|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**  **KHOA CÔNG NGHỆ THÔNG TIN**  See the source image  **BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**  **HỌC MÁY NÂNG CAO**  **ĐỀ TÀI:**  **PHÂN LOẠI EMAIL RÁC BẰNG PHƯƠNG PHÁP GIẢM CHIỀU LDA**   |  |  | | --- | --- | | **Sinh viên thực hiện** | **: BÙI QUANG TRUNG** | | **Mã sinh viên** | **: 18810310157** | | **Giảng viên hướng dẫn** | **: PHẠM ĐỨC HỒNG** | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | **Lớp** | **: D13CNPM2** | | **Khóa** | **: 2018-2023** |   **Hà Nội, tháng 12 năm 2023** |

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên sinh viên** | **Điểm** | **Chữ ký** |
| 1 | Bùi Quang Trung  18810310157 |  |  |
|  |  |  |  |
|  |  |  |  |

**Phiếu chấm điểm của giảng viên**

|  |  |  |
| --- | --- | --- |
| **Giảng viên chấm điểm** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc152679463)

[LỜI MỞ ĐẦU 2](#_Toc152679464)

[CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY 3](#_Toc152679465)

[1.1. Tổng quan 3](#_Toc152679466)

[1.2. Các thuật toán học máy 3](#_Toc152679467)

[1.2.1. Phân loại dựa trên phương thức học 3](#_Toc152679468)

[1.2.2. Phân nhóm dựa trên chức năng 4](#_Toc152679469)

[1.3. Các ứng dụng của học máy 6](#_Toc152679470)

[1.4. Một số bài toán ứng dụng của học máy 6](#_Toc152679471)

[CHƯƠNG 2: TỔNG QUAN VỀ LDA và SVM 8](#_Toc152679472)

[2.1. Tổng quan về LDA 8](#_Toc152679473)

[2.1.1. Giới thiệu 8](#_Toc152679474)

[2.1.2. Các bước của thuật toán LDA 9](#_Toc152679475)

[2.1.3. Ưu nhược điểm của LDA 10](#_Toc152679476)

[2.2. Thuật toán SVM 11](#_Toc152679477)

[2.2.1. Giới thiệu 11](#_Toc152679478)

[2.2.2. Phương thức làm việc của SVM 12](#_Toc152679479)

[CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN VÀO BÀI TOÁN 16](#_Toc152679480)

[3.1. Mô tả bài toán 16](#_Toc152679481)

[3.2. Dữ liệu thực nghiệm 16](#_Toc152679482)

[3.3. Môi trường thực nghiệm 17](#_Toc152679483)

[3.4. Kết quả thực nghiệm 17](#_Toc152679484)

[3.5. Đánh giá kết quả thực nghiệm 17](#_Toc152679485)

[KẾT LUẬN 19](#_Toc152679486)

[DANH MỤC TÀI LIỆU THAM KHẢO 20](#_Toc152679487)

# LỜI CẢM ƠN

Thực tế thì không có sự thành công nào mà không gắn liền với học tập và thực hành. Kèm theo đó chính là sự hỗ trợ, sự giúp đỡ từ giảng viên hướng dẫn và sự tìm tòi, học hỏi của bản thân. Trong suốt quá trình học tập ở giảng đường Đại học đã đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của thầy cô, gia đình và bạn bè.

Với lòng biết ơn sâu sắc nhất, em xin gửi đến thầy cô ở Khoa Công Nghệ Thông Tin - trường Đại Học Điện Lực đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Và đặc biệt, trong kỳ này, em được tiếp cận với môn học rất hữu ích đối với sinh viên ngành Công Nghệ Thông Tin. Đó là môn: “**Học máy nâng cao**”.

Chúng em xin chân thành cảm ơn thầy Phạm Đức Hồng đã tận tâm hướng dẫn chúng em qua từng buổi học trên lớp cũng như những buổi nói chuyện, thảo luận về môn học. Trong thời gian được học tập và thực hành dưới sự hướng dẫn của thầy, em không những thu được rất nhiều kiến thức bổ ích, mà còn được truyền sự say mê và thích thú đối với bộ môn “**Học máy nâng cao**”. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì chúng em nghĩ báo cáo này rất khó có thể hoàn thành được.

Mặc dù đã rất cố gắng hoàn thiện báo cáo với tất cả sự nỗ lực. Tuy nhiên, do thời gian có hạn mà đây lại là bước đầu tiên đi vào thực tế, và vốn kiến thức còn hạn chế, nhiều bỡ ngỡ, nên báo cáo về dự án “**Phân loại email rác bằng phương pháp giảm chiều LDA**” chắc chắn sẽ không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để báo cáo này được hoàn thiện hơn.

