# Introduction to Machine Learning

**Lecture 1**

**Guang Bing Yang, PhD**






- My name: Guangbing Yang
- My research interest and experience focuses on: Text summarization, Visual semantics, Natural Language Processing, Machine Learning/deep learning/ reinforcement learning algorithms, statistics, and data science.
- My Email: yguangbing@gmail.com (Chula's email Guang.B@chula.ac.th), when you send me an email to my Chula's email, please also cc to my gmail address.
- My brief CV:
  - Guangbing Yang, Ph.D. - Texas A&M University-Commerce
  - Guangbing Yang - Google Scholar






## Evaluation

- **Five assignments**:
  - Assignment 1 worth 10%, Assignment 2 to 5 worth 15%
- Final project and presentation, worth 25%
- Attendance and activity, worth 5%

**Tentative** Dates - Check the MyCourseVille, or lecture notes.

Project: Proposal Due April 9, 2021

Presentation: May 14th, 2021

Project report: May 14th, 2021






## Project

- The purpose of the final project is to provide you a bit of experience trying to do a very basic but original research in machine learning and coherently writing up your result.
- In this project, **what is expected:**
  - A simple but original idea, clearly describe and discussed.
  - Link it to existing methods
  - Implement and test (model performance evaluation) on a small scale problem
- **What is required**:
  - write some basic code to build a machine learning model and train/test it on some data
  - make some figures (e.g., architecture, system design, work flow, training/testing evaluation result plots, and others)
  - read some research papers, collect references, and
  - write an essay (no more than 3 pages) to discuss your model, algorithm, and results, etc.

# Text Books

**As reference books**

- Christopher M. Bishop (2006)

Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

- Kevin P. Murphy (2012)

Machine Learning — A Probabilistic Perspective, MIT, ISBN 978-0-262-01802-9

- Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016)

Deep Learning, MIT, ISBN: 9780262035613






# Introduction to Machine Learning

**What is Machine Learning about?**

- Definition: *the study of **computer algorithms** that **improve automatically** through **experience*** [1].
- Machine learning is a very dynamic field that lies at the intersection of Probability theory, Statistics and Computer Science
- The purpose of machine learning includes:
  - develop **algorithms** that can learn from data.
  - construct **stochastic models**.
  - make **predictions** and **decisions** with new data.

[1]. Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.






# Machine Learning's Examples

- Speech processing: - speech recognition, voice identification, e.g., Apple Siri, Amazon Alexa, and Google home, and many others
- Image processing and object identification: - Face recognition, etc
- Robotics: - autonomous car driving, planning, control, etc.
- Biostatistics / Computational Biology: Google brain projects, mining genomic data.
- Neuroscience
- Medical Imaging: - computer-aided diagnosis, image-guided therapy, etc.
- Information Retrieval / Natural Language Processing: - Semantic search, big data, machine translation, text to images, image to text, etc.






# Pattern discovery

- Huge increase in both computational power and amount of data available from web, video cameras, social medias, etc.
- Provide both capabilities and opportunities to machine learning to discover interesting underlying structure, cause, and correlations from data.

# Types of Machine Learning

Given a series of input vectors: $X_1, X_2, X_3, \ldots, X_n$

- **Supervised Learning**: the goal is to learn a **mapping** from inputs x to outputs y, given a labeled set of input-output pairs $\mathbb{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$, $\mathbb{D}$ is called training dataset.
- **Unsupervised Learning**: the goal is to learn interesting **patterns** in the data. Only inputs x are given, no labeled data provided. Sometimes, the unsupervised learning is also called **knowledge discovery**.
- **Reinforcement Learning**: the goal is to learn actions that maximize the reward in a long-term. RL is beyond the scope of this course. But we may introduce it a little if we have time.
- **Semi-supervised Learning**: given a few of labeled data, but lots of unlabeled data. (Not cover this topic in this course)

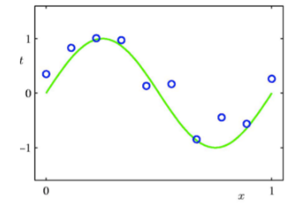
yguangbing@gmail.com,

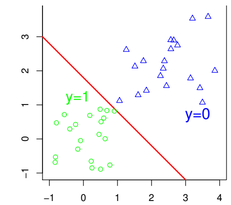




# Supervised Learning

- **Regression**: target output $y_i$ are continuous. The goal is to predict the output (real values) given new inputs
- **Classification**: target output $y_i$ are discrete class labels. The goal is to correctly classify new inputs



