PROJECT MÔN XỬ LÝ DỮ LIỆU LỚN

Giảng viên: **Tiến sĩ Bùi Thanh Hùng**
Trưởng Lab Khoa học Phân tích dữ liệu và Trí tuệ nhân tạo
Giám đốc chương trình Hệ thống thông tin
Đại học Thủ Dầu Một
hung.buithanhcs@gmail.com
Website: **https://sites.google.com/site/hungthanhbui1980**
Nộp bài: **Class room + email (hung.buithanhcs@gmail.com) (code+dữ liệu+report), deadline: 24h trước Session 15 hai ngày**
Chấm bài: **Session 15, tại B.401**

PHẦN 1 (3 điểm):

Hãy giải các bài toán sau:

Bài 1 (1 điểm):

Một ngân hàng xếp khách hàng ra thành 2 loại: tín dụng xấu và tín dụng tốt. Dựa trên thông tin trong quá khứ, ngân hàng nhận thấy rằng 1% tín dụng tốt và xy% tín dụng xấu rút quá số tiền gửi trong 1 tháng bất kỳ. Một khách hàng mới đến mở tài khoản tại ngân hàng này. Ngân hàng xác định rằng khả năng để khách hàng trở thành tín dụng tốt là 60% + xy%

1. Trong tháng đầu tiên, khách hàng này rút quá số tiền gửi. Hỏi ngân hàng sẽ xác định lại khả năng để khách hàng trở này trở thành tín dụng tốt là bao nhiêu?
2. Đến cuối tháng thứ 2, nếu khách hàng này không rút quá số tiền gửi thì ngân hàng sẽ xác định lại khả năng khách hàng này trở thành tín dụng tốt là bao nhiêu?

Với xy: là 2 số cuối mã đề của mỗi nhóm

Bài 2 (1 điểm):

Người ta xác định cứ 3000 trẻ sơ sinh thì có xy bị mất thính lực nghiêm trọng. Đây là bệnh nguy hiểm nếu để về sau nên cần phải chẩn đoán kịp thời. Với phương pháp xét nghiệm K nào đó, kết quả là 1 số thực. Sau khi biến đổi, kết quả xét nghiệm sẽ có phân phối chuẩn Gausian với trung bình (TB) và độ lệch chuẩn (ĐLC) được cho trong bảng.

Thính lực	Xét nghiệm K	
	TB	ĐLC
Bình thường	0	2
Suy giảm	3	2

1. Xây dựng một phân lớp để chẩn đoán (với xác suất lỗi nhỏ nhất) nếu biết rằng trẻ được xét nghiệm mất thính giác.
2. Xét nghiệm cho 1 trẻ được kết quả là x.y? Trường hợp này cần chẩn đoán thế nào?

Với xy là mã đề của mỗi nhóm

Bài 3 (1 điểm):

Xét bài toán phân 2 lớp ω_1, ω_2 với đặc trưng x. Cho $p(x|\omega_1)$ và $p(x|\omega_2)$ như sau:

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\}, \forall x$$

$$p(x|\omega_2) = \begin{cases} \frac{1}{4}, & -2 < x < 2 \\ 0, & x \geq 2 \ or \ x \leq -2 \end{cases}$$

1. Tìm luật phân phối thỏa cực tiểu hóa xác suất lỗi biết rằng $P(\omega_1) = P(\omega_1) = 0.ab$
2. Tìm π_1 sao cho nếu $P(\omega_1) > \pi_1$ thì phân lớp ở câu (1) luôn quyết định ω_1 bất chấp x.
3. Có tồn tại π_2 sao cho nếu $P(\omega_2) > \pi_2$ thì phân lớp ở câu (1) luôn quyết định ω_2

Với ab là mã đề của mỗi nhóm

PHẦN 2 (6 điểm)

Project

Thực hiện các yêu cầu sau và viết báo cáo theo mẫu gửi kèm:

1. Phân tích yêu cầu bài toán: Phân tích được yêu cầu của bài toán là gì (0.5 điểm)
2. Phương pháp giải quyết: Trình bày được các phương pháp giải quyết bài toán, Giải thích lý do tại sao chọn phương pháp này, Vẽ được sơ đồ tổng quát giải quyết bài toán (1.5 điểm)
3. Hiện thực - Viết code theo phương pháp giải quyết ở trên: Trình bày được cụ thể giải thuật sử dụng để giải quyết bài toán (Lưu đồ giải thuật), các tham số sử dụng, các thư viện sử dụng, code của bài toán … (3 điểm)
4. Đánh giá kết quả đạt được: So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ so sánh giữa các phương pháp theo các độ đo ví dụ như: Accuracy, MSE, RMSE, MAP, …. (hãy lựa chọn ít nhất 2 độ đo trong các độ đo phổ biến để đánh giá bài toán trên) (0.75 điểm)
5. Hướng phát triển trong tương lai: Đưa ra được hướng phát triển trong tương lai và giải thích lý do tại sao lại đưa ra hướng phát triển đó (0.25 điểm).