Chúng em xin trân trọng cảm ơn quý thầy cô giáo!

# LỜI MỞ ĐẦU

**Cùng với sự phát triển của mạng máy tính và thư điện tử, thì số lượng thư rác cũng tăng lên nhanh chóng, thậm chí vượt qua sự phát triển của công nghệ và dịch vụ chống thư rác. Vì vậy, việc nghiên cứu về bộ lọc thư rác tiên tiến và hiệu quả là điều cần thiết, không chỉ từ phía người dùng mà từ cả những nhà cung cấp.**

Phát tán thư rác là hành vi gửi thư điện tử mà người nhận không mong muốn, thường với nội dung quảng cáo, được gửi hàng loạt với số lượng lớn tới một tập hợp người nhận không phân biệt. Kẻ phát tán thư rác có thể lấy địa chỉ thư điện tử từ các trang web, phòng chat, tập dữ liệu cá nhân bị rò rỉ,…

Thư rác gây ra nhiều phiền toái và thiệt hại, bao gồm, ngăn người dùng sử dụng tối ưu thời gian, dung lượng lưu trữ và băng thông mạng. Một số lượng lớn thư rác truyền trong mạng máy tính có thể phá hủy không gian nhớ của máy chủ thư điện tử, băng thông đường truyền, tài nguyên tính toán và thời gian sử dụng của thiết bị người dùng.

Bản báo cáo này chúng em xin được trình bày về đề tài: “**Phân loại email rác bằng phương pháp giảm chiều LDA**”.

# CHƯƠNG 1: TỔNG QUAN VỀ HỌC MÁY

## Tổng quan

Học máy (Machine Learning) là một ngành khoa học nghiên cứu các thuật toán cho phép máy tính có thể học được các khái niệm (concept). Có 2 phương pháp học máy chính.

* Phương pháp quy nạp: Máy học / phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụng được nguồn dữ liệu rất nhiều và sẵn có.
* Phương pháp suy diễn: Máy học / phân biệt các khái niệm dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ máy tính.

Hiện nay, các thuật toán học máy đều dựa trên 2 phương pháp này.

Các ngành khoa học liên quan:

* Lý thuyết thống kê: các kết quả trong xác suất thống kê là tiền đề cho rất nhiều phương pháp học máy. Đặc biệt, lý thuyết thống kê cho phép ước lượng sai số của các phương pháp học máy.
* Các phương pháp tính: các thuật toán học máy thường sử dụng các tính toán số thực / số nguyên trên dữ liệu rất lớn. Trong đó , các bài toán như: tối ưu có / không ràng buộc, giải phương trình tuyến tính vv ... được sử dụng rất phổ biến.
* Khoa học máy tinh: là cơ sở đánh giá các thuật toán, đồng thời đánh giá thời gian chạy, bộ nhớ của các thuật toán học máy.

## Các thuật toán học máy

### Phân loại dựa trên phương thức học

Theo phương thức học, các thuật toán Machine Learning thường được chia làm 4 nhóm: Supervised learning, Unsupervised learning, Semi-supervised learning và Reinforcement learning. Có một số cách phân nhóm không có Semi-supervised learning hoặc Reinforcement learning.

* ***Supervised Learning (Học có giám sát):***

Supervied Learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Thuật toán này được chia làm 2 loại: Classification(Phân loại) và Regression(Hồi quy).

* ***Unsupervised Learning (Học không giám sát):***

Được chia làm 2 loại: Clustering(Phân nhóm) và Association.

* ***Semi-Supervised Learning (Học bán giám sát):***

Các bài toán khi chúng ta có một lượng lớn dữ liệu XX nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên.

* ***Reinforcement Learning (Học củng cố):***

Reinforcement Learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

### Phân nhóm dựa trên chức năng

* Thuật toán Regression:
  + Linear Regression
  + Logistic Regression
  + Stepwise Regression
* Thuật toán Classification
  + Linear Classifier
  + SVM(Support Vector Machine)
  + Kernel SVM
  + SRC (Sparse Representation-based classification)
* Thuật toán dựa trên Instance
  + K-láng giềng gần nhất(KNN)
  + Learning Vector Quantization(LVQ)
* Thuật toán Regularization
  + Ridge Regression
  + Least Absolute Shrinkage and Selection Operator (LASSO)
  + Least-Angle Regression (LARS)
* Thuật toán Bayesian
  + Naive Bayes
  + Gaussian Naive Bayes
* Thuật toán Clustering
  + K-mean
  + K-median
  + Exceptation Maximization (EM)
* Thuật toán mạng nơ ron
  + Perceptron
  + Softmax Regression
  + Multi-Layer Regression
  + Back-Propagation
* Thuật toán Dimensionality Reduction
  + Principal Component Analysis (PCA)
  + Linear Discriminant Analysis (LDA)
* Thuật toán Ensemble
  + Boosting
  + AdaBoost
  + Random Forest

## Các ứng dụng của học máy

Ứng dụng: Học máy có ứng dụng rộng khắp trong các ngành khoa học / sản xuất, đặc biệt những ngành cần phân tích khối lượng dữ liệu khổng lồ. Một số ứng dụng thường thấy.

