Since 2004
UET
ĐẠI HỌC CÔNG NGHỆ, ĐHQGHN
VNU-University of Engineering and Technology

Since 1906
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi

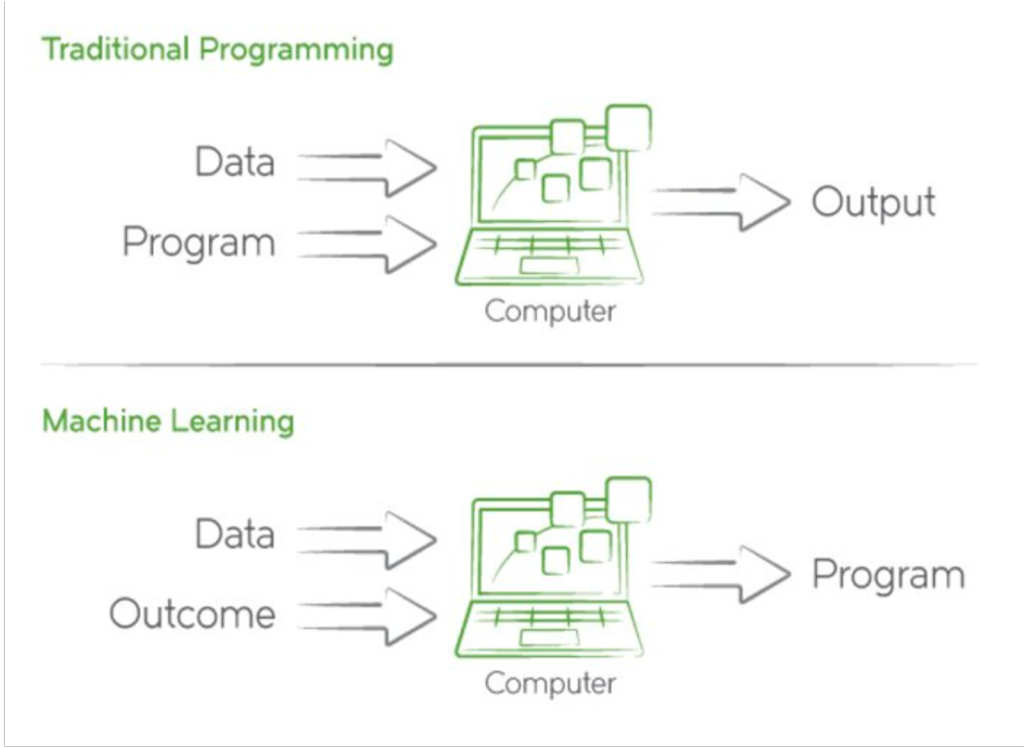
# INT3405 - Machine Learning

## Lecture 2: General Concepts for ML

**Duc-Trong Le & Viet-Cuong Ta**

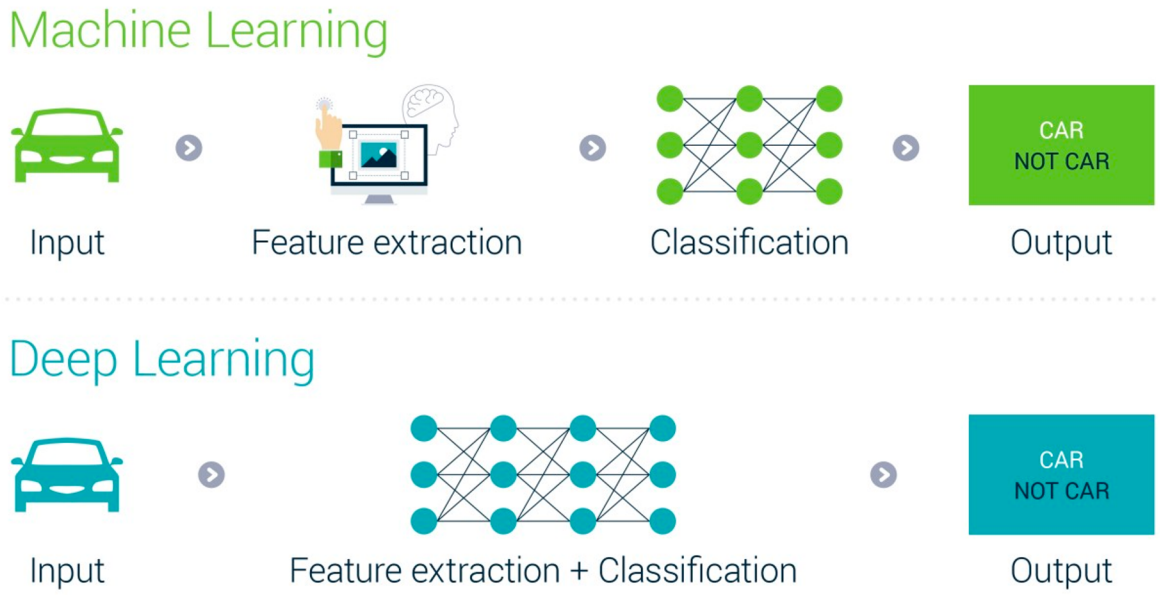
**Hanoi, 09/2023**

# Recap: Traditional Programming vs Machine Learning



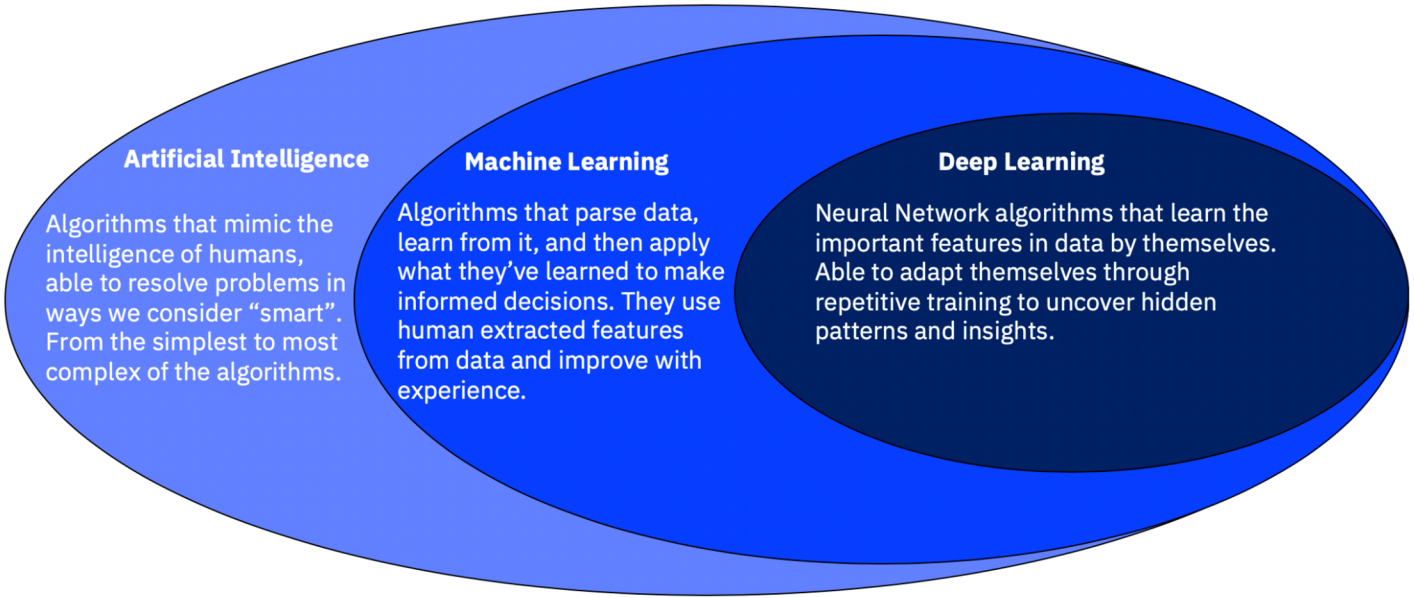

https://images.techhive.com/images/article/2017/05/traditional-programming-vs-machine-learning-100723299-large.jpg

