TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: Predict gender from voice**

Giáo viên hướng dẫn: TS. Nguyễn Huy Đức

Sinh viên/nhóm sinh viên thực hiện:

1. Lương Chung Hội, 2051060527, lớp 62THNB

2. Phạm Thanh Sơn, 2051060685, lớp 62THNB

3. Nguyễn Ngọc Ánh Dương, 2051063820, lớp 62THNB

**Hà Nội, năm 2022**

**Phần 1: Tổng quan**

***1. Giới thiệu về học máy***

* 1. **Lịch sử và vai trò của machine learning:**
* 1950 - Nhà bác học Alan Turing đã tạo ra "Turing Test (phép thử Turing)" để xác định xem liệu một máy tính có trí thông minh thực sự hay không. Để vượt qua bài kiểm tra đó, một máy tính phải có khả năng đánh lừa một con người tin rằng nó cũng là con người.
* 1952 - Arthur Samuel đã viết ra chương trình học máy (computer learning) đầu tiên. Chương trình này là trò chơi cờ đam, và hãng máy tính IBM đã cải tiến trò chơi này để nó có thể tự học và tổ chức những nước đi trong chiến lược để giành chiến thắng.
* 1957 - Frank Rosenblatt đã thiết kế mạng nơron (neural network) đầu tiên cho máy tính, trong đó mô phỏng quá trình suy nghĩ của bộ não con người.
* 1967 - Thuật toán "nearest neighbor" đã được viết, cho phép các máy tính bắt đầu sử dụng những mẫu nhận dạng (pattern recognition) rất cơ bản. Nó được sử dụng để vẽ ra lộ trình cho một người bán hàng có thể bắt đầu đi từ một thành phố ngẫu nhiên nhưng đảm bảo anh ta sẽ đi qua tất cả các thành phố khác theo một quãng đường ngắn nhất.
* 1979 - Sinh viên tại trường đại học Stanford đã phát minh ra giỏ hàng "Stanford Cart" có thể điều hướng để tránh các chướng ngại vật trong một căn phòng.
* 1981 - Gerald Dejong giới thiệu về khái niệm Explanation Based Learning (EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung để nó có thể làm theo bằng cách loại bỏ đi những dữ liệu không quan trọng.
* 1985 - Terry Sejnowski đã phát minh ra NetTalk, nó có thể học cách phát âm các từ giống như cách một đứa trẻ tập nói.
* 1990s - Machine Learning đã dịch chuyển từ cách tiếp cận hướng kiến thức (knowledge-driven) sang cách tiếp cận hướng dữ liệu (data-driven). Các nhà khoa học bắt đầu tạo ra các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra các kết luận - hay là "học" từ các kết quả đó.
* 1997 - Deep Blue của hãng IBM đã đánh bại nhà vô địch cờ vua thế giới.
* 2006 - Geoffrey Hinton đã đưa ra một thuật ngữ "deep learning" để giải thích các thuật toán mới cho phép máy tính "nhìn thấy" và phân biệt các đối tượng và văn bản trong các hình ảnh và video.
* 2010 - Microsoft Kinect có thể theo dõi 20 hành vi của con người ở một tốc độ 30 lần mỗi giây, cho phép con người tương tác với máy tính thông qua các hành động và cử chỉ.
* 2011 - Máy tính Watson của hãng IBM đã đánh bại các đối thủ là con người tại Jeopardy.
* 2011 - Google Brain đã được phát triển, và mạng deep nơron (deep neural network) của nó có thể học để phát hiện và phân loại nhiều đối tượng theo cách mà một con mèo thực hiện.
* 2012 - X Lab của Google phát triển một thuật toán machine learning có khả năng tự động duyệt qua các video trên YouTube để xác định xem video nào có chứa những con mèo.
* 2014 - Facebook phát triển DeepFace, một phần mềm thuật toán có thể nhận dạng hoặc xác minh các cá nhân dựa vào hình ảnh ở mức độ giống như con người có thể.
* 2015 - Amazon ra mắt nền tảng machine learning riêng của mình.
* 2015 - Microsoft tạo ra Distributed Machine Learning Toolkit, trong đó cho phép phân phối hiệu quả các vấn đề machine learning trên nhiều máy tính.
* 2015 - Hơn 3.000 nhà nghiên cứu AI và Robotics, được sự ủng hộ bởi những nhà khoa học nổi tiếng như Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác), đã ký vào một bức thư ngỏ để cảnh báo về sự nguy hiểm của vũ khí tự động trong việc lựa chọn và tham gia vào các mục tiêu mà không có sự can thiệp của con người.
* 2016 - Thuật toán trí tuệ nhân tạo của Google đã đánh bại nhà vô địch trò chơi Cờ Vây, được cho là trò chơi phức tạp nhất thế giới (khó hơn trò chơi cờ vua rất nhiều). Thuật toán AlphaGo được phát triển bởi Google DeepMind đã giành chiến thắng 4/5 trước nhà vô địch Cờ Vây.