DANH SÁCH ĐỀ TÀI

Đề tài 01:

51800943	Trần Hưng	Trọng
51800172	Huỳnh Hoàng	Anh

Hãy xây dựng hệ thống gợi ý môn học (Recommander System) bằng phương pháp Matrix Factorization
Tham khảo: Machine Learning cơ bản: Bài 25
Yêu cầu:

- Thu thập dữ liệu môn học của sinh viên Đại học Tôn Đức Thắng (có thể lấy giả lập)
- Xây dựng hệ thống gợi ý (cho những môn tùy chọn, trong số các môn tùy chọn nếu môn nào có điểm cao nhất, thì gợi ý môn đó để sinh viên đăng ký học)

Đề tài 02:

51703074	Nguyễn Minh	Hải
51703120	Trần Gia	Kỳ

Hãy xây dựng hệ thống gợi xem phim (Recommander System) bằng phương pháp kết hợp giữa Content –Base và Collaborative Filtering
Tham khảo: Machine Learning cơ bản: Bài 23 + 24
Yêu cầu:

- Thu thập dữ liệu xem phim tự động từ trang web xem phim bất kỳ. Có thể tham khảo bộ dữ liệu IMDB và cào dữ liệu cũng như xây dựng dữ liệu tương đương với bộ dữ liệu này.
- Xây dựng hệ thống gợi ý xem phim

Đề tài 03:

51800208	Nguyễn Hoàng	Long
51800922	Võ Quốc	Sơn

Phân lớp văn bản tiếng Việt bằng phương pháp học máy SVM.

- Hãy tự crawl dữ liệu từ trang web vnexpress.net hay vietnamenet theo các chủ đề: Có thể tham khảo dữ liệu sau để huấn luyện, nhưng vẫn phải có code cào dữ liệu tự động: https://github.com/duyvuleo/VNTC/tree/master/Data
- Sử dụng phương pháp học máy SVM để phân lớp văn bản trên

Đề tài 04:

51703070	Nguyễn Ngọc Hoàng	Gia
51703127	Lê Hữu	Luân

Hãy tìm sự tương đồng giữa 2 văn bản.

- Cho 1 văn bản tiếng Việt (Ví dụ như dạng tóm tắt (bài báo, luận văn,…).
- Hãy tìm tất cả các văn bản có độ tương đồng với văn bản trên và in ra kết quả.

Văn bản có thể gồm câu hay nhiều câu, chỉ chứa text không chứa hình ảnh, bảng biểu.

Đề tài 05:

51800930	Nguyễn Quốc	Thái
51800932	Lưu Quang	Thắng

Dữ liệu và yêu cầu đề bài Home Credit có ở đường link sau.
Hãy dự đoán khả năng trả nợ của người vay, giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

https://www.kaggle.com/c/home-credit-default-risk

Đề tài 06:

51702152	Nguyễn Thị Ý	Nhi
51703163	Hoàng Văn	Phượng

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/competitive-data-science-predict-future-sales/notebooks

Hãy dự đoán *total sales for every product and store in the next month*, giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 07:

51603101	Trương Quang	Hậu
51702190	Nguyễn Huy	Thịnh

Dữ liệu và yêu cầu đề bài có ở đường link sau.

m5-forecasting-accuracy
https://www.kaggle.com/c/m5-forecasting-accuracy

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 08:

51702088	Phạm Anh	Duy
51800950	Lê Nguyễn Minh	Tuấn

Hãy cài đặt giải thuật Mapreduce based consistent and inconsistent rule detection (MR-CIRD) algorithm theo bài báo sau:

https://www.sciencedirect.com/science/article/pii/S2314728816300460

và thực nghiệm trên bộ dữ liệu:

https://www.kaggle.com/sohier/weekly-dairy-product-prices

Đề tài 09:

51800844	Nguyễn Đoàn Công	Cẩn
51800898	Võ Hoàng	Long

Hãy cài đặt giải thuật theo bài báo sau:

https://core.ac.uk/download/pdf/82209985.pdf

và thực nghiệm trên bộ dữ liệu:

https://archive.ics.uci.edu/ml/datasets/Mushroom

Đề tài 10:

51702154	Trương Thị Huỳnh	Như
51703004	Nguyễn Huy	Cận

Dữ liệu và yêu cầu đề bài có ở đường link sau.