# Recap: Machine Learning vs. Deep Learning



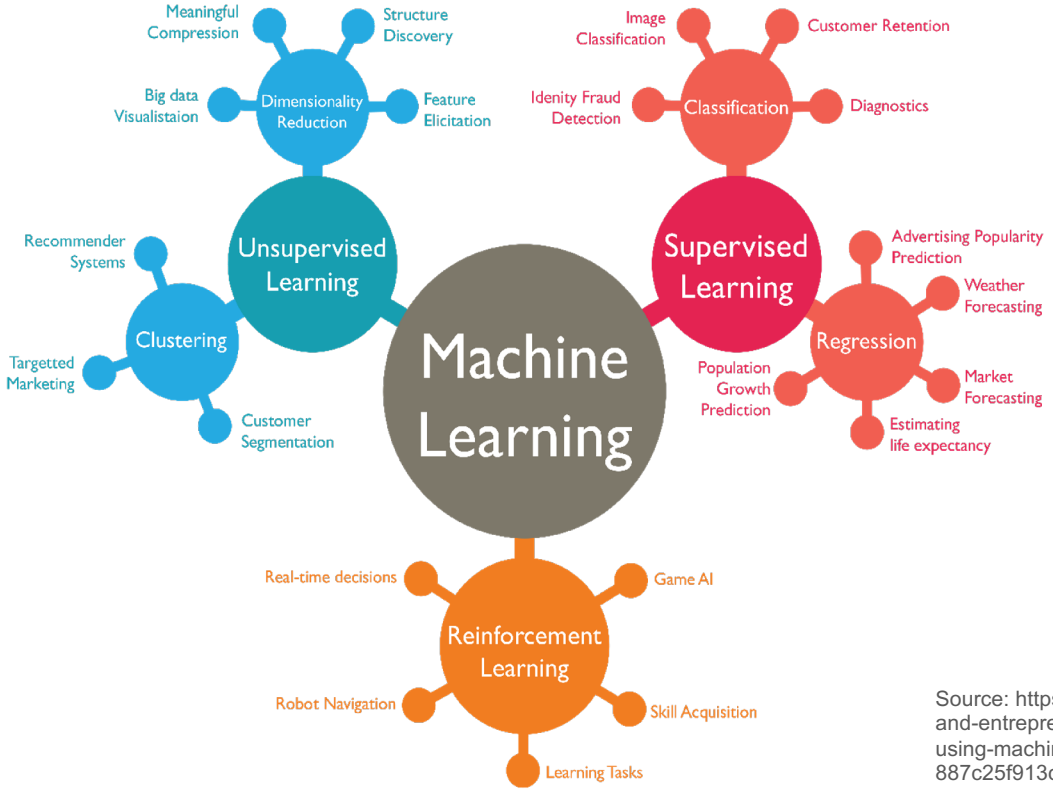

Source: https://www.linkedin.com/pulse/lets-understand-difference-between-machine-learning-vs-gauri-bapat

# Recap: Machine Learning vs. Deep Learning vs AI




Source: https://www.ibm.com/blogs/systems/ai-machine-learning-and-deep-learning-whats-the-difference/

# Recap: Types of Machine Learning




Source: https://medium.com/marketing-and-entrepreneurship/10-companies-using-machine-learning-in-cool-ways-887c25f913c3

# Outline

- Statistics - Probability
- Typical Data Distribution
- Typical Measurements
  - Entropy, Cross Entropy
  - Mutual Information
  - Kullback-Leibler Divergence
- Learning Theory

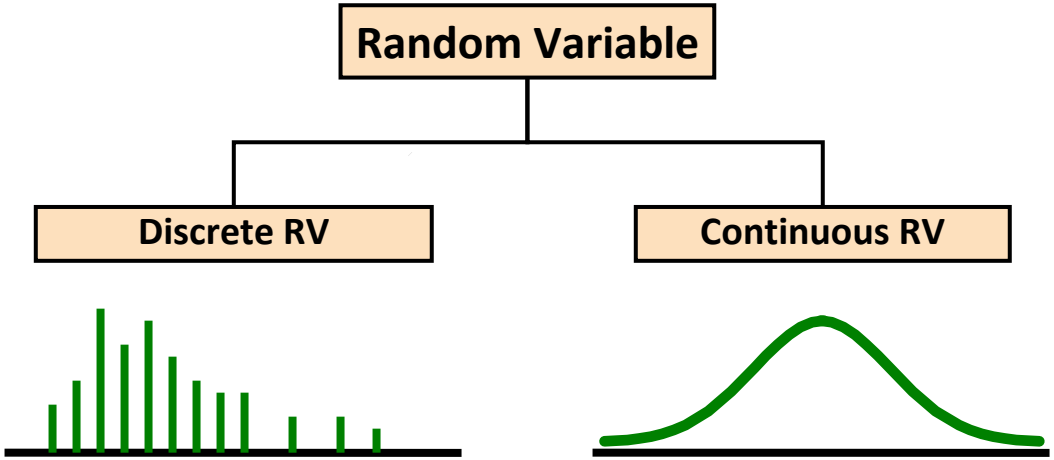
# What animal is it?

A cat?
or a tiger?

# Game: Toss a coin

# Random Variable (RV)

- Random variable (RV) is a variable whose possible values are numerical outcomes of a random phenomena.



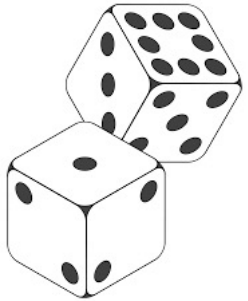
# Discrete Random Variable (DRV)

- Let toss the coin twice, and X is the number of the head occurs.

How many values of X ?

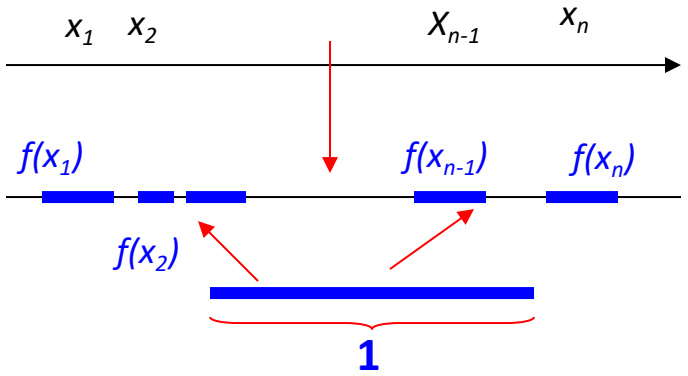
- Let toss the dice twice, and X is the sum of the dice

How many values of X?



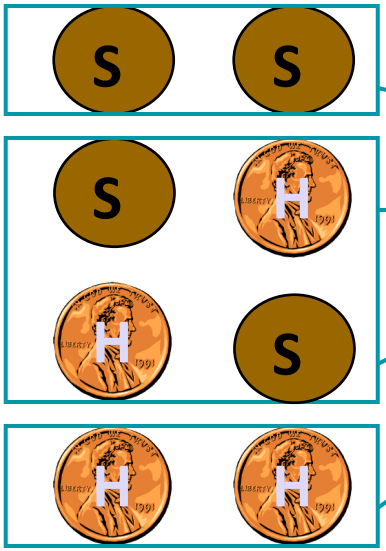
# DRV - Probability Density Function (1)

- A discrete random variable X whose values in $x_1, x_2, \ldots, x_n$.
- Probability Density Function: $f(x_i) = P(X = x_i)$
- Denotation: $p_i = f(x_i) = P(X=x_i)$
- Conditions:
  - $f(x_i) \geq 0$
  - $\sum_{i=1}^{n} f(x_i) = 1$



# DRV – Probability Density Function (2)

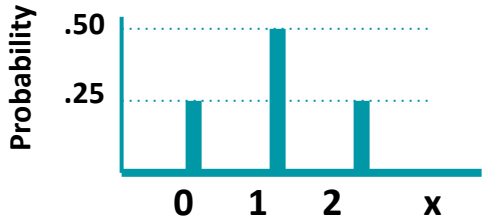
Example: Toss two coins, X is denoted as the occurrences of head.




There are 4 possible outcomes

PDF

| x | P(x) |
|---|---|
| 0 | 1/4 = .25 |
| 1 | 2/4 = .50 |
| 2 | 1/4 = .25 |



# Continuous Random Variable (CRV)

- Its value space is $\mathbb{R}$ or a subset of $\mathbb{R}$.