• Xử lý ngôn ngữ tự nhiên (Natural Language Processing): xử lý văn bản, giao tiếp người – máy…

• Nhận dạng (Pattern Recognition): nhận dạng tiếng nói, chữ viết tay, vân tay, thị giác máy (Computer Vision)…

• Tìm kiếm (Search Engine)

• Chẩn đoán trong y tế: phân tích ảnh X - quang , các hệ chuyên gia chẩn đoán tự động.

• Tin sinh học phân loại chuỗi gene, quá trình hình thành gene / protein

• Vật lý: phân tích ảnh thiên văn , tác động giữa các hạt…

• Phát hiện gian lận tài chính (financial fraud): gian lận thẻ tỉn dụng

• Phân tích thị trường chứng khoán (stock market analysis)

• Chơi trò chơi: tự động chơi cờ , hành động của các nhân ảo…

• Rôbốt: là tổng hợp của rất nhiều ngành khoa học, trong đó học máy tạo nên hệ thần kinh / bộ não của người máy.

## Một số bài toán ứng dụng của học máy

* Phân lớp(classification): trong bài toán phân lớp dữ liệu sẽ được dự đoán đầu ra từ một danh sách khả năng. Một số bài toán phân lớp tiêu biểu: một bức ảnh chụp hình động vật được phân vào lớp nào sau đây [cá heo, mèo, chuột, gà], một lá thư được gửi vào hòm thư của bạn nó là [thư spam, không phải thư spam]…
* Các thuật toán học tiêu biểu: Naive Bayes, k nearest neighbor, logistic regression, decision tree, support vector machine, neural network.
* Hồi quy(regression): đầu ra một số thực. Một số bài toán hồi quy tiêu biểu là: dự đoán thu nhập cá nhân, dự đoán giá nhà đất, dự đoán sản lượng nông nghiệp,...
* Các thuật toán học tiêu biểu: linear regression, decision tree regression, generalized linear model…
* Phân cụm(clustering): bài toán này sẽ gom những nhóm đối tượng tương tự nhau về cấu trúc (theo một ngữ cảnh). Bài toán này có nhiều ứng dụng khác nhau như: phân đoạn hình ảnh, gom nhóm các đối tượng tương tự nhau…
* Các thuật toán học tiêu biểu: k-mean, k-medium, DBSCAN, Fuzzy C-Mean, Expectation–maximization algorithm…
* Giảm chiều dữ liệu (dimensionality reduction): khi đối mặt với các bài toán cụ thể ngoài thực tế, có khả năng số lượng chiều rất lớn (như bài toán ảnh chỉ cần ảnh width:1000, height:1000 là chúng ta đã có 1000000 chiều dữ liệu), điều này gây nhiều trở ngại về tốc độ tính toán và khả năng dự đoán của mô hình.
* Các thuật toán học tiêu biểu: principal component analysis, factor analysis, rough set theory...