**Vai trò của machine learning**

- Tự động phân loại

Khối lượng thông tin đã đăng tải lên web là rất nhiều, mỗi cá nhân đều có sở thích và lựa chọn riêng, vậy nên việc tìm được thông tin đúng với nhu cầu là rất khó.

Việc ứng dụng công nghệ học máy vào phân loại các danh mục, điều hướng đơn giản chắc chắn sẽ giúp ích cho các trang web đang phát triển, hơn nữa độc giả cũng có thể tìm kiếm tin tức 1 cách đơn giản và nhanh chóng.

- Ứng dụng trong các mạng xã hội

Facebook News Feed là một ví dụ điển hình cho việc sử dụng machine learning vào các ứng dụng ngày này. Theo đó nếu bạn thường xuyên dừng lại để đọc hoặc like bài đăng của người khác thì news feed sẽ hiển thị tin về người ấy nhiều hơn.

Đằng sau hệ thống này thì phần mềm sử dụng phân tích để thống kê và dự đoán dữ liệu người dùng, từ đó sẽ đẩy mẫu này vào news feed, nếu bạn lướt qua 1 vấn đề nào đó thì trang dữ liệu sẽ được điều chỉnh cho phù hợp.

- Nhận diện hình ảnh

Đây là một trong những ví dụ về machine learning và AI phổ biến nhất. Có thể nói đây là một cách tiếp cận để xác định và phát hiện đặc trưng của 1 đối tượng trong hình ảnh kỹ thuật số. Kỹ thuật này còn có thể sử dụng để phân tích sâu hơn như nhận dạng khuôn mặt, nhận dạng ký tự quang học,...

- Chăm sóc sức khỏe

Cảm biến và các thiết bị đeo đã tạo ra được 1 lượng lớn các dữ liệu về sức khỏe. Các chương trình máy học có thể phân tích những thông tin này và hỗ trợ bác sĩ chẩn đoán và điều trị.

- Tài chính

Công nghệ máy học giúp các nhà đầu từ xác định cơ hội bằng cách phân tích thị trường chứng khoán. đánh giá quỹ phòng hộ hoặc điều chỉnh danh mục tài chính. Công nghệ machine learning cũng giúp xác định các khách hàng vay nợ có rủi ro cao, giảm bớt các trường hợp có hành vi lừa đảo.

* 1. **Ưu điểm và hạn chế của Học có giám sát và không giám sát**

**- Học máy có giám sát**

* Ưu điểm: Tồn tại khái niệm đầu ra trong suất quá trình học tập .

Thực hiện các nhiệm vụ phân loại và hồi quy

Cho phép ước tính hoặc ánh xạ kết quả tới một mẫu mới.

* Nhược điểm: Yêu cầu một tập dữ liệu được gắn nhãn.

Đòi hỏi một quá trình rèn luyện.