Weight & Height

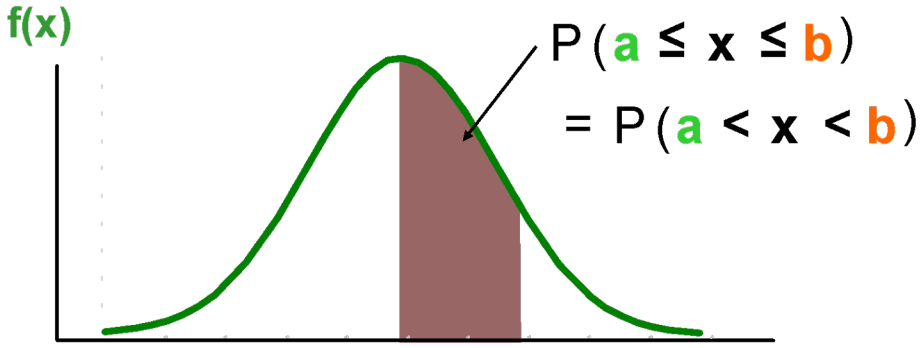

Time to
complete a task

# CRV – Probability Density Function (1)

- f(x) is the probability density function of a given continuous random variable X, if:

$$i)\ f(x) \geq 0 \quad \forall x$$

$$ii)\ \int_{-\infty}^{+\infty} f(x)dx = 1$$




$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

# Cumulative Probability Distribution (1)

- Consider a random variable X, its cumulative probability function F(x) is defined as follows:

$$F(x) = P(X \leq x)$$

- The probability of X in (a,b] as:

$$P(a < X \leq b) = F(b) - F(a)$$

# Cumulative Probability Distribution (2)

1) $0 \le F(x) \le 1$

2) F(x) is a non-decreasing function if ∀ a<b, we have F(a) ≤ F(b).

3) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$

$F(+\infty) = \lim_{x \to +\infty} F(x) = 1$

4) Probability density function f(x) = F'(x) as long as the derivative exits

# Cumulative Probability Distribution (3)

- Discrete Random Variable:

$$F(x_0) = P(X \leq x_0) = \sum_{x_i \leq x_0} p_i$$

- Continuous Random Variable:

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx$$

# Expected Value of Random Variable

- Discrete Random Variable:

$$E[X] = \sum_{x} xP(X = x)$$

- Continuous Random Variable:

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

# Expected Value - Properties

1) E[C] = C, C: constant
2) E[CX] = C.E[X]
3) E[X + Y]=E[X] + E[Y]
4) E[XY] = E[X].E[Y] if X and Y are independent
5) Given a function h(x), we have

$$Eh(x) = \sum_{i=1}^{n} h(x_i)\, p_i$$ if X is discrete

$$Eh(x) = \int_{-\infty}^{+\infty} h(x)\, f(x) dx$$ if X is continuous

# Variance of Random Variable

- Mean $\mu = \mathrm{E}[\mathrm{X}]$
- Variance: $\mathrm{Var}[\mathrm{X}] = \mathrm{E}\left[\mathrm{X} - \mathrm{E}[\mathrm{X}]\right]^2 = \mathrm{E}[\mathrm{X}^2] - (\mathrm{E}[\mathrm{X}])^2$
- Standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{Var}[\mathrm{X}]}$
- Discrete random variable:

$$\mathrm{Var}[\mathrm{X}] = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mu)^2 \mathrm{p_i}$$

- Continuous random variable:

$$\mathrm{Var}[\mathrm{X}] = \int_{-\infty}^{+\infty} (\mathrm{x} - \mu)^2 \mathrm{f}(\mathrm{x}) \mathrm{dx}$$

# Variance - Properties

1) $\mathrm{Var}[C] = 0$, C: constant
2) $\mathrm{Var}[CX] = C^2\,\mathrm{Var}[X]$

   $\mathrm{Var}[X + C] = \mathrm{Var}[X]$
3) $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ if X and Y are independent.

# Three Rules of Probability

- Basic law

$$p(X = x) \geq 0, \quad \sum_{X=x} p(X = x) = 1.$$

- Sum rule

$$p(X) = \sum_Y p(X, Y).$$

- Product rule

(Bayesian Rule)

$$p(X, Y) = p(Y|X)p(X).$$

# Joint Probability

| | |
|---|---|
| both are discrete | $\sum_{x,y} p(x,y) = 1$ |
| both are continuous | $\int p(x,y)dxdy = 1$ |
| $x$ is discrete, $y$ is continuous | $\sum_{x} \int p(x,y)dy = \int \left( \sum_{x} p(x,y) \right) dy = 1$ |

# Marginal Probability

- Discrete random variable:

$$p(x) = \sum_y p(x, y)$$

$$p(y) = \sum_x p(x, y)$$

- Continuous random variable:

$$p(x) = \int p(x, y) dy$$

$$p(y) = \int p(x, y) dx$$

# Conditional Probability

- Bayesian Rule: $p(y|x)p(x) = p(x|y)p(y)$
- Conditional Probability:

$$p(x|y=9) == \frac{p(x,y=9)}{\sum_x p(x,y=9)} = \frac{p(x,y=9)}{p(y=9)}$$

# Posterior-, Prior Probability & Likelihood

- A random variable $X$ could be predicted via the parameters $\theta$:

$$p(\theta|X) = \frac{p(X|\theta)}{p(X)} p(\theta)$$

$$p(\theta|X) \propto p(X|\theta) \times p(\theta)$$

Posterior probability

Likelihood

Prior probability

# Independence & Dependence

- If x, y are independent, so:

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- Bayesian Rule:

$$p(x, y) = p(x|y)p(y) = p(x)p(y)$$

# Exercises:

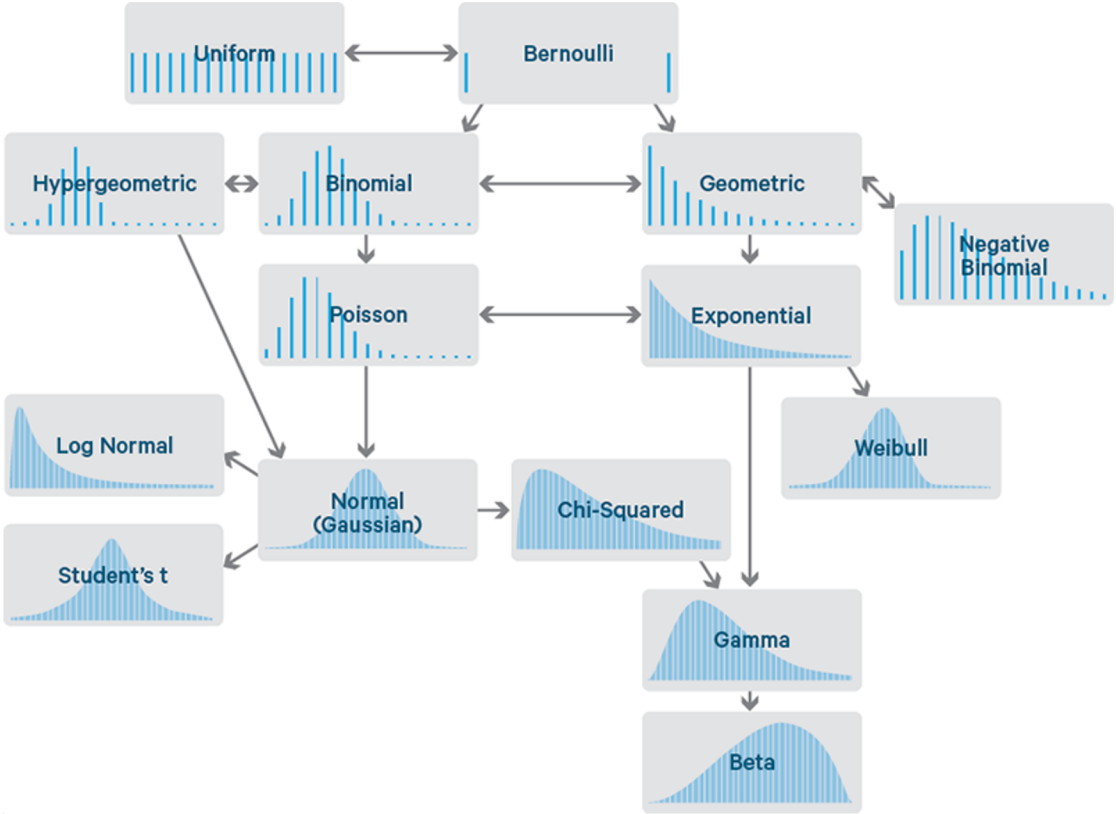
- What is the probability that the total of two dice will be greater than 9, given that the first dice is a 5?