Covid19-global-forecasting

https://www.kaggle.com/c/covid19-global-forecasting-week-4

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 11:

51703136	Phan Công	Nam
51703211	Võ Thiện	Trung

Phân tích ý kiến người dùng:

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.aivivn.com/contests/1

Đề tài 12:

51702062	Trương Đình	Ánh
51703162	Chung Quang	Phương

Xác định quan điểm theo khía cạnh
Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://vlsp.org.vn/vlsp2018/eval/sa

Đề tài 13:

51703118	Dương Quốc Anh	Kiệt
51703067	Nguyễn Anh	Duy

Tweet-sentiment-extraction
Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://www.kaggle.com/c/tweet-sentiment-extraction

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 14:

51800535	Đặng Nguyễn Thành	Đô
51800924	Đàm Đức	Tài

Vietnamese Relation Extraction
Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://vlsp.org.vn/vlsp2020/eval/re

Đề tài 15:

51800793	Trần Bảo	Long
51800954	Mai Quốc	Việt

Nhận dạng người nổi tiếng
Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://www.aivivn.com/contests/7

Đề tài 16:

51800901	Võ Hoàng	Mẫn
51800936	Lê Văn	Tiến

TREC-COVID Information Retrieval

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/trec-covid-information-retrieval

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 17:

51703076	Nguyễn Thị Lệ	Hằng
51703138	Tôn Nữ Thúy	Ngân

Hãy cài đặt giải thuật cải tiến T-Apriori
https://aip.scitation.org/doi/pdf/10.1063/1.4977361
và thực nghiệm trên bộ dữ liệu Mushroom:
https://archive.ics.uci.edu/ml/datasets/Mushroom

Đề tài 18:

51702081	Phạm Minh	Dương
51703222	Hà Huy	Tường

Nhận dạng chữ viết tay
Bengali.AI Handwritten Grapheme Classification

Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://www.kaggle.com/c/bengaliai-cv19

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 19:

51800747	Đoàn Nguyễn Văn	Hậu
51802086	Trịnh Văn	Khoa

IEEE-CIS Fraud Detection

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/ieee-fraud-detection

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 20:

51800576	Nguyễn Hoàng	Long
51800889	Lý Thiên	Lợi

Deepfake Detection Challenge

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/deepfake-detection-challenge

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 21:

51703078	Mai Vinh	Hiển
51703097	Dương Quang	Huy

Abusive Language Detection with Graph Convolutional Networks
Hãy hiện thực bài báo sau:

https://arxiv.org/abs/1904.04073

Đề tài 22

51703131	Nguyễn Chí	Minh
51703226	Lê Xuân	Vũ

A Graph to Sequence Model for AMR to text Generation
Hãy hiện thực bài báo sau:
https://arxiv.org/pdf/1805.02473.pdf

Đề tài 23:

51703114	Nguyễn Thế Anh	Khoa
51703123	Huỳnh Nguyễn Ngọc	Linh

Text Generation from Knowledge Graphs with Graph Transformers
Hãy hiện thực bài báo sau:
https://arxiv.org/abs/1904.02342

Đề tài 24:

51703130	Mai Hoàng	Minh
51703199	Nguyễn Văn	Tinh

Link Prediction Based on Graph Neural Networks

Hãy hiện thực bài báo sau:
https://arxiv.org/abs/1802.09691

Đề tài 25:

51703148	Phạm Trung	Nhân
51703181	Võ Thành	Tâm

Cài đặt giải thuật K-Meansvà K-Medoids, phân tích trên bộ dữ liệu:
dataset transaction10k of KEEL (http://sci2s.ugr.es/keel/category.php?cat=uns)

https://www.sciencedirect.com/science/article/pii/S1877050916000971

Đề tài 26:

51703102	Tống Đức	Huy
51703110	Nguyễn Lê Minh	Khang

Clustering large datasets using K-means
Hãy hiện thực bài báo sau:
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0087-2

Đề tài 27:

517H0134	Đỗ Nguyễn Đăng	Khoa
517H0184	Nguyễn Hùng	Vỹ

Nhận dạng tên riêng tiếng Việt
Dữ liệu và yêu cầu đề bài có ở đường link sau.
https://vlsp.org.vn/vlsp2018/eval/ner

Đề tài 28:

51800175	Phạm Thanh	Bình
51801019	Nguyễn Liu Tiến	Tài

Google-quest-challenge

Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/google-quest-challenge

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.

Đề tài 29:

51703086	Trần Trung	Hiếu

Kaggle-survey-2019
Dữ liệu và yêu cầu đề bài có ở đường link sau.

https://www.kaggle.com/c/kaggle-survey-2019

Giải pháp đưa ra phải cao hơn so với kết quả cao nhất từ notebook được nhiều người follow nhất.