- **Học máy không giám sát**

* Ưu điểm: Nó không yêu cầu một dữ liệu đâò tạo phải đưọc gắn nhãn. Việc dán nhãn tự động cho tập dữ liệu đào tạo giúp tiết kiệm thời gian phân loại thủ công. Nhiệm vụ phân loại nhanh chóng

Nhược điểm: Không cho phép ước tính hoặc ánh xạ kết quả tới một mẫu mới. Các kết quả khác nhâu đáng kể khi có sự xuất hiện của ngoại lai. Nó chỉ thực hiện các nhiệm vụ phân loại.

***2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn***

* 1. **Phương pháp Cart (Classification and regression tree (cây phân loại và hồi quy))**

***2.1.1 Khái niệm***

CART hay Classification and regression tree (cây phân loại và hồi quy) là một thuật toán cây quyết định, được giới thiệu bởi Leo Breiman. Nó có thể giải quyết cả hai vấn đề phân loại và hồi quy.

***2.1.2 Cách hoạt động của CART***

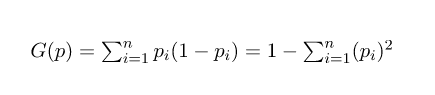
CART thường sử dụng phương pháp Gini để tạo các điểm phân chia.

*Gini impurity*

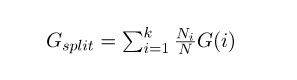
Là phương pháp hướng đến đo lường tần suất một đối tượng dữ liệu ngẫu nhiên trong tập dữ liệu ban đầu được phân loại không chính xác, trên cơ sở đối tượng dữ liệu đã nằm trong một tập con được phân ra từ tập dữ liệu ban đầu, có dán nhãn thể hiện thuộc tính chung bất kỳ của các đối tượng còn lại trong tập con này, giá trị phân loại chính là nhãn của tập con.

Gini impurity chính là chỉ số đo lường mức độ đồng nhất hay nhiễu loạn của thông tin, hay sự khác biệt về các giá trị mà mỗi điểm dữ liệu trong một tập con, hoặc một nhánh của cây quyết định. Công thức Gini có thể dùng cho cả dữ liệu rời rạc và liên tục. Nếu điểm dữ liệu thuộc về một node và có chung thuộc tính bất kỳ thì node này thể hiện sự đồng nhất lúc này gini=0gini=0 và ngược lại gini sẽ lớn.

Công thức tổng quát của Gini:



Công thức trên để tính độ vẩn đục của một node, khi có nhiều cách phân nhánh mỗi cách có thể phân ra một số node nhất định. Cho nên, lúc này có thêm công thức thứ 2 để tìm ra các phân chia tối ưu nhất:



Trong đó:

* Ni là số điểm dữ liệu có trong node của nhánh được phân
* N là số điểm dữ liệu có trong node được dùng để phân nhánh
* Hệ số Gsplit càng nhỏ thì cách phân nhánh đó càng tối ưu.
  + 1. ***Đánh giá***
* Ưu điểm :
* Cây quyết định có thể thực hiện phân loại đa lớp.
* Cung cấp hầu hết khả năng diễn giải mô hình bởi vì chúng đơn giản như là một loạt các điều kiện if-else.
* Có thể xử lý cả dữ liệu số và dữ liệu phân loại.
* Mối quan hệ phi tuyến (Nonlinear relationships) giữa các tính năng không ảnh hưởng đến hiệu suất của Cây quyết định
* Nhược điểm :
* Nhược điểm lớn nhất của Cây quyết định là vấn đề Overfitting.
* Một thay đổi nhỏ trong bộ dữ liệu có thể làm cho cấu trúc cây không ổn định có thể gây ra phương sai.
* Cây quyết định có thể bị underfit nếu dữ liệu mất cân bằng. Do đó, nên cân bằng tập dữ liệu trước khi phù hợp với Cây quyết định