- Suppose there is a school having 60% boys and 40% girls as students. The girls wear trousers or skirts in equal numbers; all boys wear trousers. An observer sees a (random) student from a distance; all the observer can see is that this student is wearing trousers. What is the probability this student is a girl?

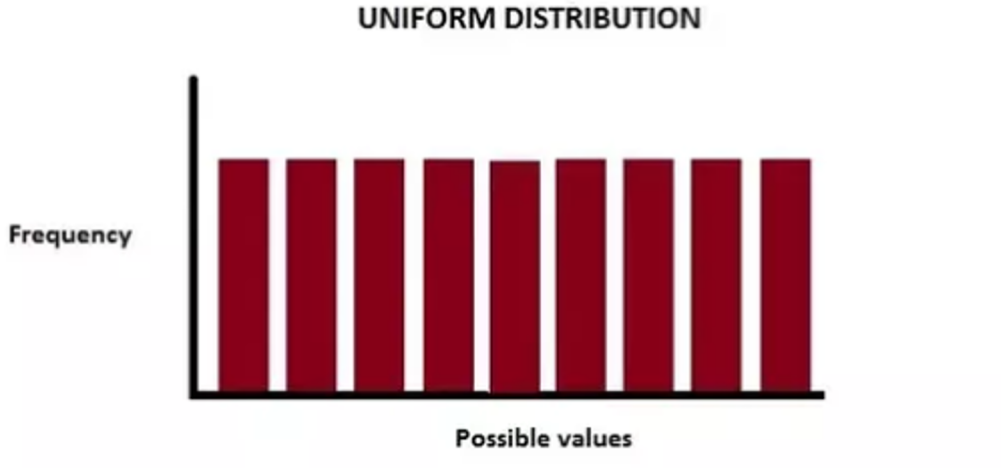
# Outline

- Statistics - Probability
- Typical Data Distribution
- Typical Measurements
  - Entropy, Cross Entropy
  - Mutual Information
  - Kullback-Leibler Divergence
- Learning Theory

# Typical Data Distributions



# Uniform Distribution



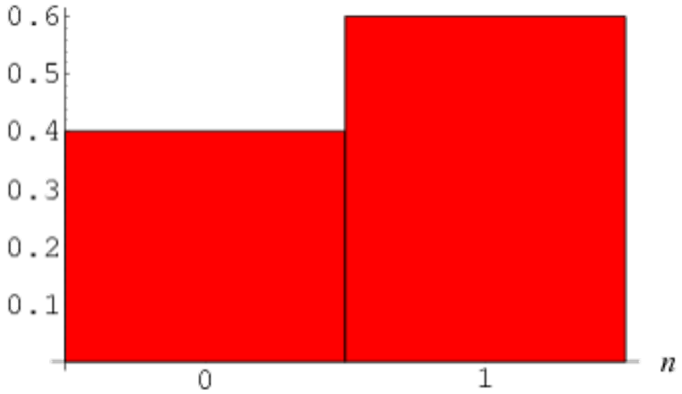

P(X) = 1 / total number of possible outcomes

# Uniform Distribution – Example

The probability of doomsday being Monday is 1/7


The probability that a numeric letter appearing in 500k VNĐ note is 1/10

# Bernoulli Distribution



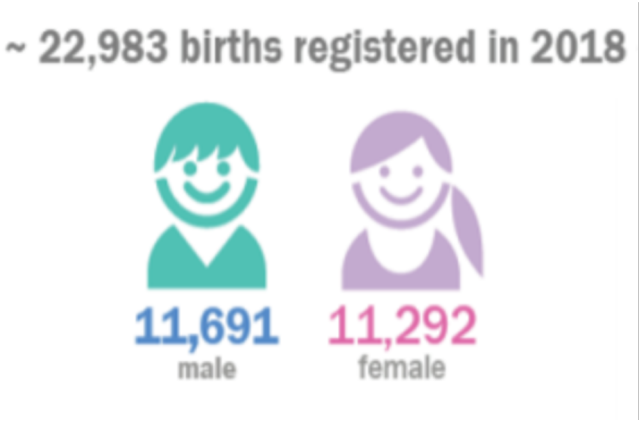

$$=> \text{Bern}_x[0.6]$$

$$P(x) = \text{Bern}_x[\lambda] = \lambda^x(1-\lambda)^{1-x}$$

$$x \in \{0,1\}, \lambda \in [0,1]$$

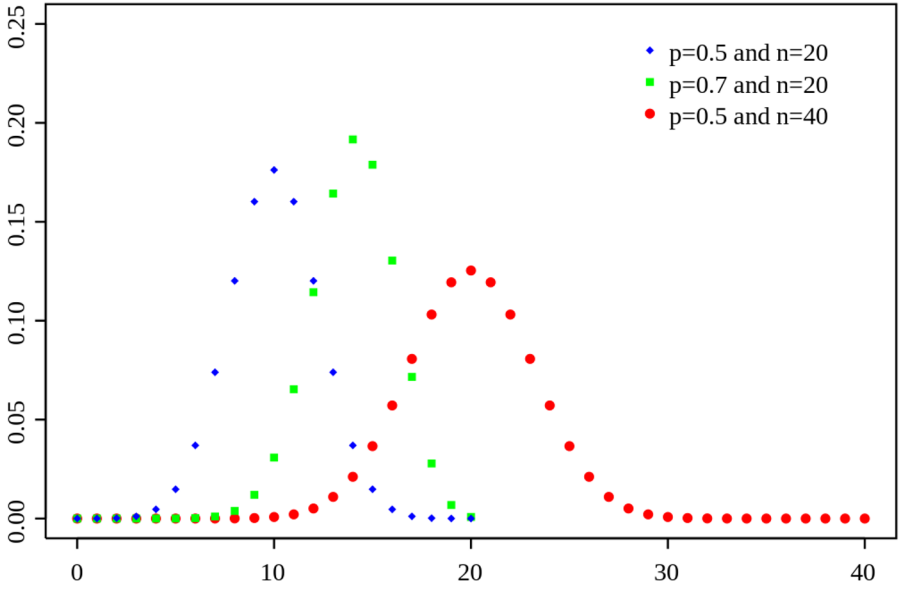
# Bernoulli Distribution – Example

The probability of head/tail $\lambda \approx 0.5$




The probability of male/female baby

# Binary Distribution




$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{voi } k = 0, 1, 2, \ldots, n$$

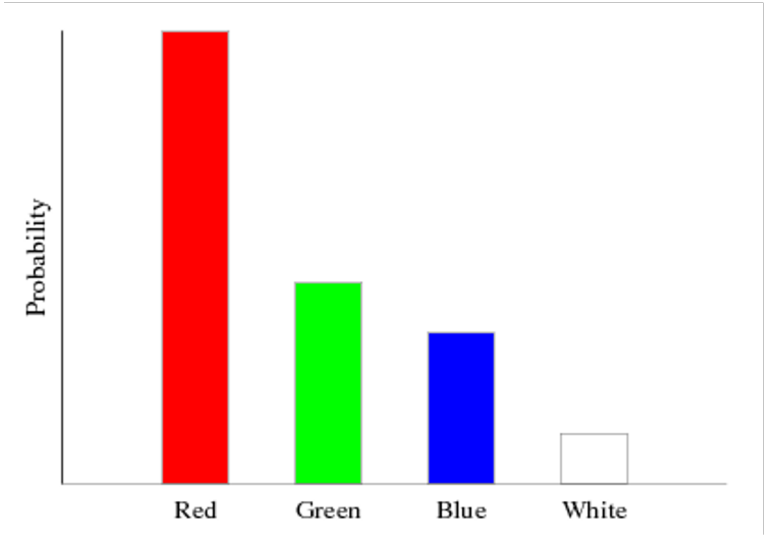
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

