**Introduction to Machine Learning and Data Mining**

Hệ số: 2-2-6

BTL: Capstone Project

Đề xuất đề tài trước 10/11

**Lec1: Intro**

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

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

Methodology: product – driven

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 machne 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**

1. Data crawling

2. Pre-processing

Data collection

Sampling

Techniques:

Crowd-sourcing

Logging

Scrapping

**Lec3: Linear regression**

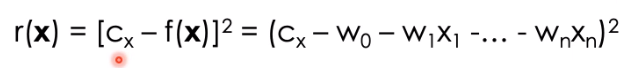
1. Regression problem

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

Hỉearchical clustering

Mixture models

Deep clustering

2. K-means

Đưa dữ liệu về dạng vecto

Mỗi cụm có một điểm tâm là centroid

Độ đo d(x,y)

**Lec6: KNN**

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

3.

Graphical user interface, text, application

Description automatically generated

4.

**Lec7: Random forest**

1. Phương pháp phi tham số

(ý tưởng gần với: Manifold learning)

2 Hai thành phần quan trọng

ID3 (Iterative Dichotomiser 3):

Cần cắt tỉa để tránh over fitting

Random forests

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

Combination: dữ 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)**

1. Siêu phẳng 

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

2. Mức lề (Margin)

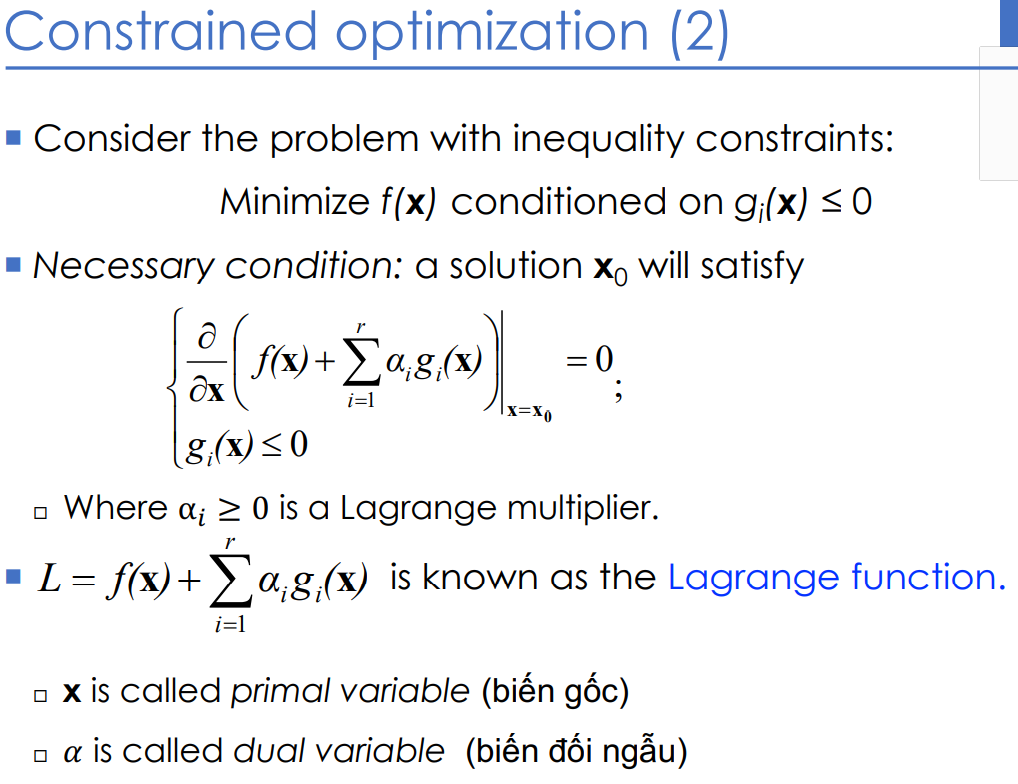
Khoảng cách từ x\_i đến siêu phẳng 

Graphical user interface, text, application

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, letter

Description automatically generated

3. SVM phi tuyến

Text, letter

Description automatically generated

Text, letter

Description automatically generated

**Lec9: Model assessment**

1. Graphical user interface, application

Description automatically generated

**Lec10: ANN (Artificial Neural Network)**

Diagram

Description automatically generated

**Lec10: Prob Models (Probabillistic Modeling)**

- 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)

- 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\*

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

\*) Expectatioon Maximization

GMM revisit

K-means

Lower Bound

**Lec10: Regularization**

- Phương pháp Drop out

- Phương pháp Batch Normalization

- Data augmentation

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

\*) Đánh giá bằng Hệ số tương quan