**2.2. PCA - Principal Components Analysis**

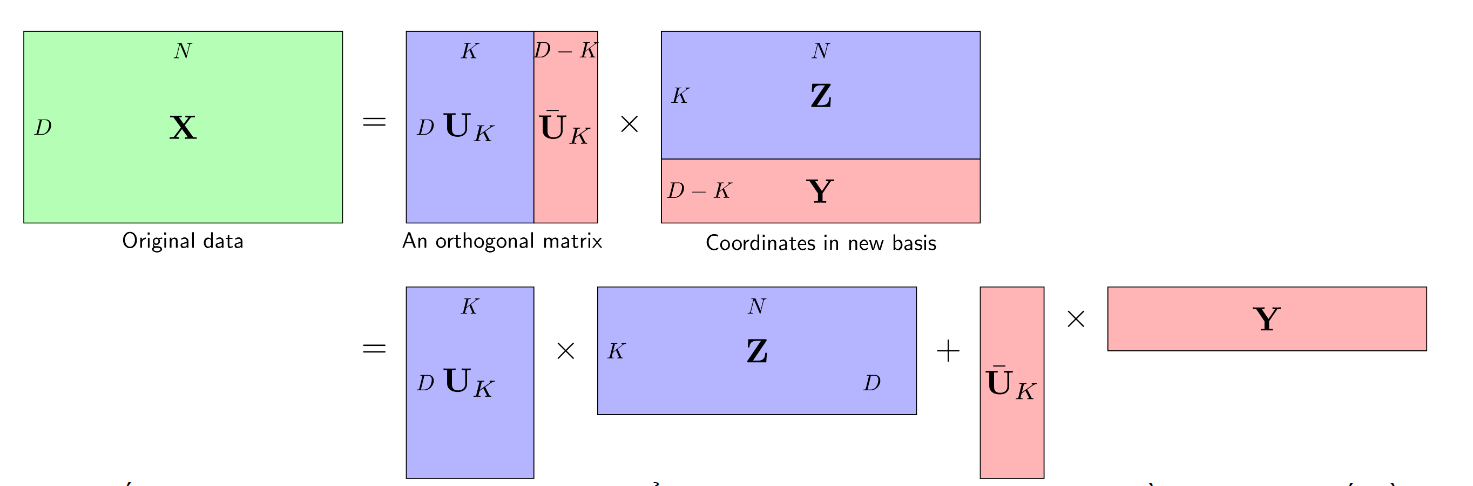
Dimensionality Reduction (giảm chiều dữ liệu) là một trong những kỹ thuật quan trọng trong Machine Learning. Các feature vectors trong các bài toán thực tế có thể có số chiều rất lớn, tới vài nghìn. Ngoài ra, số lượng các điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều cao này thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Vì vậy, giảm số chiều dữ liệu là một bước quan trọng trong nhiều bài toán. Đây cũng được coi là một phương pháp nén dữ liệu.

*Dimensionality Reduction*, nói một cách đơn giản, là việc đi tìm một hàm số sao cho:

* Input: x ∈ RD , với D rất lớn
* Output: z ∈ RK , với K < D

Cách đơn giản nhất để giảm chiều dữ liệu từ D về K (K < D) là chỉ giữ lại K phần tử *quan trọng nhất*. Tuy nhiên, việc này chắc chắn chưa phải tốt nhất vì:

* Chưa xác định được thành phần nào quan trọng hơn.
* Trong trường hợp xấu nhất, lượng thông tin mà mỗi thành phần mang là như nhau, bỏ đi thành phần nào cũng dẫn đến việc mất một lượng thông tin lớn.
* Cách giải quyết:
* Biểu diễn các vector dữ liệu ban đầu trong *một hệ cơ sở mới* mà trong hệ cơ sở mới đó, tầm quan trọng giữa các thành phần là khác nhau rõ rệt.
* Bỏ qua những thành phần ít quan trọng nhất.
  + 1. **Phương pháp *Principal Component Analysis* (PCA):**
* Tức là phân tích thành phần chính.
* Dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/ mặt đặc biệt nào đó.
* Xem xét một trường hợp đặc biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con (subspace).
* *Mục đích*: Đi tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong K thành phần đầu tiên.
* *Cách thực hiện*:
* Giả sử hệ cơ sở trực chuẩn mới là U, ta muốn giữ lại K toạ độ trong hệ cơ sở mới.