# Binary Distribution – Example

The probability getting X times of head in 10 times tossing a coin

| X | P(X) |
|---|---|
| 0 | 0.000977 |
| 1 | 0.009766 |
| 2 | 0.043945 |
| 3 | 0.117188 |
| 4 | 0.205078 |
| 5 | 0.246094 |
| 6 | 0.205078 |
| 7 | 0.117188 |
| 8 | 0.043945 |
| 9 | 0.009766 |
| 10 | 0.000977 |

# Categorical Distribution

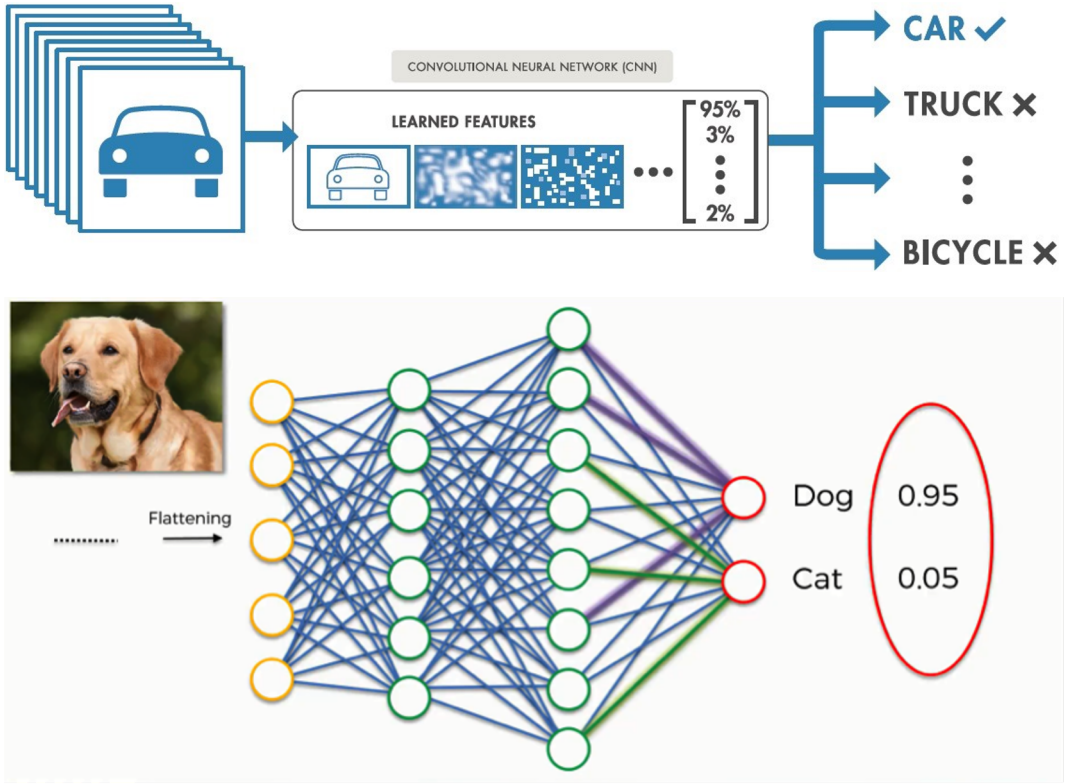



$$P(x) = \mathrm{Cat}_x[\boldsymbol{\lambda}]$$

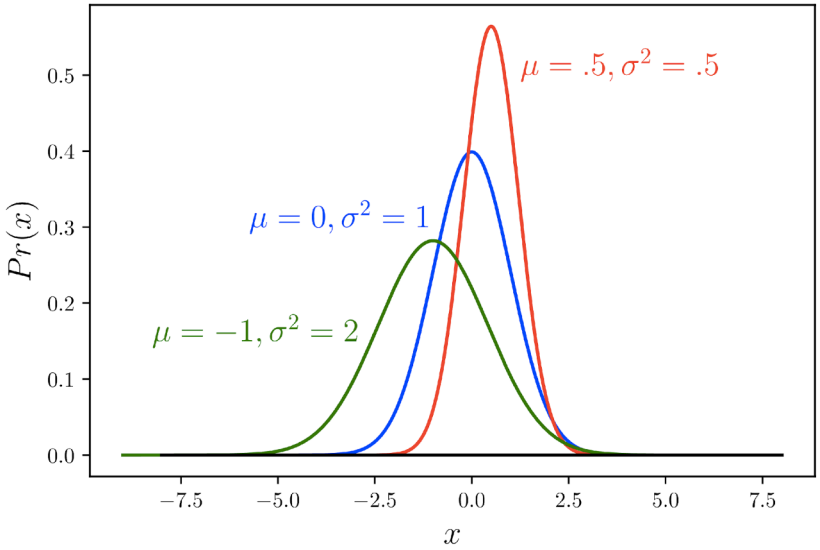
$$\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_{|C|}\}; \lambda_k \in [0,1]$$

It is general form of Bernoulli distribution with more than 2 possible outcomes

# Categorical Distribution - Example



# Univariate Normal Distribution

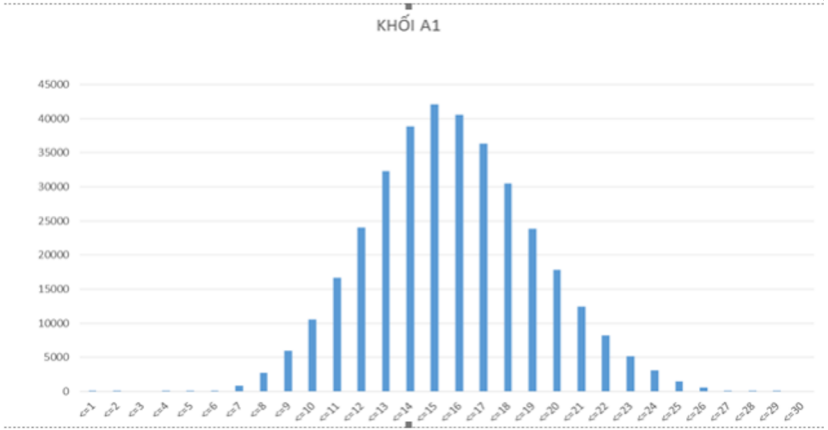



Normal distribution = Gaussian distribution

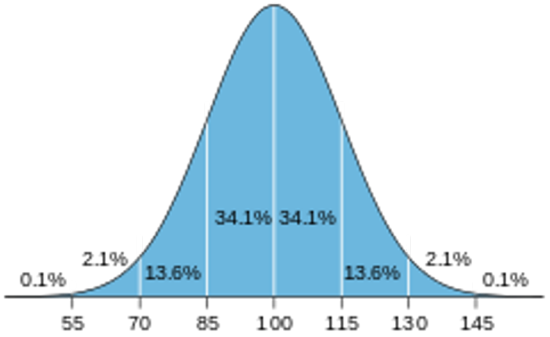
$$P(X) = \text{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Mean $\mu$, Variance $\sigma^2$

