**Introduction to Machine Learning and Data Mining**

Hệ số: 2-2-6

BTL: Capstone Project

Đề xuất đề tài trước 10/11

|  |  |
| --- | --- |
| Học có giám sát | Học không giám sát |
| Linear Regression  KNN (K-nearest neighbors)  Random Forests  SVM (Support Vector Machines)  ANN (Artificial Neural Network) | K-means |

**Lec1: Intro**

Một số thành tựu:

- IBM’s Watson

- Sự phát triển của Amazon

- Mạng GAN (2014)

- Máy đánh cờ AlphaGo (2016)

- Sinh văn bản GPT-3 (2020)

1. Machine Learning vs Data Mining

- Machine Learning:

Xây dựng những hệ thống mà có khả năng tự cải thiện bản thân bằng cách học từ dữ liệu.

- Data Mining:

Tìm ra/ Khai phá tri thức mới và hữu dụng từ các tập dữ liệu lớn.

- Data: Methodology: insight – driven

Diagram

Description automatically generated

Data collection >> Data processing >> Data vizualization & Grasping >> Analysis, hypothesis testing, & ML >> Insight & Policy Decision

Methodology: product – driven

Diagram

Description automatically generated

Business understanding >> Analytic approach >> Data requirements >> Data collection >> Data understanding >> Data preparation >> Modeling >> Evaluation >> Deployment >> Feedback

- What is Machine Learning?

Hệ thống tự cải thiện khả năng của bản thân dựa trên dữ liệu trong quá khứ.

Một máy học nếu mà nó có thể cải thiện bản thân **(P)** một cách tin cậy trong nhiệm vụ **(T)** nào đấy, dựa vào kinh nghiệm **(E)** nào đó.

(P – Performance, T – Task, E – Experience)

- What dose a machine learn?

F: x => y

Model:

- Where does a machine learn from?

Training set, training example

After learning: y = f(x)

- Two basic learning problems:

Supervised learning:

Learn y = f(x) from set X, Y sao cho yi =~ f(xi)

Classification

Multiclass (nhiều lớp)

Multilabel (đa nhãn)

Regression

Unsupervised learning:

Learn y = f(x) from set X

Clustering

Community detection

Trends detection

=> Semi-supervised learning, reinforcement learning, ...

- Design a learning system

Select a training set

Determine the type of the function to be learned

Select a representation/approximation (model) h for the unknown function f

Select a good algorithm to learn the model h, function f

- Design a learning system

Learning algorithm

No-free-lunch theorem

Training data

Over-fitting:

Hàm h là overfitting nếu tồn tạo hàm g tệ hơn h khi training nhưng cho kết quá tốt hơn h trong tương lai.

=> Regularization

**Lec2: Data crawling and pre-processing**

Cần tiền xử lý để thuận tiện lưu trữ, truy vấn, dể làm việc với các mô hình học máy.

1. Data crawling

2. Pre-processing

Data collection

Sampling: WHAT, WHY, HOW (Variety vs Biaas)

Techniques: Crowd-sourcing (khảo sát), Logging (lưu lại lịch sử), Scrapping (lọc tách từ nhiều nguồn,…)

3. Xử lý dữ liệu

Lọc nhiễu, làm sạch, số hóa.

4. Data “rawness”

Table

Description automatically generated with low confidence

5. Cleaning

Điền giá trị thiếu bằng Hồi quy, Suy diễn Bayes, …

6. Intergrating (biến đổi json sang dataframe, table)

7. Transforming: trích xuất đặc trưng ngữ nghĩa, chuẩn hóa (Semantics).

Trích chọn đặc trưng/ Giảm chiều/ Trừu tượng hóa

**Lec3: Linear regression**

Cần một hàm đánh giá lỗi để tìm được hàm f tốt nhất cho bài toán hồi quy.

Hàm lỗi thực nghiệm (residual sum of quares)

- Phương pháp bình phương tối thiếu OLS (orginary least squares).

=> Nhược điểm là  phải khả nghịch.

Cần tính tonas ma trận có rất nhiều chiều => lâu

Dễ ovefit vì tập trung vào tối ưu lỗi trên tập train.

**\*) Phương pháp hồi quy Ridge (thêm lambda)**

- Tránh được việc  phải khả nghịch => luôn hoạt động

- Giảm overfit

=> Nhược điểm là lỗi trên tập train có thê lớn hơn OLS.

Phụ thuộc vào việc lựa chọn lambda.