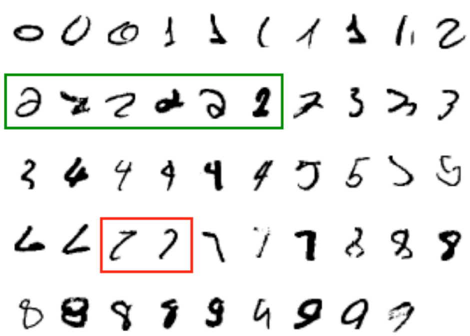




yguangbing@gmail.com,

© GuangBing Yang, 2021. All rights reserved.



# Examples of classification

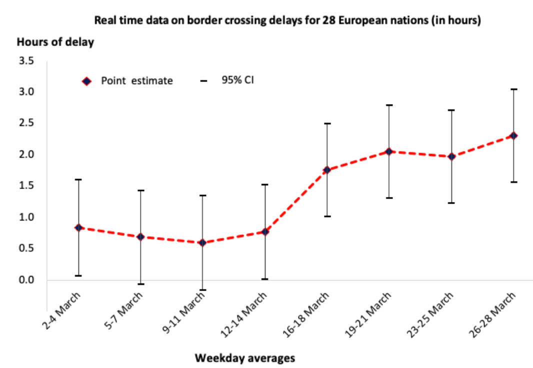


yguangbing@gmail.com,

© GuangBing Yang, 2021. All rights reserved.

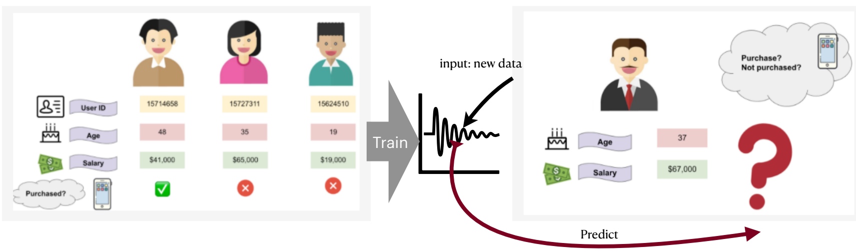


# Examples of regression




Reference: https://voxeu.org/article/covid-concussion-and-supply-chain-contagion-waves

# Examples of Logistic Regression (classification)




Given same labeled data: client information with purchased (targets/labels), to predict a new customer whether or not buy a new phone.






# Popular Algorithms

**For Regression Problems**

- Linear regression: Objective function: $min_w[Xw - y]^2$, no regularization
- Ridge regression: The ridge coefficients minimize a penalized residual sum of squares: $min_w[Xw - y]^2 + \alpha[w]^2$, L2 regularization
- Lasso: a linear model that estimates sparse coefficients. $min_w \frac{1}{2n_samples}[Xw - y]^2 + \alpha w$, L1 regularization
- Elastic-Net: a linear regression model trained with both L1 and L2 regularization of the coefficients. $min_w \frac{1}{2n_samples}[Xw - y]^2 + \alpha\rho w + \frac{\alpha(1-\rho)}{2}w^2$, L1 plus L2 regularizations.
- Bayesian regression: a fully probabilistic model with normal distribution around Xw, $p(y \mid X, w; \alpha, \sigma) = \mathbb{N}(y \mid Xw, \alpha, \sigma^2)$






# Popular Algorithms

**For Classification Problems**

- Logistic regression
- Naive Bayes
- Decision Trees
- k-Nearest Neighbours
- Random Forests
- Gradient Tree Boosting
- XGBoost






# Brief History of Machine Learning

- Alan Turing's paper "Computing Machinery and Intelligence" in 1950, probably the earliest start the research of machine learning.
- Arthur Samuel in 1959 first time, stated the term "machine learning"
- Tom M. Mitchell in 1997 provided a widely quoted, more formal definition of ML
- After 1990s, machine learning became a separated filed from Artificial Intelligence.
- Nowadays, ML is an essential part of the AI.
- After 2010, Deep Learning and Reinforcement Learning, which are core part of ML, are more popular.