* Quan sát hình vẽ trên với cơ sở mới U = [, là một hệ trực chuẩn, với là ma trận con tạo bởi K cột đầu tiên của U.
* Với cơ sở mới này, ma trận dữ liệu có thể được viết thành:

X =

* Từ đây ta suy ra:
* Mục đích của PCA là đi tìm ma trận trực giao U sao cho phần lớn thông tin được giữ lại ở phần màu xanh và phần màu đỏ sẽ được lược bỏ và thay bằng một ma trận không phụ thuộc vào từng điểm dữ liệu.
  + 1. ***Các bước thực hiện PCA***

Fireworks in the sky

Description automatically generated with medium confidence

1. Tính vector kỳ vọng của toàn bộ dữ liệu:

Schematic

Description automatically generated with medium confidence

1. Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu:



1. Tính ma trận hiệp phương sai:

A picture containing diagram

Description automatically generated

1. Tính các trị riêng và vector riêng có norm bằng 1 của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
2. Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận Uk có các cột tạo thành một hệ trực giao. K vector này, còn được gọi là các thành phần chính, tạo thành một không gian con gần với phân bố của dữ liệu ban đầu đã chuẩn hoá.
3. Chiếu dữ liệu ban đầu đã chuẩn hoá  xuống không gian con tìm được .Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới

A picture containing text

Description automatically generated

1. Dữ liệu ban đầu có thể tính được xấp xỉ theo dữ liệu mới như sau:

Logo

Description automatically generated with medium confidence

* + 1. ***Đánh giá***
* Ưu điểm :
* Tìm được các đặc tính tiêu biểu của đối tượng cần nhận dạng mà không cần phải xác định các thành phần và mối quan hệ giữa các thành phần đó
* Giúp giảm chiều dữ liệu để dễ dàng quan sát và trực quan hóa , dữ liệu mới dễ dàng phân tích
* Thuật toán có thể thực hiện tốt với các cảnh có độ phận giải cao , do PCA sẽ thu gọn ảnh thành một ảnh có kích thước nhỏ hơn.
* PCA có thể kết hợp với nhiều phương pháp khác như mạng Noron , SVM … để mang lại hiệu quả nhận dạng cao hơn .
* Nhược điểm :
* Có thể xảy ra mất thông tin .
* Nhạy cảm với nhiễu .
* PCA phân loại theo chiều lớn nhất của tập vector . Tuy nhiên , chiều phân bố lớn nhất không phải lúc nào cũng mang lại hiệu quả tốt nhất cho bài toán nhận dạng

***3.Trình bày bài toán***

* **Tên bài toán:** Predict gender from voice
* **Tóm tắt công việc:** Nhận biết giới tính bằng giọng nói bằng phương pháp CART và phân tích thành phần chính (PCA).
* **Input:** Bộ dữ liệu bao gồm 3.168 mẫu giọng nói được ghi lại, được thu thập từ những người nói nam và nữ. Mỗi mẫu gồm các các đặc tính âm thanh của mỗi giọng nói được đo và đưa vào file csv
* **Ouput:** Xác định một giọng nói là nam hay nữ

**Phần 2: Thực nghiệm**

***1.Mô tả tập dữ liệu của bài toán***

- Mỗi mẫu Vector gồm 21 thuộc tính gồm:

