#### **Introduction to Machine Learning**

Thanh-Nghi Do Can Tho University *dtnghi@cit.ctu.edu.vn* 

> Can Tho Dec. 2019

#### Content

Motivation

Definitions and problemsApplications

#### Content

#### Motivation

Definitions and problemsApplications

- Definitions & problems
- Applications

## Motivation

#### • Why machine learning is important?

- Growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.
- Physical, biological, financial phenomenon, etc.: too complex to be described in a deterministic way
- Machine learning techniques
- Computers can learn without being explicitly programmed to perform specific tasks
- Voluntary to replace human experts

- Definitions & problems
- Applications

## Motivation

#### • Why machine learning is important?

- Adapt to the environment
- Deal with new problems
- Data analysis
- Pattern recognition
- Computer vision
- Control robots, self-driving cars
- Text mining, chatbots
- Information retrieval
- Game, etc

#### Content

Motivation

#### **Definitions and problems**

- Motivation
- Definitions & problems
- Applications

#### Wikipedia

• Learning is the process of acquiring new, or modifying existing, knowledge, behaviors, skills, values, or preferences

#### Arthur L. Samuel, AI pioneer, 1959

- Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed
- Tom Mitchell, Professor at Carnegie Mellon University
  - A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E* 7

- Motivation
- Definitions & problems
- Applications

#### **The Traditional Programming Paradigm**



- Motivation
- Definitions & problems
- Applications

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed – Arthur Samuel (1959)



- Motivation
- Definitions & problems
- Applications

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

#### Handwriting Recognition Example:





- Motivation
- Definitions & problems
  - Applications

# Machine Learning, AI, Deep Learning



- Motivation
- Definitions & problems
- Applications

#### Problems

#### Supervised learning

• Given a set of *m* training examples

$$D_{Train} = \{(X_1, Y_1), \dots, (X_m, Y_m)\} \text{ such that}$$
$$X_i \in \mathbb{R}^n \text{ is the feature vector of the } i\text{-th example and}$$
$$Y_i \text{ is its label (i.e., target)}$$

- A learning algorithm tries to find a function  $\phi: X \rightarrow Y$  $Y = \phi(X) + \varepsilon$
- Methods: decision trees, neural networks, support vector machines, *k* nearest neighbors, naive Bayes, etc

- Motivation
- Definitions & problems
- Applications

#### Classification problem



- Motivation
- Definitions & problems
- Applications

#### Classification problem



- Motivation
- Definitions & problems
- Applications

### Classification problem



- Motivation
- Definitions & problems
- Applications

#### Regression problem



- Motivation
- Definitions & problems
- Applications

#### Regression problem



- Motivation
- Definitions & problems
- Applications

#### Regression problem



- Motivation
- Definitions & problems
- Applications

### Problems

#### Unsupervised learning

• Given a set of *m* training examples without labels  $D_{Train} = \{X_1, \ldots, X_m\}$  such that  $X_i \in \mathbb{R}^n$  is the feature vector of the *i*-th example

• A learning algorithm tries to find the hidden structure (i.e. clusters) in unlabeled data according to similarities, patterns and differences without any prior training of data

- Motivation
- Definitions & problems
- Applications

## Unsupervised learning: clustering







- Motivation
- Definitions & problems
- Applications

#### Unsupervised learning: clustering



- Motivation
- Definitions & problems
- Applications

## Challenges

#### Statistical techniques

- *m*, *n* are reasonable; under assumption of the linear model and the data distribution is known (Gaussian, Binomial, Poisson)
- It seems to be very difficult
  - Small samples
  - Unknown data distribution
  - Unknown target function
  - Non-linear models
  - Very large data

- Motivation
- Definitions & problems
- Applications

## Challenges

#### Method selection

- Many learning algorithms
- No free lunch in machine learning
- Given a problem => model selection
- Empirical test, tuning hyper-parameters, comparison
- Training time, testing time, model quality

# Bias and Variance in Machine Learnin $g_{pplications}$ Error = Bias<sup>2</sup> + Variance



Motivation

# Bias and Variance in Machine Learning $g_{pplications}$ Error = Bias<sup>2</sup> + Variance

- Let *Y* be the true value, and let *Y*' be an estimator of *Y* based on a sample of data
  - Bias(Y') = E(Y') Y
  - Low bias value means that the model will closely match the training dataset.
  - High bias value means that the model will not match the training dataset closely.