## Current & Future

- Due to globalization, the majority of jobs moved to "knowledge work" from "manual labor".
- The massive amounts of data and information available to us from the web make the jobs of knowledge workers even harder.
- Making sense of all the data with our job in mind is becoming a more essential skill.
- Machine learning will help you get through all data and extract some information.
- Machine learning becomes the essential skills.
- It has a very bright future!






## Reviews

- Linear Algebra
  - Matrix Multiplication
    - Vector-Vector Multiplication
    - Matrix-Vector Multiplication
    - Matrix-Matrix Products
    - Operations and Properties
  - Matrix Calculus
- Probability Theory
  - General Concepts
  - Expected Values
  - Common Probability Distributions






### Linear Algebra Brief Review

- **Linear algebra** is a branch of mathematics providing a concise way to represent and operate on a set of linear equations via vectors and matrices.
- For example, system equations:

$$2x_1 + 3x_2 = 15$$
$$-x_1 + 2x_2 = 6$$

can be represented using matrix and matrix operations

$$Ax = y$$

$$A = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}, y = \begin{bmatrix} 15 \\ 6 \end{bmatrix}$$

- To solve this system equation, many steps may be needed, but later, you will see we can get it quickly and easily using matrix operations.






### Matrix Multiplication

- The **product** of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix:

  $C = AB \in \mathbb{R}^{m \times p}$

  where $C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$

  Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B.

### Vector - Vector Multiplication

- The **product** of two vectors, $a, b \in \mathbb{R}^n$, the outcome of the product $a^Tb$, also called the ***inner product*** or ***dot product*** of vectors, is a real number calculated by
- In this case, $a^Tb = b^Ta$ because the size of the vector a and b are the same, which is n.

$$a^Tb \in \mathbb{R} = [a_1 a_2 ... a_n] \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_i b_i = (a_1b_1 + a_2b_2 + ... + a_nb_n)$$

yguangbing@gmail.com, January 22nd, 2021 



### Vector - Vector Multiplication

- In contrast, for two vectors, $a \in \mathbb{R}^m, b \in \mathbb{R}^n$ with different size, the ***outer product*** of $ab^T \in \mathbb{R}^{m \times n}$, is defined as,
- In this case, $ab^T$ is a $m \times n$ matrix rather than a real scaler value.
- For others, matrix - vector, vector - matrix and matrix - matrix, please see lecture note: 'review-linear-algebra.pdf'

$$ab^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_m \end{bmatrix} [b_1 b_2 ... b_n] = \begin{bmatrix} a_1b1 & a_1b_2 & ... & a_1b_n \\ a_2b_1 & a_2b_2 & ... & a_2b_n \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ a_mb_1 & a_mb_2 & ... & a_mb_n \end{bmatrix}$$

yguangbing@gmail.com, January 22nd, 2021 



### Operations and Properties

- **Identity Matrix & Diagonal Matrix**
  - a square matrix with ones on the diagonal and zeros everywhere else. It is denoted as: $I \in \mathbb{R}^{n \times n}, I_{ij} = \begin{cases} 1 & \text{if i = j} \\ 0 & \text{if } i \neq j \end{cases}$
  - property of identity matrix: for all $A \in \mathbb{R}^{m \times n}, AI = A = IA$
  - diagonal matrix is a matrix where all non-diagonal elements are zeros, denoted as $D = \text{diag}(d_1, d_2, \ldots, d_n)$. It is not necessary a square matrix, with $D_{ij} = \begin{cases} d_i & \text{if i = j} \\ 0 & \text{if } i \neq j \end{cases}$

yguangbing@gmail.com, January 22nd, 2021 



### Operations and Properties

- **The Transpose**
  - The transpose of a matrix results from flipping the rows and columns.
  - Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by $(A^T)_{ij} = A_{ji}$
  - The properties:
    - $(A^T)^T = A$
    - $(AB)^T = B^TA^T$
    - $(A + B)^T = A^T + B^T$