# CHƯƠNG 2: TỔNG QUAN VỀ LDA và SVM

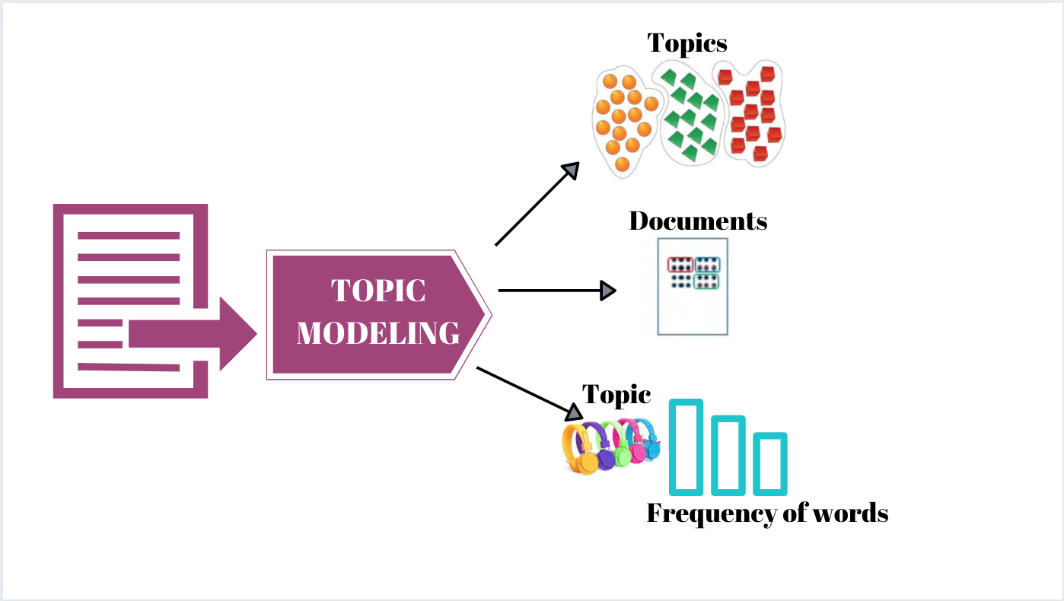
## 2.1. Tổng quan về LDA

Trong thời đại dữ liệu bùng nổ như ngày nay, dữ liệu ta thu thập được rất lớn. Trong thực tế, các vector đặc trưng (feature vectors) có thể có số chiều rất lớn, tới vài nghìn. Đồng thời, lượng điểm dữ liệu cũng rất lớn. Điều đó sẽ gây khó khăn cho việc lưu trữ và tính toán. Vì vậy, một trong những bước quan trọng trong nhiều bài toán học máy là ta phải giảm chiều dữ liệu (dimentionality reduction).

Giảm chiều dữ liệu còn là phương pháp được sử dụng để giảm vấn đề quá khớp (overfitting), nó có hai hướng là hướng lựa chọn đặc trưng (feature selection) và hướng trích xuất đặc trưng (feature extraction). Hôm nay ta sẽ tìm hiểu về một thuật toán theo hướng trích xuất đặc trưng là Latent Dirichlet Allocation (LDA).

### 2.1.1. Giới thiệu

Thuật toán Linear Discriminant Analysis (LDA) là một phương pháp thuộc họ phân tích biểu tượng tuyến tính. LDA được áp dụng rộng rãi trong các bài toán giảm chiều và phân lớp dữ liệu.



Hình 2.1. Ví dụ minh họa LDA

LDA giải quyết bài toán phân lớp bằng cách tìm ra một tập các hyperplane tối ưu phân chia các lớp dữ liệu. Hyperplane tối ưu ở đây là hyperplane làm tăng khoảng cách giữa các lớp và làm giảm khoảng cách trong từng lớp. LDA tiền đề rằng mẫu dữ liệu các lớp tuân theo phân phối Gauss và không bị chevron.

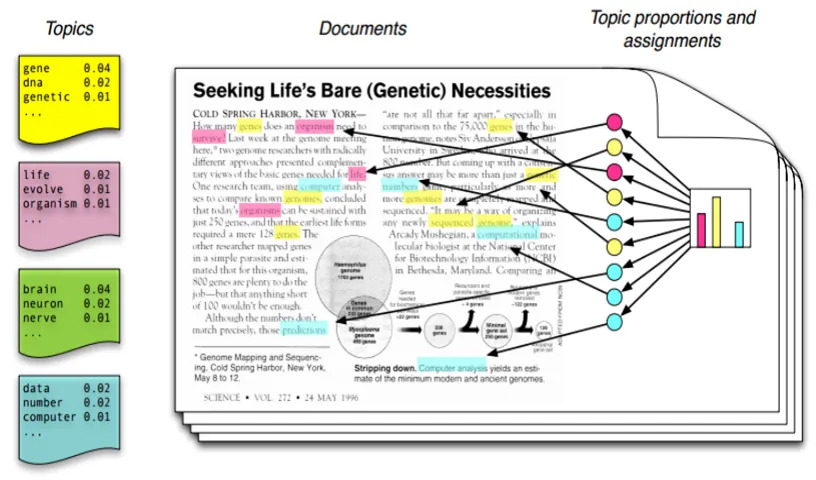
Để tìm ra các hyperplane phân lớp, LDA dựa trên việc tính toán ma trận giá trị riêng và vectơ riêng của ma trận phân tách giữa các lớp và trong lớp. Trên cơ sở đó, phương trình phân biệt được xây dựng để phân lớp dữ liệu. Mô hình LDA được huấn luyện bằng cách tính toán ma trận W dựa trên dữ liệu huấn luyện.

Ngoài ra, bởi LDA giả định phân phối dữ liệu nên nó có thể áp dụng cho bài toán giảm chiều dữ liệu. Cụ thể, LDA sẽ chọn ra các vector riêng có giá trị riêng lớn nhất để giảm chiều dữ liệu.

