BÀI THỰC HÀNH

IT6077

PHÂN TÍCH DỮ LIỆU LỚN

BÀI 1

TIỀN XỬ LÝ DỮ LIỆU VỚI CÔNG CỤ

Dữ liệu:

- **File hotel_bookings.txt**
- **File bank-data.csv**

Phần 1: Cài đặt môi trường thực hành và thao tác với các định dạng dữ liệu.

Sinh viên tiến hành cài đặt các công cụ phục vụ cho việc thực hành.

Sinh viên thao tác trên các định dạng dữ liệu khác nhau.

Yêu cầu:

1. Chuyển đổi định dạng dữ liệu .txt sang csv
2. Cài đặt Excel add-in cho phân tích dữ liệu.
3. Cài đặt và làm quen với giao diện weka, cài đặt các gói bổ sung.
4. Cài đặt Python, Pycharm và các gói cần thiết.

Phần 2. Tóm lược dữ liệu

- Thực hiện lập bảng tóm lược dữ liệu bằng Excel

Phần 3. Làm sạch dữ liệu

- Xử lý giá trị khuyết với WEKA (bank-data.csv)
- Xử lý giá trị khuyết với Excel (bank-data.csv)

Phần 4. Trích chọn thuộc tính

- Giảm số thuộc tính của bộ dữ liệu **hotel_bookings.txt** xuống còn 5 thuộc tính bằng phương pháp PCA.

BÀI 2

TIỀN XỬ LÝ DỮ LIỆU VỚI PYTHON – PHẦN 1

Dữ liệu:

- **File hotel_bookings.csv**
- **File abalone.csv**
- **File train.csv** (Titanic - Machine Learning from Disaster)

Phần 1. Cài đặt môi trường thực hành

- Python
- Numpy
- Pandas

Viết chương trình bằng ngôn ngữ Python để:

Phần 2. Làm sạch dữ liệu (file DataClean.py)

- Đọc file *hotel_bookings.csv* lên data frame của pandas
- Xóa bỏ cột *Company* và một số cột kiểu phi số khác (tùy ý).
- Xóa bỏ các dòng có khuyết dữ liệu
- Lưu kết quả ra file csv mới với tên *hotel_bookings_ok.csv*
- Đọc file *abalone.csv*, thực hiện điền khuyết cho cột "*Height*" bằng `interpolate()` hoặc bằng giá trị trung bình của cột.

Phần 3. Tóm lược dữ liệu (file DataSum.py)

- Đọc file *hotel_bookings_ok.csv*
- In ra bảng tóm lược dữ liệu cho các cột kiểu số, bao gồm: count, mean, std, min, Q1, Q2, Q3, max,
- Vẽ biểu đồ histogram cho cột "*stays_in_week_nights*".

Phần 4. Thực hiện các yêu cầu của phần 2 và phần 3 trên các bộ dữ liệu còn lại.

BÀI 3

TIỀN XỬ LÝ DỮ LIỆU VỚI PYTHON - PHẦN 2

Dữ liệu:

- **File abalone.csv**
- **Dữ liệu breast_cancer của sklearn.**

Phần 1. Chuẩn hóa dữ liệu (file DataNorm.py)

- Viết hàm: đọc file *abanlone.csv* lên data frame của pandas.
- Viết hàm chuẩn hóa một cột dữ liệu (kiểu số thực) theo min-max normalization.
- Viết hàm chuẩn hóa một cột dữ liệu (kiểu số thực) theo unit norm normalization.
- Viết hàm chuẩn hóa một cột dữ liệu (kiểu số thực) theo z-score normalization.
- Sử dụng các hàm đó để chuẩn hóa dữ liệu (theo từng cách) và lưu dữ liệu kết quả ra tệp .csv.

Phần 2. Trích chọn thuộc tính (file FeatureExt.py)

- Viết hàm cho phép trích rút thuộc tính bằng PCA từ file dữ liệu breast_cancer của sklearn (số thuộc tính trích rút tùy ý, tức là tham số của hàm).
- Thực hiện trích rút thuộc tính từ file dữ liệu breast_cancer và lưu kết quả vào tệp .csv.

BÀI 4

PHÂN TÍCH MÔ TẢ

Dữ liệu:

- **File mtcars.csv**

Phần 1. Thống kê mô tả (file Descriptive.py)

- Đọc toàn bộ nội dung file mtcars.csv lên data frame của pandas.
- Tính và in ra trung bình cộng theo hàng, theo cột của bộ dữ liệu.
- Tính và in ra median, mode của từng cột dữ liệu.
- Tính Max, Min, Q1, Q3 cho từng cột dữ liệu.
- Tính phương sai và độ lệch chuẩn cho từng thuộc tính.
- Tạo bảng thống kê mô tả cho từng thuộc tính và lưu vào tệp dạng text:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
count											
min											
max											
median											
mode											
Q1											
Q2											
Q3											
IQR											
variance											
stdev											

Phần 2. Vẽ các loại biểu đồ (file DescriptiveChat.py)

- Vẽ biểu đồ histogram cho các thuộc tính kiểu nguyên của bộ dữ liệu.
- Vẽ biểu đồ hộp BoxPlot cho từng thuộc tính.