yguangbing@gmail.com, January 22nd, 2021 

## Operations and Properties

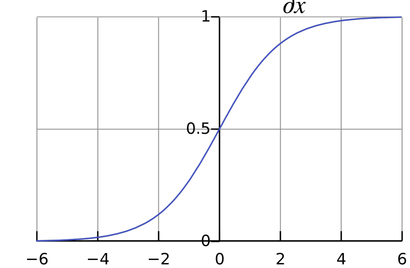
- **The Inverse**
  - The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted as $A^{-1} \in \mathbb{R}^{n \times n}$.
  - It is unique and $A^{-1}A = I = AA^{-1}$
  - Note that not all matrices, including same square matrices, have inverses. By definition, non-square matrices have no inverses.
  - Particularly, if $A^{-1}$exists, we say A is ***invertible*** or ***non-singular***, otherwise, it is ***non-invertible*** or ***singular***. Also, the determinant of the A (detA) is not zero, and is vice versa.
  - The properties:
    - $(A^{-1})^{-1} = A$
    - $(AB)^{-1} = B^{-1}A^{-1}$
    - $(A^{-1})^T = (A^T)^{-1}$

yguangbing@gmail.com, January 22nd, 2021 



## Sigmoid function & its derivative

- **The Sigmoid function**
  - $S(x) = \frac{1}{1+e^{-x}}$ is denoted as $x \in \mathbb{R}^n$.
  - Its derivative or gradient defined as $\frac{\partial S(x)}{\partial x} = S(x)(1 - S(x))$
  - 




Sigmoid function-chart from Wikipedia page: Sigmoid function

yguangbing@gmail.com, January 22nd, 2021 



## Probability Theory Brief Review

- **Probability Theory** studies the **uncertainty.**
- **Statistics** somehow apply probability theory to explain the variation in some measure of interest. In other words, probability quantifies uncertainty, statistics explains variation.
- For example,
  - Roll a 6-side die. What is the probability of obtaining a 6? (a probability problem)
  - Observe the variation of the annual income of a person. What factors explain the variation in a person's income.(a statistic problem, which is 'variation = factors of observation + random errors'. Clearly, no way to account for all factors that affect person's income, have to leave any remaining variation to uncertainty.
- In later lectures, you will see a loss function (or cost function) can be taken as the random error in the variation expression list above.
- That is why we say machine learning use statistics and probability to address the uncertainty problem.

yguangbing@gmail.com, January 22nd, 2021 



## Probability Theory Brief Review

- **Elements of Probability**
  - **Sample space:** the sample space is denoted by $\Omega$**,** it is the event set
  - **Events**: a particular subset of $\Omega$, denoted as A, $A \subseteq \Omega$.
  - **Probability measure**: A function $P : F \to R$ that satisfies the following properties:
    - $P(w) \geq 0$
    - $\sum_{w \in \Omega} P(w) = 1$
    - If $A_1$ and $A_2$ are disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, more generally, if $A_1, A_2, ...A_n$ are mutually disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

yguangbing@gmail.com, January 22nd, 2021 

## Probability Theory Brief Review

- **Properties of Probability:**
  - If $A \subseteq B, P(A) \leq P(B)$
  - $P(A \cap B) = min(P(A), P(B))$
  - $P(A \cup B) \leq P(A) + P(B)$ - (Union Bound)
  - $P(A^c) = 1 - P(A)$ since $A^c$ is called A`s complement, $A^c$ and A are disjoint. $A^c \cup A = \Omega$, and $P(A^c \cup A) = P(\Omega) = 1 = P(A^c) + P(A)$
  - If $A_1, \ldots, A_k$ are a set of disjoint events such that $\sum_{i=1}^{k} A_k = \Omega$, then $\sum_{i=1}^{k} P(A_k) = P(\Omega) = 1$ (Law of Total Probability)
  - Independence: A and B are said to be independent events if $P(A \cap B) = P(A)P(B)$, or
    - conditional independence: $P(A \mid B) = P(A)$ or, $P(B \mid A) = P(B)$