Như vậy, LDA là một phương pháp hiệu quả trong xử lý các bài toán phân lớp và giảm chiều với đầu vào là dữ liệu phân phối chuẩn.

### 2.1.2. Các bước của thuật toán LDA

1. Chuẩn bị dữ liệu:
   * Tiền xử lý văn bản: Loại bỏ các ký tự đặc biệt, chuyển thành chữ thường, loại bỏ stopwords, và thực hiện các bước khác để làm sạch văn bản.
   * Chuyển đổi văn bản thành biểu diễn vector: Sử dụng phương pháp như Bag-of-Words (BoW) hoặc TF-IDF để biểu diễn mỗi văn bản thành một vector.
2. Xác định số lượng chủ đề (topic):
   * Chọn số lượng chủ đề (topic) mong muốn trong mô hình LDA. Đây là một tham số quan trọng cần xác định trước.
3. Khởi tạo các tham số:
   * Khởi tạo ma trận phân phối chủ đề-từ (topic-word matrix) và ma trận phân phối văn bản-chủ đề (document-topic matrix) ban đầu với các giá trị ngẫu nhiên.
4. Huấn luyện mô hình:
   * Sử dụng phương pháp học gom nhóm, thường là học gom nhóm tiên nghiệm (Expectation-Maximization, EM), để cập nhật các ma trận phân phối chủ đề-từ và văn bản-chủ đề.
   * Lặp lại quá trình này cho đến khi hội tụ hoặc đạt đủ số lần lặp.
5. Kết quả:
   * Sau quá trình huấn luyện, thuật toán LDA sẽ cho ra ma trận phân phối chủ đề-từ và ma trận phân phối văn bản-chủ đề.
   * Các chủ đề (topic) có thể được xác định dựa trên các từ có độ quan trọng cao trong ma trận phân phối chủ đề-từ.
   * Mỗi văn bản có thể được biểu diễn bằng một vector phân phối chủ đề, cho ta cái nhìn về cấu trúc chủ đề của văn bản đó.



Hình 2.2 Thuật toán LDA

### 2.1.3. Ưu nhược điểm của LDA

LDA (Latent Dirichlet Allocation) là một mô hình phân loại và phân tích chủ đề mạnh mẽ, nhưng cũng có một số ưu và nhược điểm cần lưu ý:

**Ưu điểm:**

* Phát hiện chủ đề: LDA cho phép xác định các chủ đề chính trong tập dữ liệu, giúp hiểu và tổ chức thông tin một cách tự động.
* Phân loại hiệu quả: LDA có thể phân loại các văn bản mới dựa trên các chủ đề chính đã được xác định, giúp tăng tính hiệu quả và tự động hóa quá trình phân loại.
* Tính diễn giải: LDA cung cấp khả năng diễn giải chủ đề, giúp người dùng hiểu rõ hơn về nội dung và xu hướng trong tập dữ liệu.
* Hỗ trợ khám phá dữ liệu: LDA giúp khám phá các mối quan hệ và liên kết giữa các chủ đề trong tập dữ liệu, đóng góp vào việc phân tích và khám phá dữ liệu.

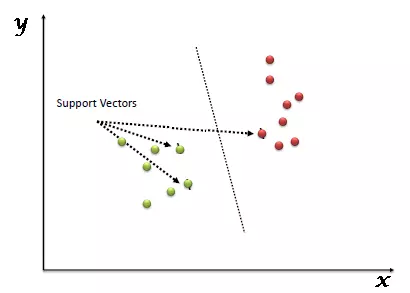
**Nhược điểm:**

* Số lượng chủ đề cần xác định trước: LDA yêu cầu người dùng xác định số lượng chủ đề trước khi huấn luyện. Việc xác định số lượng chủ đề phù hợp có thể khó và yêu cầu sự hiểu biết sâu về dữ liệu.
* Không xem xét tương quan giữa từ: LDA chỉ dựa trên tần suất xuất hiện của các từ để xác định chủ đề. Điều này có thể dẫn đến việc chủ đề không đại diện tốt nếu có các từ tương tự nhưng không liên quan xuất hiện trong các chủ đề khác nhau.
* Phụ thuộc vào tiền xử lý dữ liệu: LDA yêu cầu quá trình tiền xử lý dữ liệu đúng và hiệu quả để loại bỏ các từ không liên quan và chuyển đổi dữ liệu thành dạng phù hợp.

## 2.2. Thuật toán SVM

### 2.2.1. Giới thiệu

**SVM(Support Vector Machine)** là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.



Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Vector Machine là một biên giới để chia hai lớp tốt nhất.