# Univariate Normal Distribution - Example

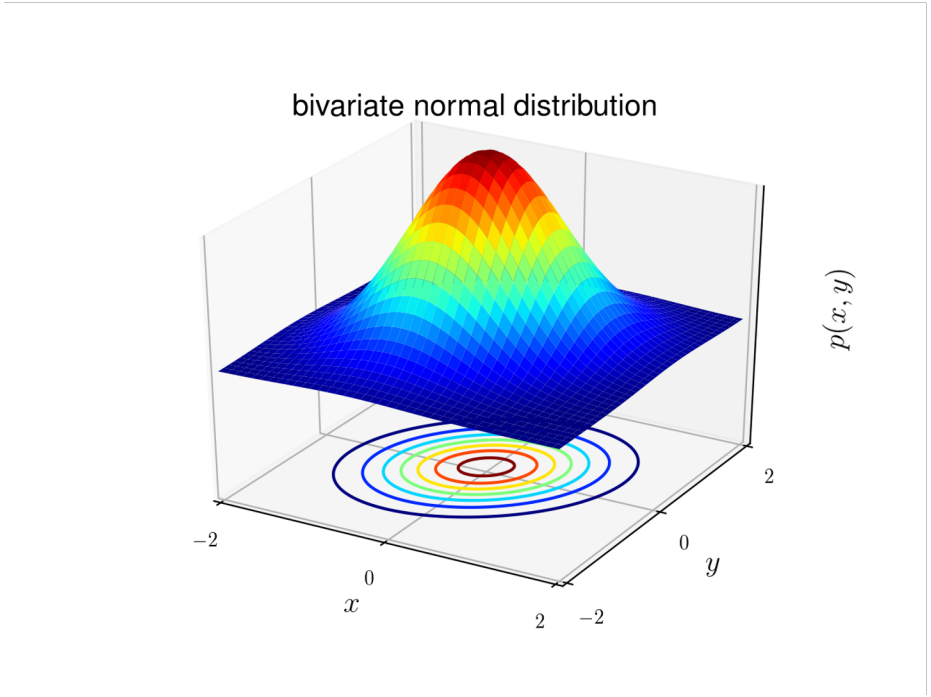



Score Distribution of a University Entrance Exam



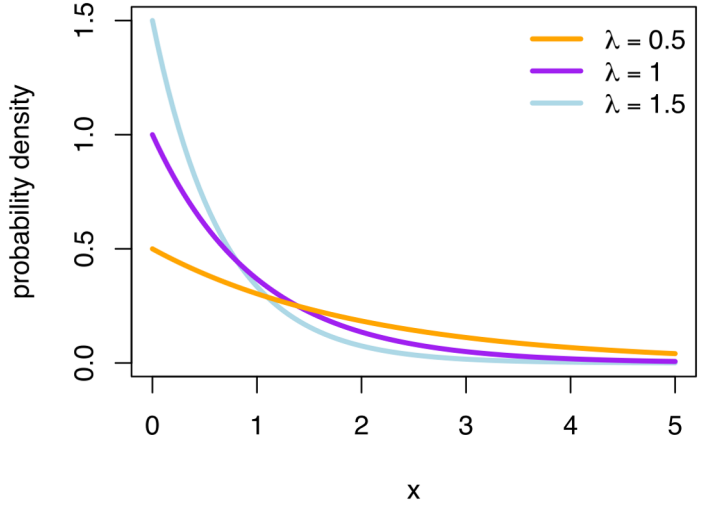

IQ Distribution

# Multivariate Normal Distribution



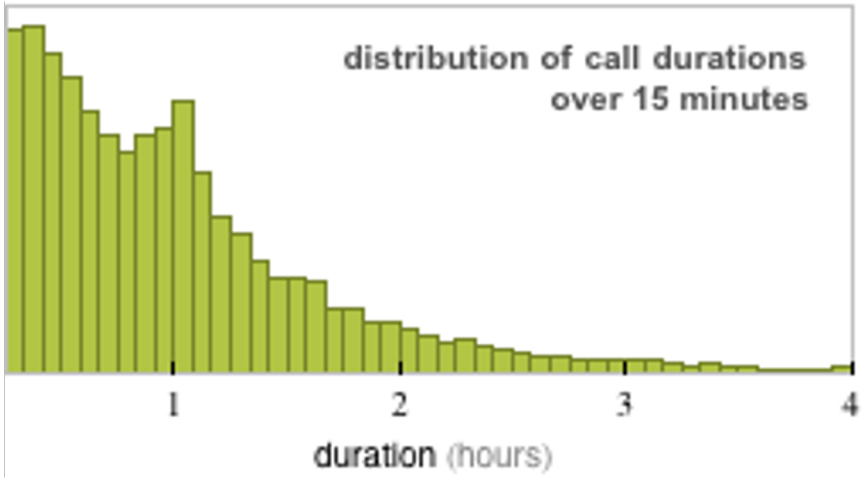

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

# Exponential Distribution




$$P(x) = \text{Exp}(\lambda) = \lambda e^{-\lambda x} \quad x \geq 0$$

# Exponential Distribution - Example




Time distribution of call duration over 15 minutes in a Telecom company

# Outline

- Statistics - Probability
- Typical Data Distribution
- Typical Measurements
  - Entropy, Cross Entropy
  - Mutual Information
  - Kullback-Leibler Divergence
- Learning Theory

# Typical Measurement - Entropy

- Measure the ‘uncertainty’ or ‘surprise’ in data
- Discrete random variable:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

- Continuous random variable:

$$h(X) = -\int_{\mathcal{X}} f(x) \log f(x)\, dx$$

# Typical Measurement – Mutual Information

- Measure the mutual dependence between two random variables.

$$
\begin{aligned}
\mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X|Y) \\
&\equiv \mathrm{H}(Y) - \mathrm{H}(Y|X) \\
&\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\
&\equiv \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X)
\end{aligned}
$$

# Typical Measurement – Cross Entropy

- Measure the difference between two distributions
- Discrete random variable:

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

- Continuous random variable:

$$H(p, q) = -\int_{\mathcal{X}} P(x) \log Q(x)\, dr(x)$$

# Typical Measurement – KL Divergence

- Measure the difference between two distributions
- Discrete random variable:

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

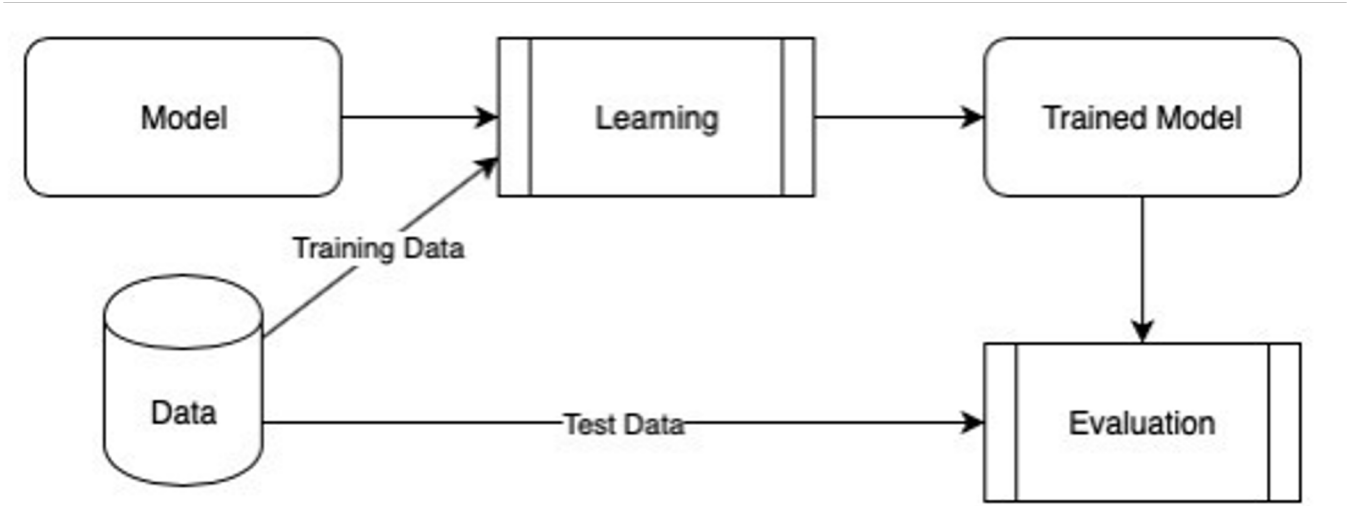
- Continuous random variable:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

# Outline

- Statistics - Probability
- Typical Data Distribution
- Typical Measurements
  - Entropy, Cross Entropy
  - Mutual Information
  - Kullback-Leibler Divergence
- Learning Theory

# Inductive Learning



# The Statistical Learning Framework

The formal definition of the learner's input:

- **Domain Set**: An arbitary set $\mathcal{X}$. $\mathcal{X}$ represents the space contain the objects we want to learn
- **Label Set**: A set of definitions $\mathcal{Y}$ which defines on $\mathcal{X}$. In the binary case, think of $\{0, 1\}$ or $\{-1, 1\}$ or $\{class_1, class_2, class_3\}$ or a *Real Value* if the task is regression task. It can be extended to $n$ classes.
- **Training data**: A sequence $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ in the domain $\mathcal{X} \times \mathcal{Y}$.

In general, $S$ is the specific definition of the experience E in **supervised classification**.