**\*) Phương pháp hồi quy LASSO (đổi chuẩn 2 về chuẩn 1)**

**=>**

- Phức tạp hơn Ridge vì vùng điều kiện dạng L1 khó hơn dạng L2

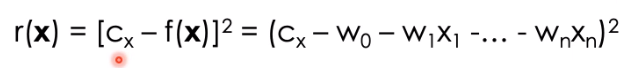
1. Regression problem

Linear model

f(x,w) = w0 + w1 x1 + w2 x2 + … + wn xn

=> Find W(w0, w1, … wn)

Loss/definition function



Expected lost (risk)

A picture containing text

Description automatically generated

Text

Description automatically generated

Khó khăn: Tại sao không làm việc với bài toán này

Lỗi thực nghiệm (Empirical loss)

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

\*) Phương pháp bình phương tối thiểu (ordinary least squares)

Text

Description automatically generated with medium confidence

=> (1) là Hàm lồi (convex function)

Text

Description automatically generated

\*Nhược điểm của OLS

\*Phương pháp giảm đạo hàm (Gradient descent)

- Ridge regression

Text, letter

Description automatically generated

Text, letter

Description automatically generated

- LASSO

Text, letter

Description automatically generated

Text

Description automatically generated

**Lec4: K-Mean**

1. Clustering

Partitoin-based clustering (K-Means)

Hierarchical clustering

Mixture models

Deep clustering

2. K-means

Đưa dữ liệu về dạng vecto

Mỗi cụm có một điểm tâm là centroid

Số K là hằng số được chọn từ trước

Độ đo d(x,y)

Tính toán phân cụm lặp đến khi hội tụ

Hội tụ khi không có sự thay đổi cụm các các điểm, không có sự thay đổi tâm các cụm và không có sự biến đổi của hàm lỗi.

Quan trọng là tính toán được tâm cụm và khoảng cách.

Những điểm ngoại lai (Outliers) rất nhạy cảm và có tác động đến kết quả phân cụm. (rất có thể outliers là nhiễu)

Khử ngoại lai: Outlier Removal (loại bỏ các điểm quá xa tâm) hoặc Random sampling (phân cụm một phần K clusters, sau đó gán dữ liệu còn lại).

Khởi tạo K tâm cụm cũng rất quan trọng, cần thử nhiều trường hợp để cho ra kết quả đúng nhất.

Một cách khởi tạo là lần lượt chọn điểm xa nhất với các cụm đã có làm tậm cụm tiếp theo.

Cụm cong là một vấn đề khó giải quyết nếu sử dụng khoảng cách không gian thông thường.

**\*) Online K-means:**

Mõi vòng lặp, một thực tế được khai thác để cập nhật các cụm có sẵn sử dụng đạo hàm Q stochastic.

**\*) Revisiting K-means:**

Text

Description automatically generated

Lặp lại đạo hàm bằng hồi quy để giảm hàm loss.

Text

Description automatically generated

- Chọn Learning rate

Text

Description automatically generated

**Lec5: KNN**

Instance-based learning/ Lazy learning/ Memory-based learning.

Called a non-parametric method.

Neighbor-based learning gồm 2 thành phần chính:

- Độ đo tương đồng (khoảng cách) giữa các thực thể/đối tượng.

- Láng giềng để dự đoán.

Có thể sử dụng hàm tối ưu lỗi Bayes

~ gần với **Maniford learning.**

1. Phương pháp phi tham số

(ý tưởng gần với: Manifold learning)

2 Hai thành phần quan trọng

Độ đo tương đồng (similarity measure)

Các láng giềng

Graphical user interface, text, application

Description automatically generated

Một vài độ đo khoảng cách:

Geometric distance

Hamming distance

Sosince measure

Normailization để đảm bảo tỷ lệ các tham số trong khoảng nhất định, không tồn tại một tham số nào chiếm tỷ trọng quá lớn. Mean = 0, Variance = 1.

Cần trọng số ứng với từng thuộc tính để đảm bảo vai trò của thuộc tính đó. Có thể 1 vài thuộc tính quan trọng hơn, 1 vài lại ít quan trọng hơn. Ví dụ, láng giếng gần hơn có trọng số lớn hơn láng giềng xa hơn. Trong số thường tỉ lệ nghịch với khoảng cách.

Có thể dùng 1 số độ đo rất linh hoạt như phân kỳ Kullback-Leibler, phân kỳ Bregman divergence,…

**Lec6: Random forest**

**\*) Decision tree**

ID3 (Iterative Dichotomiser 3):

Mô hình top-down

Tại bước N, chọn thuộc tính kiếm tra là A, giúp phân biệt tốt nhất dữ liệu trong bước N. Sinh nhánh cho mỗi giá trị của A và sau đó chia dữ liệu vào từng nhánh phù hợp tương ứng.