### 2.2.2. Phương thức làm việc của SVM

* Kernel

Quá trình xây dựng siêu phẳng phân cách trong SVM được thực hiện qua các phép biển đổi đại số. Với kernel dạng tuyến tính (linear kernel) công thức dùng để dự đoán những điểm dữ liệu mới là: thực hiện tích vô hướng giữa đầu vào (x) với mỗi support vector (xi) như sau: f(x) = B(0) + sum(ai (x, xi)). Các hệ số B0 và ai (cho mỗi đầu vào) phải được ước tính từ dữ liệu học. Với kernel dạng đa thức (polynomial kernel) có thể được viết dưới dạng: K(x, xi) = 1 + sum(x, xi) ^ d. Còn với kernel dạng lũy thừa (exponential kernel) có dạng: K(x, xi) = exp(-gamma \* sum((x - xi2)).

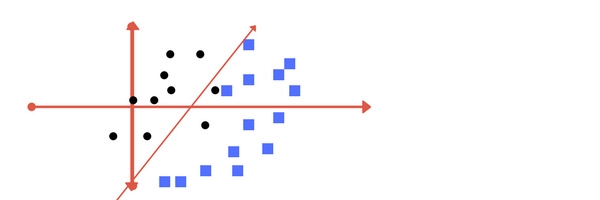
Kernel dạng đa thức và dạng lũy thừa tính toán đường phân cách ở những chiều không gian cao hơn và được gọi là kernel trick.

* Regularization

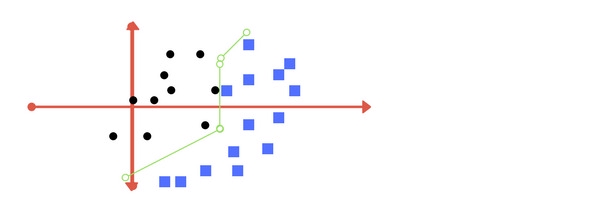
Tham số Regularization ( được nhắc đến trong thư viên sklearn là tham số C) điều chỉnh việc có nên bỏ qua các điểm dữ liệu bất thường trong quá trình tối ưu mô hình SVM. Nếu tham số này có giá trị lớn, quá trình tối ưu sẽ chọn một siêu phẳng sao cho siêu phẳng này phân cách tất cả các điểm dữ liệu một cách tốt nhất, từ đó khoảng cách giữa siêu phẳng tới các điểm dữ liệu của 2 lớp sẽ có giá trị nhỏ (small-margin). Ngược lại, khi tham số này có giá trị nhỏ, siêu phẳng sẽ được xây dựng sao cho khoảng cách với các điểm dữ liệu của 2 lớp có giá trị lớn (large-margin), kể cả khi siêu phẳng này sẽ phân loại sai nhiều điểm dữ liệu hơn.

Dưới đây là các ví dụ về 2 trường hợp chọn tham số C:

Tham số C có giá trị nhỏ

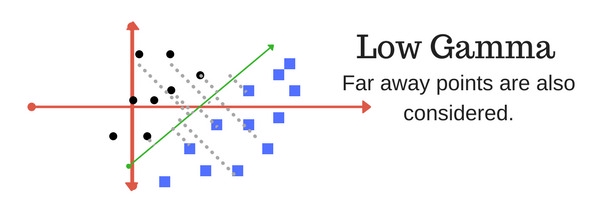
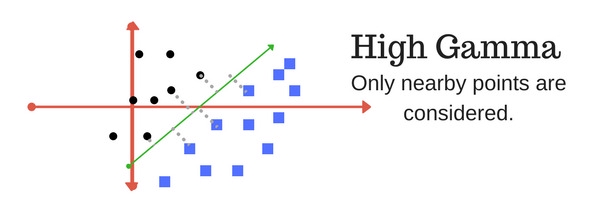


Tham số C có giá trị lớn



* Gamma

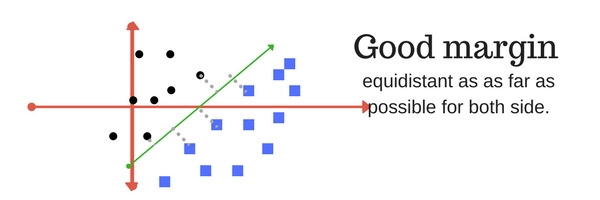
Tham số gamma xác định việc sử dụng bao nhiêu điểm dữ liệu cho việc xây dựng siêu phẳng phân cách. Với giá trị gamma nhỏ, các điểm dữ liệu nằm xa đường phân cách sẽ được sử dụng trong việc tính toán đường phân cách. Ngược lại, với giá trị gamma lớn, chỉ những điểm nằm gần đường phân cách mới được sử dụng để tính toán.