BÀI 5

PHÂN TÍCH DỰ BÁO – PHẦN 1

Dữ liệu:

- **File weatherHistory.csv**

Phần 1. Xử lý dữ liệu (file RegData1.py)

- Đọc toàn bộ nội dung tệp weatherHistory.csv lên data frame của pandas.
- Vẽ đồ thị dạng scatter với hai trục là dữ liệu của hai cột Humidity và Temperature.

Phần 2. Mô hình hồi quy tuyến tính (file LinearReg1.py)

- Tạo mô hình hồi quy tuyến tính đơn biến, với biến phụ thuộc là Temperature và biến độc lập là Humidity để dự báo nhiệt độ khi biết độ ẩm.
- Nhập vào một giá trị của độ ẩm Humidity, in ra nhiệt độ dự báo tương ứng bằng mô hình hồi quy tuyến tính vừa lập.

Phần 3. Mô hình hồi quy tuyến tính – ôn tập (file LinearReg2.py)

- Tạo mô hình hồi quy tuyến tính đơn biến, với biến phụ thuộc là Visibility và biến độc lập là Wind speed để dự báo về tầm nhìn khi biết tốc độ gió.
- Nhập vào một giá trị của tốc độ gió (Wind speed), in ra dự báo về tầm nhìn tương ứng bằng mô hình hồi quy tuyến tính vừa lập.

BÀI 6

PHÂN TÍCH DỰ BÁO – PHẦN 2

Dữ liệu:

- File Summary of Weather.csv
- File Candy-data.csv

Phần 1. Xử lý dữ liệu (file RegData2.py)

- Đọc toàn bộ nội dung tệp "Summary of Weather.csv" lên data frame của pandas.
- Vẽ đồ thị dạng scatter với hai trục là dữ liệu của hai cột MinTemp và MaxTemp, xem xét mức độ phụ thuộc tuyến tính của MaxTemp vào MinTemp.

Phần 2. Mô hình hồi quy tuyến tính - tiếp (file LinearReg3.py)

- Tạo mô hình hồi quy tuyến tính đơn biến, với biến phụ thuộc là MaxTemp và biến độc lập là MinTemp để dự báo nhiệt độ cao nhất khi biết nhiệt độ thấp nhất.
- Nhập vào một giá trị của nhiệt độ thấp nhất MinTemp, in ra nhiệt độ cao nhất dự báo tương ứng bằng mô hình hồi quy tuyến tính vừa lập.

Phần 3. Hồi quy logistic (file LogReg.py)

- Đọc nội dung tệp candy-data.csv lên data frame của pandas.
- Xóa bỏ các cột đầu tiên "competitorname" khỏi dữ liệu.
- Tách tệp dữ liệu thành hai tệp với 70% dữ liệu train và 30% dữ liệu test.
- Xây dựng mô hình dự báo xem một loại kẹo có chocolate hay không từ các thuộc tính còn lại bằng hồi quy logistic.
- Sử dụng mô hình để dự báo cho dữ liệu test, in kết quả là độ chính xác.

BÀI 7

PHÂN TÍCH DỮ LIỆU MẠNG

Dữ liệu:

- **File KAR.net (Karate)**
- **File BOK.net (Book)**
- **File KAR.pairs (Karate)**

Phần 1. Hiển thị thông tin dữ liệu (file GraphDis.py)

- Đọc tệp dữ liệu Karate.net. In thông tin về dữ liệu đọc được ra màn hình.

Phần 2. Phân cụm dữ liệu mạng (file GraphPartition.py)

- Phân cụm dữ liệu bằng giải thuật maximize modularity, in kết quả các cụm ra màn hình, cho đồ thị KAR.net.
- Tính và in ra màn hình giá trị modularity của partition thu được.
- Phân cụm dữ liệu BOK.net bằng *community_louvain*: hiển thị đồ thị, phân cụm, in kết quả phân cụm, tính và in ra modularity của partition kết quả.

Phần 3. Phân cụm dữ liệu mạng bằng igraph và networkx (file IgraphPartition.py)

- Đọc dữ liệu Karate có sẵn trong networkx bằng `nx.karate_club_graph()`
- Hiển thị đồ thị lên màn hình.
- Phân cụm dữ liệu bằng edge betweeness theo thuật toán của Girvan Newman.
- In đồ thị đã được phân cụm ra màn hình.
- Phân cụm dữ liệu bằng thuật toán tham lam greedy của Newman, maximize modularity: `greedy_modularity_communities` và in kết quả.

BÀI 8

ÔN TẬP

Dữ liệu:

- **File banking.txt**

Yêu cầu:

- Tiền xử lý dữ liệu banking.txt sao cho có thể sẵn sàng để phân tích dữ liệu. Chuyển định dạng file về .csv.
- Tạo bảng mô tả dữ liệu cho các cột kiểu số.
- Xây dựng mô hình phân tích hồi quy, dự báo duration khi biết age. Đánh giá mô hình.
- Vẽ đồ thị mô tả sự phụ thuộc của duration vào age.