

KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

Mã đề thi: **01-01**

Ngày thi: 03/04/2025

ĐỀ THI KẾT THÚC HỌC PHẦN

Tên học phần: **Phân tích dữ liệu lớn**

Thời gian làm bài: 75 phút

Loại đề thi: **Tự luận**

Câu 1 (1.5 điểm): Hãy trình bày khái niệm và quy trình hoạt động của MapReduce trong xử lý dữ liệu lớn ?.

Câu 2 (1.5 điểm): Nêu tổng quan về các đặc điểm và thách thức khi xử lý dữ liệu lớn ?.

Câu 3 (1.0 điểm): Thuật toán phân cụm K-means với khoảng cách Euclid được sử dụng để phân nhóm khách hàng dựa trên dữ liệu mua sắm (số lần mua, số lượng sản phẩm mua) trong tháng. Sau khi chạy thuật toán với $k = 3$, ta thu được các tâm cụm như sau: (20; 30), (50; 60), (90; 100). Hãy xác định cụm mà điểm dữ liệu (45; 55) thuộc về.

Câu 4 (3.0+1.0 điểm): Cho bảng dữ liệu huấn luyện về các dòng máy tính:

ID	Nhà sản xuất	Dung lượng RAM (GB)	Hiệu suất cao (yes/no)
1	Dell	8	yes
2	HP	16	yes
3	Lenovo	4	no
4	Asus	32	yes
5	Acer	8	no

- a. Dựa vào thuật toán cây ra quyết định sử dụng Information Gain, hãy xây dựng cây với độ cao bằng 1 (chỉ có node gốc và lá không có node trung gian) để phân loại máy tính có hiệu suất cao hay không?.
- b. Tính độ chính xác của cây ra quyết định trên khi phân loại dữ liệu kiểm tra sau:

ID	Nhà sản xuất	Dung lượng RAM (GB)	Hiệu suất cao (yes/no)
1	HP	16	yes
2	Lenovo	4	no
3	Acer	8	no

Câu 5 (1.0+1.0 điểm):

- a. Hãy giới thiệu tóm tắt về các phương pháp trực quan hóa phân bố (phối) của dữ liệu.
- b. Cho dữ liệu kích thước bộ nhớ RAM (GB) của các mẫu máy tính theo thứ tự tăng dần: 4, 4, 6, 8, 8, 8, 12, 16, 16, 32, 32, 32, 64, 128, 128, 128. Hãy vẽ boxplot của dữ liệu này ?.

..... HẾT

Ghi chú: + Cán bộ coi thi không phải giải thích gì thêm

+ **Sinh viên không được sử dụng tài liệu**

Cán bộ ra đề
Nguyễn Hoàng Huy

Cán bộ duyệt đề
Trần Thị Thu Huyền

KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

Mã đề thi: **01-02**

Ngày thi: 03/04/2025

ĐỀ THI KẾT THÚC HỌC PHẦN
Tên học phần: **Phân tích dữ liệu lớn**
Thời gian làm bài: 75 phút
Loại đề thi: **Tự luận**

Câu 1 (1.5 điểm): Hãy nêu và giải thích một số công cụ phổ biến trong hệ sinh thái xử lý dữ liệu lớn Hadoop.

Câu 2 (1.5 điểm): Trình bày tổng quan các phương pháp thu thập và tiền xử lý dữ liệu trong các hệ thống phân tích dữ liệu lớn ?.

Câu 3 (1.0 điểm): Thuật toán phân cụm K-means với khoảng cách Euclid được sử dụng để phân nhóm khách hàng dựa trên dữ liệu mua sắm (số lần mua, tổng giá trị mua) trong tháng. Sau khi chạy thuật toán với $k = 3$, ta thu được các tâm cụm như sau: (20; 305), (50; 600); (90; 1009). Hãy xác định cụm mà điểm dữ liệu (45; 559) thuộc về.

Câu 4 (3.0+1.0 điểm): Cho bảng dữ liệu huấn luyện về các dòng điện thoại:

ID	Thương hiệu	Dung lượng pin (mAh)	Hiệu suất cao (yes/no)
1	Apple	3000	no
2	Samsung	4500	yes
3	Xiaomi	5000	yes
4	Oppo	3500	no
5	Realme	6000	yes

- Dựa vào thuật toán cây ra quyết định sử dụng Information Gain, hãy xây dựng cây với độ cao bằng 1 (chỉ có node gốc và lá không có node trung gian) để phân loại điện thoại có hiệu suất cao hay không?.
- Tính độ chính xác của cây ra quyết định trên khi phân loại dữ liệu kiểm tra sau:

ID	Thương hiệu	Dung lượng pin (mAh)	Hiệu suất cao (yes/no)
1	Apple	3000	no
2	Samsung	4500	yes
3	Oppo	3500	no

Câu 5 (1.0+1.0 điểm):

- Hãy giới thiệu tóm tắt về các phương pháp trực quan hóa phân bố (phối) của dữ liệu.
- Cho dữ liệu về dung lượng pin (mAh) của các mẫu điện thoại theo thứ tự tăng dần: 2000, 2500, 3000, 3100, 3500, 4000, 4200, 4500, 5000, 5500, 6000, 6500. Hãy vẽ boxplot của dữ liệu này ?.

..... HẾT

Ghi chú: + Cán bộ coi thi không phải giải thích gì thêm

+ **Sinh viên không được sử dụng tài liệu**

Cán bộ ra đề
Nguyễn Hoàng Huy

Cán bộ duyệt đề
Trần Thị Thu Huyền