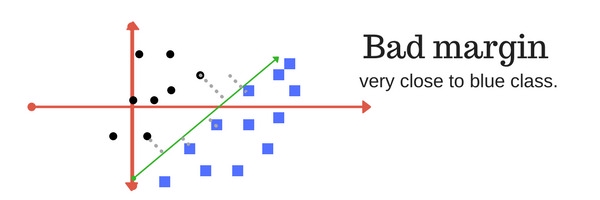
* Margin

Margin trong SVM là khoảng cách giữa siêu phẳng phân cách với các điểm dữ liệu gần nó nhất. Khoảng cách này đối với các điểm dữ liệu gần nhất của cả 2 lớp càng lớn thì mô hình càng phân loại chính xác. Các ví dụ về margin:

SVM có margin tốt : khoảng cách lớn và cân bằng giữa siêu phẳng và các điểm dữ liệu của 2 lớp



SVM có margin tồi: khoảng cách không cân bằng & nghiêng hẳn về 1 phía



# CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN VÀO BÀI TOÁN

## 3.1. Mô tả bài toán

Thư rác gây ra nhiều phiền toái và thiệt hại, bao gồm, ngăn người dùng sử dụng tối ưu thời gian, dung lượng lưu trữ và băng thông mạng. Một số lượng lớn thư rác truyền trong mạng máy tính có thể phá hủy không gian nhớ của máy chủ thư điện tử, băng thông đường truyền, tài nguyên tính toán và thời gian sử dụng của thiết bị người dung. Bằng cách xác định các chủ đề chính trong các email, LDA giúp tự động phân loại và quản lý email spam một cách hiệu quả. Quá trình này giúp tiết kiệm thời gian và công sức của người dùng, đồng thời cung cấp khả năng diễn giải chủ đề để hiểu rõ hơn về nội dung của các email spam.

## 3.2. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm là file spam.csv .‎Bộ dữ liệu này được tạo ra bởi các đóng góp cộng đồng trên Github. Mục tiêu là tạo ra được data ảo với khoảng hơn 5000 bản ghi được sử dụng với mục đích học tập.‎

**Thống kê dữ liệu thực nghiệm:**

|  |  |
| --- | --- |
| Tổng dữ liệu | 5572 |
| Số lượng email spam | 747 |
| Số lượng email không phải spam | 4825 |
| Số lớp trong tập dữ liệu | 2 |

**Tiền xử lý dữ liệu:**

* Dữ liệu văn bản (nội dung email) được chuyển đổi thành vector sử dụng CountVectorizer. Đây là một quá trình biểu diễn văn bản thành dạng số để có thể sử dụng các thuật toán máy học.
* Trong quá trình chuyển đổi, stopwords (các từ thông thường như "the", "and", "is",...) được loại bỏ để giảm nhiễu và tăng hiệu suất của mô hình.

## 3.3. Môi trường thực nghiệm

Sử dụng Visual Studio Code để lập trình và biên tập code python

Sử dụng ngôn ngữ lập trình Python.

## 3.4. Kết quả thực nghiệm

Sử dụng phương pháp học máy là SVM sử dụng phương pháp giảm chiều dữ liệu LDA. Dữ liệu được chia thành 80% huấn luyện và 20% test. Sử dụng 3 độ đo là Percision, Recall và F1-Score.

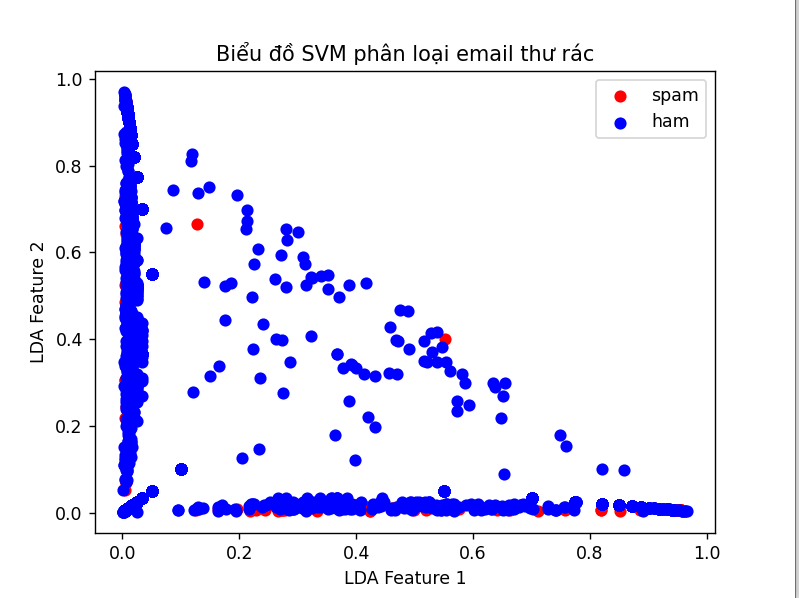
Độ đo Percision: TP/(TP+FP)

