trong thư mục có tên MSSV1_MSSV2, nén lại thành tập tin .zip hoặc .rar. Đại diện thay mặt nhóm để nộp ở link tương ứng trên Moodle.
- Sinh viên có thể viết chương trình bằng ngôn ngữ C/C++/C#.
- Các bài làm giống nhau hay chép code từ nơi khác sẽ bị 0 điểm.