Dừng lại khi phân loại tất cả dữ liệu train or tất cả thuộc tính được dùng.

Cần cắt tỉa để tránh over fitting

**\*) Làm giàu thông tin: Entropy**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated  
Example:

Text

Description automatically generated

=> Nhược điểm của ID3: có thể overfit với tập train

Thiếu dữ liệu là vấn đề lớn

Cách tính nào khác cách tính Gain?

Chưa đánh giá được chi phí của thuộc tính trong ID3

Tránh overfit bằng:

Dừng học sớm trước khi quá phù hợp với dữ liệu train.

Phát triền cây đến mức đầy đủ rồi cắt tỉa.

Text

Description automatically generated

**\*) Random forests**

Dự đoán dựa vào sự kết hợp giữa nhiều cây quyết định, bằng cách lấy trung bình của tất cả các dự đoán riêng. Mỗi cây là một cây đơn gian nhưng ngâu nhiên. Mỗi cách phát triển dựa vào các thuộc tính và dữ liệu train.

3 thành phần cơ bản:

- Randomization and no pruncing: đưa ngẫu nhiên hóa, không dùng cắt tỉa

- Combination: dự đoán sau được lấy bởi trung bình những dự đoán trước

- Bagging: tập dữ liệu huấn luyện được sinh ra bởi lấy mẫu từ tập dữ liệu gốc

Huấn luyện K cây độc lập với nhau: Phán đoán dựa vào trung bình kết quả của các cây

**Lec8: Support Vector Machines (Máy vector hỗ trợ)**

Tìm siêu phẳng (phân loại tuyến tính) để chia đôi các lớp dữ liệu.

Sử dụng hàm nhân để chuyển đổi dữ liệu từ phi tuyến sang một trong gian khác mà ở đó chúng được chia cắt tuyến tính.

Linear SVM uses a linear kernel.

=> Ưu điểm: SVM có lý thuyết vững chắc để đảm bảo hiệu quả.

Đáp ứng tốt với các bài toán đa chiều.

Phương pháp phổ biến, mạnh mẽ.

Phân loại văn bản, SVM tuyến tính cho kết quả tốt.

Giả thuyết là tồn tại một siêu phẳng phân tách được 2 lớp.

1. Siêu phẳng  (mặt phẳng, bề mặt quyết định)

Các siêu phẳng tổng quát hóa cao để phân biệt các lớp (các miền, các vùng)