## Bias and Variance in Machine Learning $g_{pplications}$ Error = Bias<sup>2</sup> + Variance

- Let *Y* be the true value, and let *Y*' be an estimator of *Y* based on a sample of data
  - *Variance* =  $E[(Y' E(Y'))^2]$
  - Low variance value means that the model is less sensitive to changes in the training data and can produce consistent estimates of the target function with different subsets of data from the same distribution.
  - High variance value means that the model is very sensitive to changes in the training data and can result in significant changes in the estimate of the target function when trained on different subsets of data from the same distribution.

# Bias and Variance in Machine Learning $g_{pplications}$ Error = Bias<sup>2</sup> + Variance

- Four combinations between bias and variance
  - High Bias, Low Variance: A model with high bias and low variance is said to be underfitting.
  - High Variance, Low Bias: A model with high variance and low bias is said to be overfitting.
  - High Bias, High Variance: A model has both high bias and high variance, which means that the model produces inconsistent and inaccurate predictions on average.
  - Low Bias, Low Variance: A model that has low bias and low variance means that the model produces consistent and accurate predictions. But in practice, it's not possible.

- Motivation
- Definitions & problems
- Applications

## Modeling errors in learning process



**Definitions & problems** 

# Statistical learning theory (Vapnik, 2000)

Four elements of empirical inference theory

- Consistency of empirical risk minimization (ERM) inference
- Non-asymptotic error bounds on ERM inference
- Controlling the accuracy of inference
- Constructing algorithms that control the accuracy of inference

**Definitions & problems** 

# Statistical learning theory (Vapnik, 2000)

Four elements of empirical inference theory

• Consistency of empirical risk minimization (ERM) inference

#### Expected Risk:

The theoretical goal of learning is to find a function h from the hypothesis space  $\mathcal{H}$  that minimizes the expected risk:

$$R(h) = \mathbb{E}_{(x,y) \sim P}[L(h(x),y)]$$

- R(h): Expected risk.
- L(h(x), y): Loss function measuring the error of h(x) with respect to the true label y.
- P(x, y): Unknown joint distribution of input-output pairs.

However, since P(x, y) is unknown, R(h) cannot be computed directly.

**Definitions & problems** 

# Statistical learning theory (Vapnik, 2000)

Four elements of empirical inference theory

• Consistency of empirical risk minimization (ERM) inference **Empirical Risk**:

Instead, we approximate the expected risk using the **empirical risk**, defined as the average loss over the training dataset:

$$R_{ ext{emp}}(h) = rac{1}{N}\sum_{i=1}^N L(h(x_i),y_i)$$

- $(x_i, y_i)$ : Training data points.
- N: Number of training samples.

 $R_{
m emp}(h)$  serves as an estimate of R(h) based on the observed data.

Definitions & problems

# Statistical learning theory (Vapnik, 2000)

Four elements of empirical inference theory

• Non-asymptotic error bounds on ERM inference Which value of risk can one guarantee for the chosen function. That is, find the smallest value A for which with probability  $1 - \eta$  the inequality

$$R(h) \le A$$

holds true.

**Definitions & problems** 

# Statistical learning theory (Vapnik, 2000)

#### • Four elements of empirical inference theory

• Controlling the accuracy of inference

Which principle of inference guarantees the smallest error bound. It means ensuring that a model reliably predicts outcomes with minimal error, both during training and when applied to unseen data.

• Constructing algorithms that control the accuracy of inference How to construct algorithms of empirical inference that guarantee the smallest error bounds. It involves designing models and training processes that balance empirical performance with generalization.

#### Content

Motivation

Definitions and problems

- Motivation
- Definitions & problems
- Applications

#### Bioinformatics

- Gene expression classification
- Diagnostic support system
- Computer vision
- Handwriting character recognition
- Speech recognition
- Fraud detection
- Financial market analysis

- Motivation
- Definitions & problems
- Applications

#### Information retrieval

- Analyzing and indexing images and vidéos
- Content-based image retrieval
- Games
- Natural language processing
- Recommendation system
- Sentiments analysis
- Robotics
- Self-driving cars
- Cyber-security

#### etc.

- Motivation
- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications

# Applications Hight quality ICH image DB Visual features



40

#### Motivation

- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications



Use your voice to send messages, set reminders, search for information, and more.