* ***meanfreq:*** *tần số trung bình (tính bằng kHz)*
* ***sd:*** *độ lệch chuẩn của tần số*
* ***median:*** *tần số trung bình (tính bằng kHz)*
* ***Q25:*** *lượng tử đầu tiên (tính bằng kHz)*
* ***Q75:*** *lượng tử thứ ba (tính bằng kHz)*
* ***IQR:*** *dải ký tự (tính bằng kHz)*
* ***skew:*** *độ lệch*
* ***kurt:*** *kurtosis (see note in specprop description)*
* ***sp.ent:*** *entropy quang phổ*
* ***sfm:*** *độ phẳng quang phổ*
* ***mode:*** *tần số chế độ*
* ***centroid:*** *tần số trung tâm (see specprop)*
* ***meanfun:*** *trung bình của tần số cơ bản được đo trên tín hiệu âm thanh*
* ***minfun:*** *tần số cơ bản tối thiểu được đo trên tín hiệu âm thanh*
* ***maxfun:*** *tần số cơ bản tối đa được đo trên tín hiệu âm thanh*
* ***meandom:*** *trung bình của tần số chi phối được đo trên tín hiệu âm thanh*
* ***mindom:*** *tần số tối thiểu được đo trên tín hiệu âm thanh*
* ***maxdom:*** *tối đa của tần số ưu thế được đo trên tín hiệu âm thanh*
* ***dfrange:*** *dải tần số chính được đo trên tín hiệu âm thanh*
* ***modindx:*** *chỉ số điều chế. Được tính bằng hiệu số tuyệt đối tích lũy giữa các phép đo lân cận của các tần số cơ bản chia cho dải tần số*

- Mô tả ma trận dữ liệu (X): 20 cột và 3168 hàng

- Mô tả ma trận nhãn lớp (Y): 1 cột và 3168 hàng

- Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình: 2117 vector, 30% dùng để kiểm tra sự phù hợp của mô hình: 951 vector.

***2. Mô tả cách giải bài toán bằng phương pháp học máy***

**-** Đọc dữ liệu từ tập dữ liệu : voice.csv

- Dùng phương pháp Phân tích thành phần chính (PCA) để lựa chọn các tập thuộc tính tốt nhất cho bài toán.

- Từ tập training data và test data ban đầu , sử dụng các thành phần chính tốt nhất đã chọn để tạo ra tập training data và test data mới theo tỉ lệ 7:3 (70% để huấn luyện ; 30% để kiểm tra ).

- Dùng phương pháp học máy Cây phân lớp và hồi quy ( CART) để xây dựng mô hình và đánh giá mô hình trên tập dữ liệu mới.

- Tính độ tin cậy của mô hình dự đoán (Accuracy , Precison , Recall , F1 – score)

- Cho người dung nhập vào dữ liệu mới để mô hình dự đoán kết quả của dữ liệu vừa nhập.

- In ra kết quả dự đoán

***3. Đánh giá mô hình***

* **Kiểm tra tập dữ liệu có bị thiếu giá trị hay không:**

Chart

Description automatically generated with low confidence

* Dữ liệu không bị thiếu
* **Kiểm tra tập dữ liệu dán nhãn :**

Chart, bar chart, treemap chart

Description automatically generated

* Nhãn 0 có 1584 dữ liệu
* Nhãn 1 có 1584 dữ liệu

**Kết quả phương pháp học máy**

* + Tỷ lệ dự đoán đúng: 88% - 92%
  + Precision là tỉ lệ số điểm true positive trong số những điểm được phân loại positive (TP + FP) : 87% - 94% => Precision càng cao thì độ chính xác của các điểm tìm được càng cao
  + Recall là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP+FN) : 89% - 90% => Recall càng cao thì tỉ lệ bỏ sót cấc điểm thực sự positive càng thấp
  + Trung bình điều hòa (F1-score): 88 =>92%

**Kết luận**

Qua các kiến thức được học về Machine Learning và tìm hiểu thêm em đã thực hiện thành công bài toán “Nhận biết giới tính bằng giọng nói.” với mô hình bằng phương pháp CART và phân tích thành phần chính (PCA)

Chúng em nhận thấy với bài toán mô hình dự đoán đoán bằng phương pháp CART cho dữ liệu

**Tài liệu tham khảo**

* Slide bài giảng Nguyễn Huy Đức
* Machine Learning cơ bản : <https://machinelearningcoban.com/>