# The Statistical Learning Framework

The formal definition of the learner's output:

- A prediction rule:

$$h : \mathcal{X} \to \mathcal{Y} \tag{1}$$

- $h$: classifier, predictor, hypothesis or mapping function. For example, $h$ is the linear function with thresholding in previous lesson.
- Given an algorithm $A$, we denote $A(S)$ is the set of classifier/hypothesis generated by apply $A$ on $S$

$h$ could be more general than the definition of a function
One example is: h is the stochastic function

# The Statistical Learning Framework

A simple **data-generation pipeline**:

- Assume we have a probability distribution $\mathcal{D}$ over $\mathcal{X}$,
- Given $\mathcal{D}$, we draw a sample $x_i \sim \mathcal{D}$
- Assume we have a **correct** label function $c$:

$$c : \mathcal{X} \to \mathcal{Y} \tag{2}$$

that $y_i = c(x_i)$, and $S = ((x_1, y_1), \ldots, (x_m, y_m))$

Notes:

1. $c$ is function we want to recover. However, we do not know $\mathcal{D}$ and $c$, our learning algorithm $A$ can only access $\mathcal{X}$, $\mathcal{Y}$, $S$
2. Taken into account that, every time the **data-generation pipeline** runs, we can get a new training data $S$

# The Statistical Learning Framework

Given domain set $\mathcal{X}$ and label set $\mathcal{Y}$, the data distribution $\mathcal{D}$ and the labeling function $c$.

The error of a prediction rule $h : \mathcal{X} \to \mathcal{Y}$ is defined as:

$$L_{\mathcal{D},c}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq c(x)] \tag{3}$$

However, we do not know $\mathcal{D}$ and $c$ in practice

# The Statistical Learning Framework

The Empirical Risk Minimization (ERM) framework learns a learner

$$h_S : \mathcal{X} \to \mathcal{Y}$$

that minimize the **the training error** without knowing $\mathcal{D}$ and $c$.

The **training error** is defined as

$$L_S(h) \overset{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} \tag{4}$$

Note that: we use $h_S$ to emphasize our learners depends on $S$, a specific $S$

# The Statistical Learning Framework

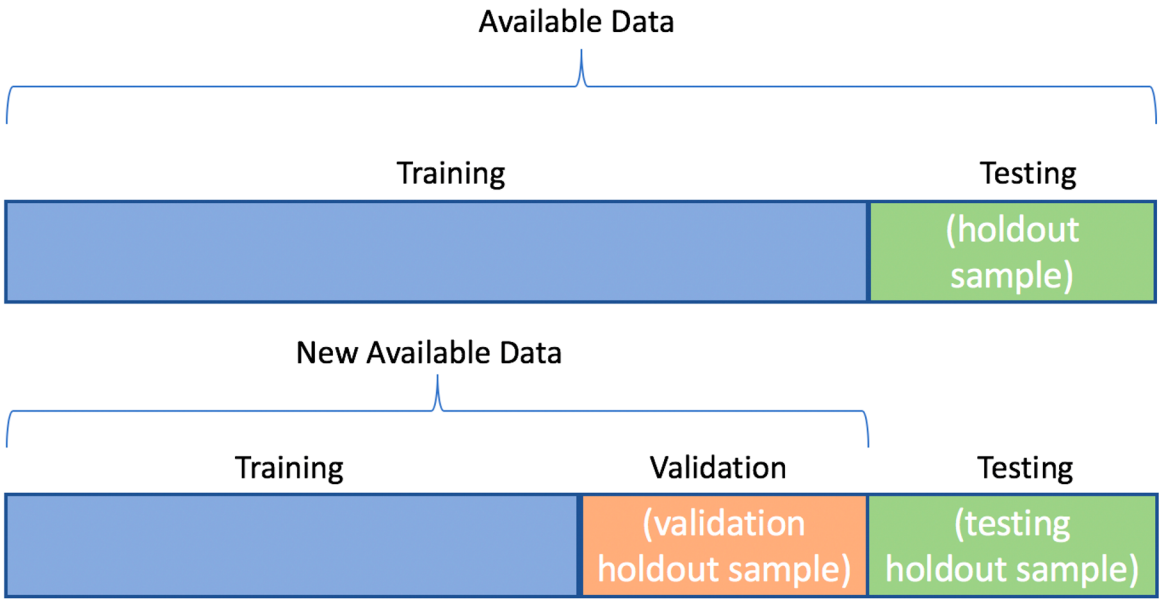
Assume we have two dataset $S_1$, $S_2$, which both sample from $\mathcal{D}$
We have the error on $S_1$, $S_2$ is defined as:

$$L_{S_1}(h) \stackrel{\text{def}}{=} \frac{|\{(x_i, y_i) \in S_1, i \in [m_1] : h(x_i) \neq y_i\}|}{m_1} \tag{5}$$

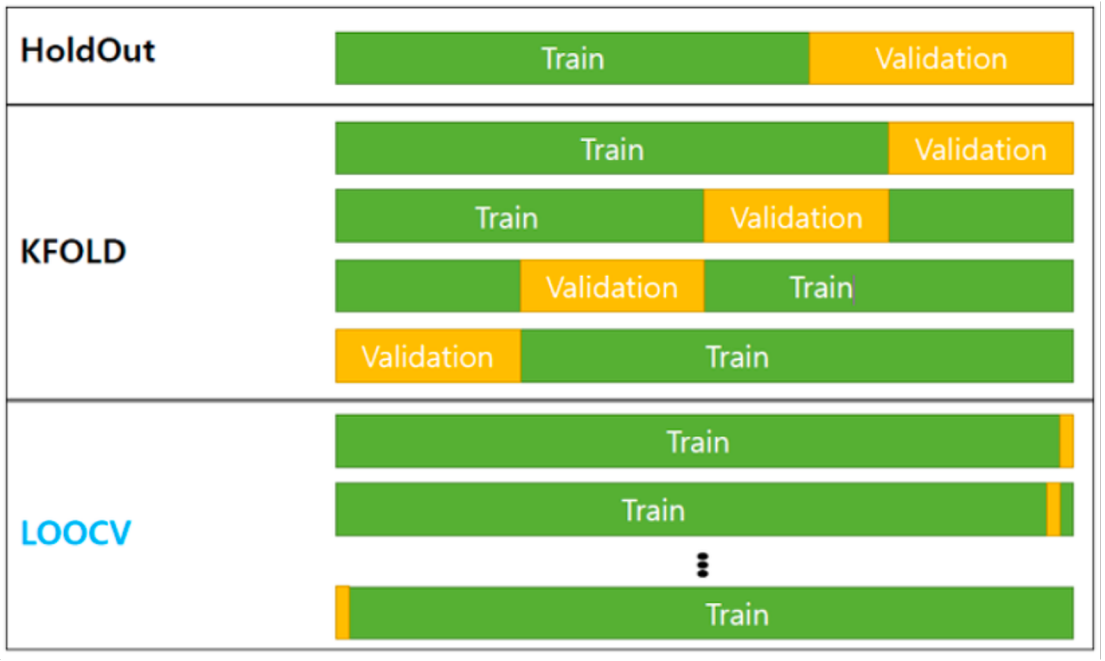
$$L_{S_2}(h) \stackrel{\text{def}}{=} \frac{|\{(x_i, y_i) \in S_2, i \in [m_2] : h(x_i) \neq y_i\}|}{m_2} \tag{6}$$

1. Overfitting: we have $L_{S_1}(h)$ near 1 and $L_{S_2}(h)$ near 0.5
2. Generalization: we have $L_{S_1}(h) \sim L_{S_1}(h)$