- Motivation
- Definitions & problems
- Applications





- Definitions & problems
- Applications



- Motivation
- Definitions & problems
- Applications



#### ⑤ OpenAI

## ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to <u>InstructGPT</u>, which is trained to follow an instruction in a prompt and provide a detailed response.

- Motivation
- Definitions & problems
- Applications

- Motivation
- Definitions & problems
- Applications

Write a Python script that does the following:

- loads a dataset from csv on disk to Pandas DataFrame
- ensures that dataset is entirely numeric; converts string values to numeric and tracks this
- conversion in a dictionary to reference later
- checks for missing values in the DataFrame and imputes missing values for any record

that has one missing value; entirely drops a record that has more than one missing value

- splits the dataset into test and training sets

#### The Response

```
import pandas as pd
# Load dataset from csv to DataFrame
df = pd.read_csv('dataset.csv')
# Initialize dictionary to track string to numeric conversions
conversions = {}
# Convert string values to numeric and track conversions in dictionary
for col in df.columns:
    if df[col].dtype == object:
        conversions[col] = {val: i for i, val in enumerate(df[col].unique())}
    df[col] = df[col].map(conversions[col])
# Check for missing values
missing values = df.isna().sum()
```

- Definitions & problems
- Applications

Nhờ chatGPT viết thư bày tỏ tình cảm với cô bạn gái cùng lớp

Yêu cầu: Viết cho tôi một lá thư để bày tỏ tình cảm của tôi với cô bạn gái cùng lớp. Sau đây là các đặc điểm của tôi.1: Tôi là một học sinh rất giỏi môn Tin học 2: Gia đình tôi là một gia đình truyền thống, bố mẹ tôi đều là giáo viên 3: Bố mẹ tôi thích bạn gái của tôi xuất thân từ một gia đình truyền thống, biết kính trên nhường dưới. 4: Tôi thích chơi đàn guitar cổ điển và có thể độc tấu guitar 5: Tôi yêu thích nhạc tình ca của Ngô Thụy Miên. 6: Tôi cũng yêu thích nhạc của Trịnh Công Son về thân phận con người và đât nước. 7:Tôi ít tiếp xúc với bạn gái 8: Bạn gái của tôi cũng học giỏi và rất nết na.

#### ChatGPT:

Kính gửi [Tên cô bạn gái],

Tôi viết thư này để bày tỏ cảm tình của tôi với cô. Tôi là một học sinh rất giỏi môn Tin học, và gia đình tôi là một gia đình truyền thống. Bố mẹ tôi đều là giáo viên và họ thích cô xuất thân từ một gia đình truyền thống, biết kính trên nhường dưới.

Ngoài việc học tập, tôi cũng thích chơi đàn guitar cổ điển và có thể độc tấu guitar. Tôi yêu thích nhạc tình ca của Ngô Thụy Miên và cũng rất thích nhạc của Trịnh Công Son về thân phận con người và đất nước.

Tôi biết rằng tôi chưa có nhiều cơ hội để tiếp xúc với cô, nhưng tôi muốn bắt đầu một mối quan hệ với cô. Tôi biết cô cũng là một học sinh rất giỏi và rất nết na, và tôi tin rằng chúng ta có thể học hỏi nhiều điều từ nhau.

Tôi mong sẽ có cơ hội được gặp cô trong tương lai và tìm hiểu thêm về cô.

Trân trọng, [Tên của bạn]

- Motivation
- Definitions & problems
- Applications



- CACGTAGACTGAGGACGAGAAG CACGTAGACTGAGGACGACTCCTCTTC (transcription) CUGCAUCUGACUCCUGAGGAGAAG (translation) ···· V H L T P E E K ··· protein



Cancer?

- Motivation
- Definitions & problems
- Applications

- Motivation
- Definitions & problems
- Applications