Độ đo Recall: TP/(TP+FN)

Độ đo F1-Score: 2/([1/Percision]+[1/Recall])

Bảng đánh giá kết quả của mỗi phương pháp

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **Accuracy** | **Percision (spam)** | **Recall (spam)** | **F1-Score (spam)** |
| **SVM có LDA** | 86.73% | 100.00% | 0.67% | 1.33% |



Biểu đồ SVM phân loại email thư rác

## 3.5. Đánh giá kết quả thực nghiệm

Quá trình xử lý email spam là một vấn đề quan trọng trong lĩnh vực phân loại và lọc thư rác. Trong nỗ lực giải quyết vấn đề này, code nhóm em đã cố gắng thiết kế để thực hiện các bước tiền xử lý dữ liệu, huấn luyện mô hình và đánh giá hiệu suất.

Mô hình LDA được áp dụng để giảm chiều dữ liệu và tìm ra các chủ đề tiềm ẩn trong email. Mô hình SVM được sử dụng để phân loại email là spam hay không, dựa trên các đặc trưng được trích xuất từ mô hình LDA.

Kết quả của mô hình được đánh giá thông qua các độ đo như độ chính xác, precision, recall và F1-score trên tập kiểm tra. Các kết quả này cung cấp thông tin về hiệu suất của mô hình trong việc phân loại email.

Cuối cùng, mô hình SVM được trực quan hóa bằng biểu đồ scatter plot, giúp hình dung cách mô hình phân loại các điểm dữ liệu spam và không phải spam dựa trên các đặc trưng LDA. đã thực hiện một quy trình toàn diện để xử lý email spam và đánh giá hiệu suất của mô hình. Nó đóng góp vào việc phát hiện và loại bỏ thư rác trong hộp thư đến, đồng thời cung cấp một phương pháp tiếp cận hiệu quả trong lĩnh vực này.

# KẾT LUẬN

Như vậy, nhóm chúng em đã thành công trong việc chứng minh phương pháp giảm chiều LDA không quá tốt trong những tập dữ liệu có quá ít thuộc tính, và các thuộc tính ảnh hưởng nhiều tới kết quả đầu ra. Việc làm giảm số thuộc tính trong các trường hợp này thực sự không ổn. Vì vậy không nên sử dụng LDA trong các trường hợp như vậy.

Qua việc thực hiện nghiên cứu đề tài “**Phân loại email rác bằng phương pháp giảm chiều LDA**”, nhóm chúng em đã được biết thêm rất nhiều kiến thức về thuật toán cũng như bước đầu nắm bắt được ứng dụng của môn học Học máy nâng cao. Bên cạnh đó, việc làm nghiên cứu giúp chúng em đoàn kết hơn, rèn luyện cho chúng em kỹ năng làm việc nhóm. Trong quá trình thực hiện đề tài có rất nhiều ý tưởng hay và độc đáo. Nhưng do kiến thức của chúng em còn hạn hẹp và thời gian không cho phép nên chúng em chưa thể thực hiện được những ý tưởng đó. Tuy nhiên chúng em đã cố gắng để xây dựng một chương trình hoàn chỉnh và đẹp nhất để đưa tới Thầy/Cô. Trong quá trình xây dựng chương trình nhóm chúng em khó tránh khỏi những sai xót còn tồn tại. Vì vậy chúng em rất mong được nhận lời góp ý và chỉnh sửa từ Thầy/Cô để có thể hoàn thành chương trình một cách hoàn chỉnh nhất.

Chúng em một lần nữa xin cảm ơn thầy giáo Phạm Đức Hồng đã tận tình giảng dạy cũng như hướng dẫn chúng em làm sản phẩm kết thúc học phần trong môn học **Học máy nâng cao**, thầy đã giúp đỡ chúng em trong quá trình nghiên cứu đề tài và chia sẻ những tài liệu hay cũng như các kỹ năng lập trình cần thiết.

# DANH MỤC TÀI LIỆU THAM KHẢO

[1]. Slide giảng dạy của thầy Phạm Đức Hồng.

[2]. Vũ Hữu Tiệp; Machine Learning cơ bản; Last update: March8, 2018.

[3]. Nguyễn Thanh Tuấn; Deep Learning cơ bản; Last update: October 2019.

[4]. https://machinelearningcoban.com/2017/06/30/lda/