2. Mức lề (Margin)

Text, letter

Description automatically generated

Khoảng cách từ x\_i đến siêu phẳng 

Graphical user interface, text, application

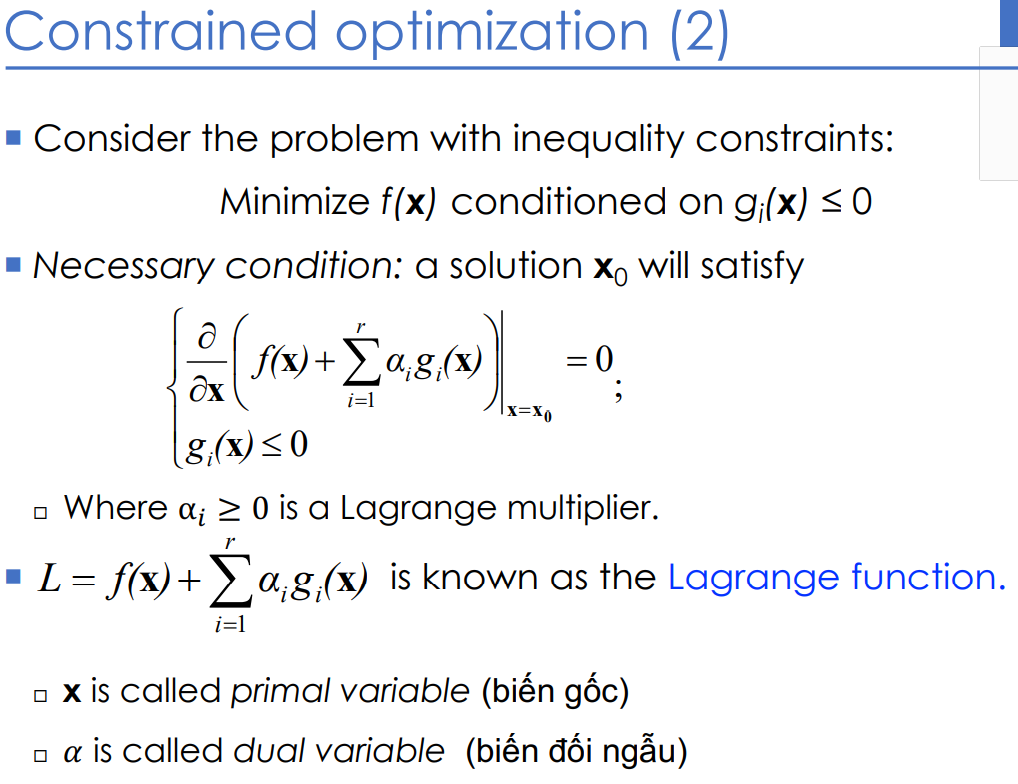
Description automatically generated

Text

Description automatically generated

Text

Description automatically generated



Text, letter

Description automatically generated

- Điểm yên ngựa: min với biến này nhưng mà max với biến khác saddle points

- Điểm dừng: stationary points

- Nghiệm Nash

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Sau 1 thôi 1 hồi đạo hàm theo w và b ta được

Text, letter

Description automatically generated

- Bài toán mục tiêu của Đối ngẫu

Text

Description automatically generated

Dạng Parabol úp ngược (hàm lõm - concave function)

Projected Gradient descent (giảm đạo hàm)

Proximal Gradient descent (giảm đạo hàm chiếu)

Text

Description automatically generated

Text, letter

Description automatically generated

2. Soft-margin SVM

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Table

Description automatically generated with medium confidence

Text, letter

Description automatically generated

**3. SVM phi tuyến**

Bước 1: chuyển đôi đầu vào sang không gian khác có nhiều chiều hơn.

Bước 2: sử dụng linear svm trong không gian mới

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Lec9: Model assessment / Model evaluation**

Hiệu năng của mổ phương pháp ML phụ thuộc vào

- Sự phân bố dữ liệu

- Tập train

- Sự đại diện của tập train so với toàn bộ không gian

Gồm:

- Đánh giá lý thuyết: learning rate, lượng train đủ chưa?, sự chính xác của phán đoán? Chịu nhiễu?

- Đánh giá thực nghiệm: quan sát kết quả thực tế trong vài tình huống như sử dunngj vài dataset và một vài độ đo hiệu năng. Tổng hợp đánh giá từ các quan sát đó.

- Đánh giá mô hình: đánh giá hiệu năng của phương pháp, mô hình chỉ dựa trên bộ dữ liệu đã quan sát.

Hold-out: chia tập train/test ngẫu nhiên

(Repeated hold-out: lấy hold-out nhiều lần rồi tính trung bình)

Graphical user interface, text

Description automatically generated

Statified sampling (lấy mẫu phân tầng): không áp dụng cho hồi quy và học không giám sát.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

**\*) Tiêu chí đánh giá**

Độ chính xác/ Tính hiệu quả/ Khả năng chống nhiễu/ Tính khả mở/ Độ phức tạp

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Lec9: ANN (Artificial Neural Network)**

Một sự đột phá khổng lồ (huge)Diagram

Description automatically generated

**Net input** thường có thêm w0 để có thể phân tách các lớp tốt hơn.

**\*) Một số hàm kích hoạt**

- hard-limited: ngưỡng (0,1) (Binary) hoặc (-1,1) (Bipolar)

Diagram

Description automatically generated with medium confidence

- threshold logic: liên tục những không hơn

Diagram

Description automatically generated

- Sigmoid

Diagram

Description automatically generated

- Tanh

Diagram

Description automatically generated

- ReLu

Graphical user interface

Description automatically generated with low confidence

Kiến trúc ANN

A picture containing diagram

Description automatically generated

Graphical user interface, text

Description automatically generated

**\*) Ý tưởng**

Học các trọng số ***w*** của mạng từ tập train ***D***

Môt vài phương pháp dựa trên đạo hàm

Lan truyền ngược (Back-propagation)

Stochastic Gradient decent (SGD)

Adam

AdaGrad

Text

Description automatically generated with medium confidence

- Perceptron: sơ khai của ANNs, output chỉ có (-1,1)

Text, letter

Description automatically generated

=> Nhược điểm:

Chỉ phân biệt được dữ liệu phân chia tuyến tính.

Logo

Description automatically generated with low confidence

**\*) Hàm lỗi**

- Đạo hàm

(Nếu chia nhỏ tập dữ liệu thành nhiều lô nhỏ, trọng số sẽ được cập nhật theo mini-batch traning)

- Lan truyền ngược

Có thể học bằng hàm phi tuyến.

Dựa vào cấu trúc mạng đã chọn trước.

Mỗi nơ-ron, hàm kích hoạt phải khác biệt.

Sử dụng chiến lược đạo hàm để cập nhật trọng số.