## Probability Theory Brief Review

- **Product Law**:
  - Let A and B be events and assume $P(B) \neq 0$, then $P(A \cap B) = P(A \mid B)P(B)$.
- **Random Variable**: In probability, it is a **function** from $\Omega$ to a real number. Because the outcome of the experiment with sample space $\Omega$ is random, the number produced by the function is also random as well.
  - Consider an experiment in which a coin is flipped three times, and the sequence of heads and tails is observed as: $\Omega = \{hhh, hht, htt, hth, ttt, tth, thh, tht\}$.
  - Define the random variables such as
    - (1) the total number of heads,
    - (2) the total number of tails, and
    - (3) the number of heads minus the number of tails.
  - Each of these is a real-valued function defined on $\Omega$. In other words, each of them is a rule that assigns a real number to every point $w \in \Omega$.



## Probability Theory Brief Review

- **Bayes Rule**:
  - defined as: Let A and $B_1, \ldots, B_n$ be events where the $B_i$ are disjoint, $\cup_{i=1}^{n} B_i = \Omega$, and $P(B_i) > 0$ for all i, Then,
  - $P(B_j \mid A) = \dfrac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$
  - Where $\sum_{i=1}^{n} P(A \mid B_i)P(B_i) = P(A)$ is called evidence or marginal distribution of joint probability of A and B over B.
  - If let $P(B_j)$ as the **prior** probability, and $P(A \mid B_j)$ as the likelihood function, then the **posterior** probability is given by: $P(B_j \mid A) = \dfrac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$. Since the evidence P(A) does not change with B, so $P(B_j \mid A) \propto P(A \mid B_j)P(B_j)$
  - Based on above expression, we often say '$posterior \propto likelihood \times prior$' — This is a very important concept in machine learning, particularly in *generative* approaches.



## Probability Theory Brief Review

- **Expected Values**:
  - **Expectation:** If X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, g(X) can be considered a random variable, the expected value of g(X), denoted by $E(g(X))$ is
    - $E(g(X)) = \sum_{x \in X} g(x)p_X(x)$, provided that $\sum_{x \in X} |g(x)| p_X(x) < \infty$. If the sum diverges, the expectation is undefined.
    - For continues random variables: $E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$
  - **Variance**: $Var(X) = E\{[X - E(X)]^2\}$ or $Var(X) = E[X^2] - [E(X)]^2$

## Probability Theory Brief Review

- **Common probability distributions**:
  - **Discrete distribution:**
    - *Bernoulli distribution:* $p(x) = \begin{cases} p & \textbf{if } p = 1 \\ 1-p & \textbf{if } p = 0 \end{cases}$, or $p(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if x = 0 or x=1} \\ 0 & \text{otherwise} \end{cases}$.
  - **Continuous distribution**:
    - Uniform: (where $a < b$): equal probability density to every value between $\boldsymbol{a}$ and $\boldsymbol{b}$ on the real line. $f(x) = \begin{cases} \frac{1}{b-a} & \text{if a } \le x \le \text{b} \\ 0 & \text{otherwise} \end{cases}$
    - Normal distribution or Gaussian distribution: $f(x) = \dfrac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
    - It depends on two parameters: $\mu$, called **mean**, and $\sigma$, called **variance**.



# Assignment 1

- Assignment 1 worth 10%, and is about reviews of mathematics and Python programming. Very easy (bonus!).
- Google Colab is a Jupiter Notebook running in the cloud.
- Copy and download my Colab to your Google drive (Important note: Don't modify my Colab notebook, otherwise other classmates will see your work.)
- Working on your copy of the Colab notebook. Don't forget to add your name and student id in it.
- After finishing it, share it with me (only me, do not share your work with others.)
- All programming exercises MUST be running correctly in Colab without any errors and exceptions. If your code cannot run at all, and I cannot see any kind of outputs, you receive no grade points for that part.
- Before you submit your Colab notebook, make sure to leave the outputs (results) of the functions in the notebook. I ONLY review the outputs of your functions or the final results.
- The assignment due at Feb 5th, 2021. It is an individual assignment.
- Just remind you to beware of academic integrity and responsible behaviour.
- Academic dishonesty or academic misconduct is cheating. The minimum penalty is Failing grade (F) for assignment/project.
- The consequences for any of academic dishonesty can be very serious based on university's regularization.



# Any questions?
# Next section, the lab