Gồm 2 pha:

- Thông qua chuyển tiếp: đi từ input đến output qua các tầng ẩn hidden layers.

- Truyền ngược lỗi: Tính lỗi các giá trị output, lan truyền ngược về input.

- Lỗi lan truyền ngược được thực hiên bằng tính toán hồi quy giá trị đạo hàm cục bộ của mỗi nơ-ron.

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Diagram

Description automatically generated

Diagram

Description automatically generated

Diagram

Description automatically generated

\*) Momentum

- Thêm quán tính để khống chế đạo hàm khi learning rate thay đổi.

Các yêu tố như:

Khởi tạo trọng số nhỏ để tránh điểm yên ngựa hoặc điểm đứng im

Learning rate phù hợp + hằng số quán tính alpha.

A picture containing diagram

Description automatically generated

Số lượng nơ-ron

Giới hạn học

Text

Description automatically generated

**Lec10: Prob Models (Probabillistic Modeling) Mô hình Xác suất**

- Mô hình xác suất để thích ứng với môi trường không ngừng biến đổi trong thực tế.

- Text, letter

Description automatically generated

- Một khu vực sai số (không chắc chắn) cho một phán đoán.

- Cơ bản về Lý thuyết xác suất

Biến ngẫu nhiên

Xác suất kết hợp

Xác suất có điều kiện

Độc lập xác suất

Luật Bayes

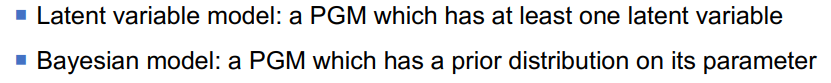
Mô hình đồ thị xác suất

Mô hình Gauss

Mô hình hỗn hợp (đánh trọng số cho từng cụm Phân bố Gauss)

**\*) Một vài loại mô hình**

Mô hình đồ thị xác suất (Probabilistic graphical model – PGM): Graph + Probability Theory.



**\*) Gaussian mixture model (GMM)**

Chart

Description automatically generated with medium confidence

\*) Một vài mô hình xác suất PGM

GMM (Gaussian mixture model): mô hình dữ liệu thực

Latent Dirichlet allocation (LDA): mô hình dữ liệu chủ đề ẩn trong văn bản

Hidden Markov model (HMM): mô hình time-series

Conditional Random Field (CRF): phán đoán có cấu trúc

Deep generative models (DGM): mô hình dữ liệu có cấu trúc và nhân tạo

\*) Hai vấn đề

- Suy diễn

- Học (ước lượng)

- Hướng tiếp cận suy diễn

Maximum Likelihood Estimation (MLE, cực đại hoá khả năng)

Maximum a Posterior Estimation (MAP, cực đại hoá hậu nghiệm): tìm cực đại tại điểm h\*

MLE là trường hợp đặc biệt của MAP

Full Bayesian inference (Gaussian Naïve Bayes): tìm cực đại của cả hàm

**\*) Expectation Maximization (EM algorithm)**

GMM revisit: Đánh giá học GMM với K phân bố Gauss

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Chart, scatter chart

Description automatically generated

**Lec11: Regularization để tránh overfit**

Dùng các chuẩn 0, chuẩn 1, chuẩn 2, chuẩn p…

Hồi quy Ridge là một phiên bản chính quy của OLS (còn có hồi quy LASSO).

Regularization trong SVM: MAP (cực đại hóa hậu nghiệm)

MAP in Ridge, MAP in Lasso

- Phương pháp Drop out

- Phương pháp Batch Normalization

- Data augmentation

- Early stopping (Thường dùng trong cây quyết định – Decision Tree)

- Thêm hàm mất mát

\*) Đánh giá bằng Hệ số tương quan

**Lec12: Data mining**

Knowledge discovery and data mining KDD (trích xuất tri thức ấn trong dữ liệu lớn). Data => Information => Knowledge => Meta-knowledge

Dựa vào dữ liệu cũ để phán đoán xu hướng tương lai.

Ngoại lai => những bất thường trong thực tế

Descriptive mining: summarization/ depencdency modeling

KDD: data type (có giám sát/có nhãn, không giám sát/không nhãn) – metadata

Diagram

Description automatically generated

**Lec13: Association rule + Apiori**

-

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

- Lưu ma trận hoặc lưu ma trận thưa

=> Thuật toán Priori (Agrawal and Srikant) – IBM

Một tập lớn là tập thường xuyên thì tập con cũng là tập thường xuyên.

Nếu một tập chứa item có support bé thì không quan tâm đến tập nào lớn hơn chứa item đó.

Diagram

Description automatically generated