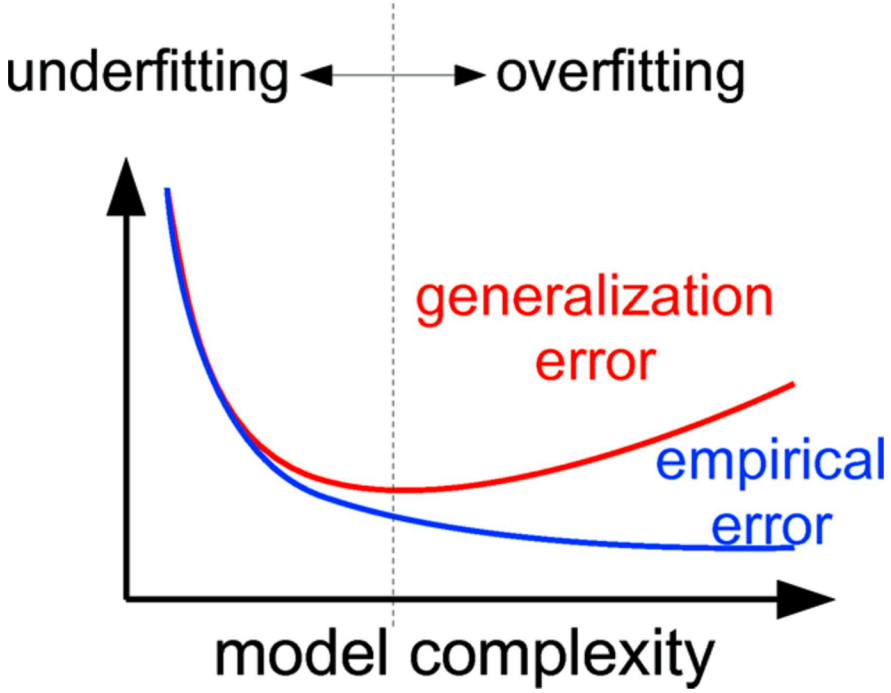
# Train/Test/Validation



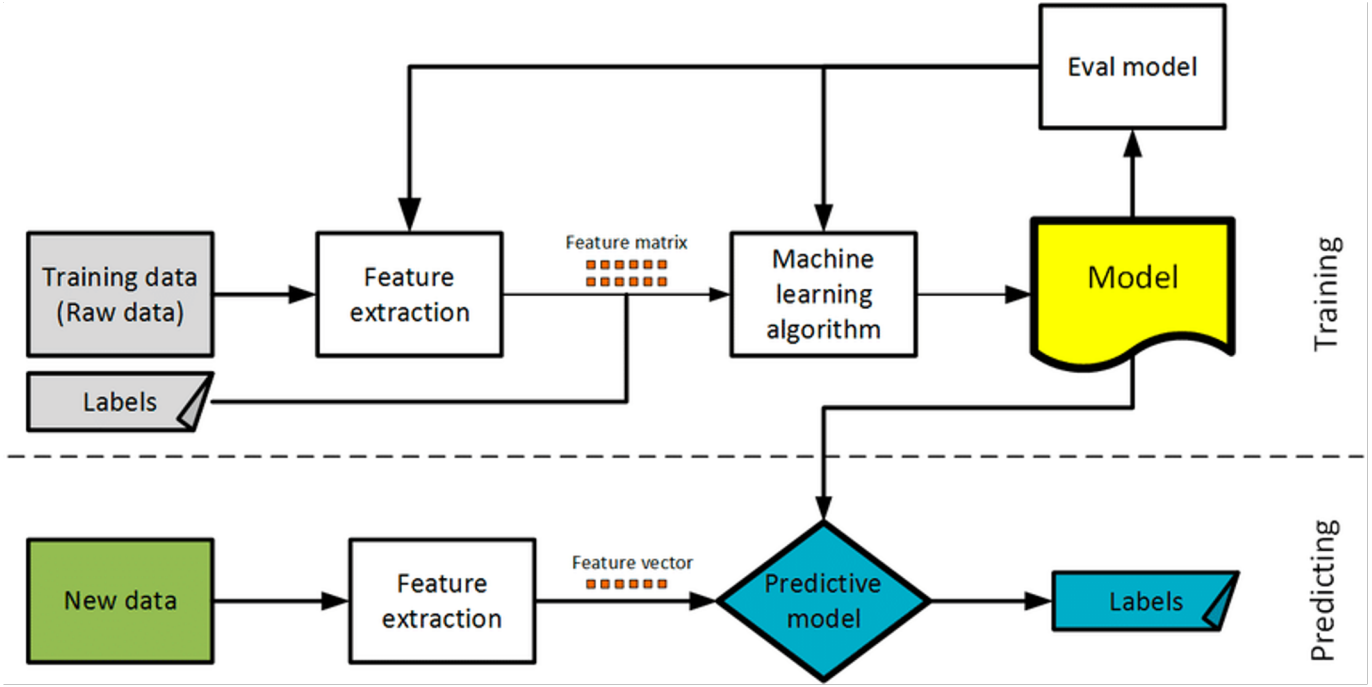
# Cross Validation



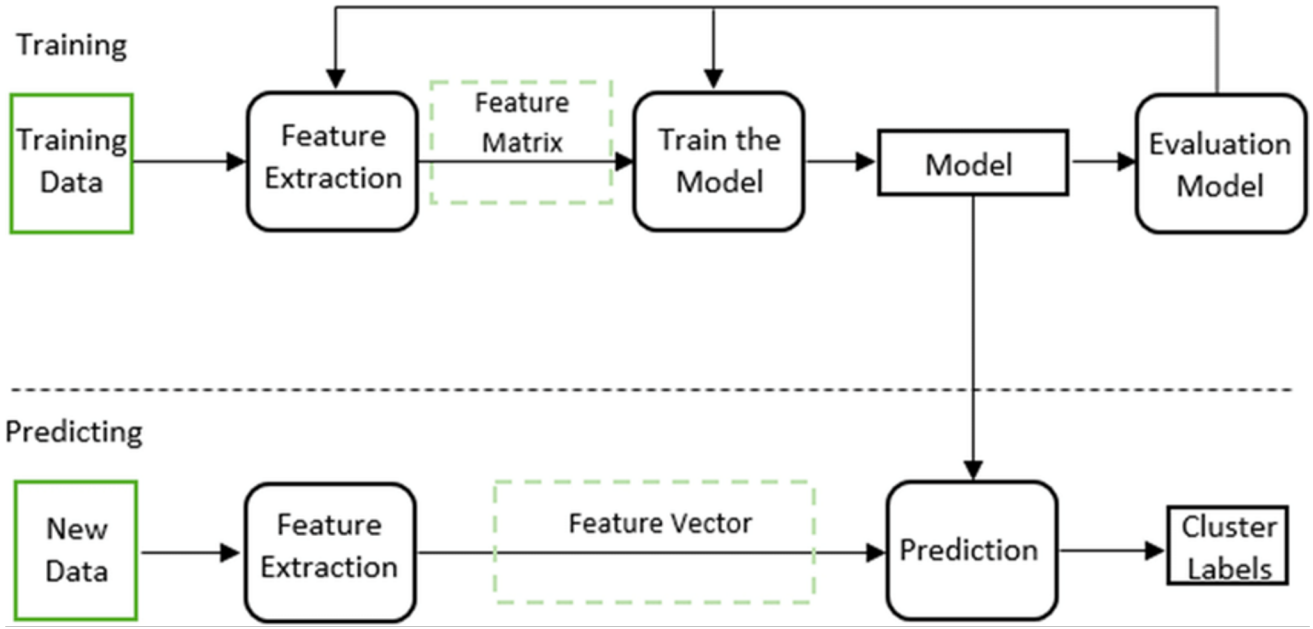
# Generalization- & Empirical Error



# Supervised Learning Workflow



# Unsupervised Learning Workflow



# Summary

- Statistics - Probability
- Typical Data Distribution
- Typical Measurements
  - Entropy, Cross Entropy
  - Mutual Information
  - Kullback-Leibler Divergence
- Learning Theory

# Homework: Parameter Estimation

$X \in \{0, 1\}$ with probability $P(X = 1) = \theta$, written as $X \sim Ber(\theta)$.
We also have $P(X = 0) = 1 - \theta$.

- A biased coin: $\theta =$ probability of head
- Binary classification: $P(y|x) = Ber(\theta(x))$
  $\rightarrow$ probability of class 1 is a function of input

# Homework: Parameter Estimation

Toss a coin (sampling) $N$ times, the number of times heads come up (number 1) is $s$ times, what is the parameter $\theta$ of the coin (Bernoulli distribution)?
An intuitive guess: $\theta = \frac{s}{N}$, why does this number make sense?
Let $x_i \in \{0, 1\}$ is the values from the $i^{th}$ toss.
The probability of the data $D = \{x_1, x_2, \ldots, x_N\}$ under the model $X \sim Ber(\theta)$ is

$$L(\theta) = P(D) = P(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} P(x_i)$$
$$= \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

# Homework: Parameter Estimation

$L(\theta)$ is the *likelihood* of $\theta$ with respect to the dataset $D$
MLE: Find $\theta$ for which $L(\theta)$ is maximized.

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} x_i \log \theta + (1 - x_i) \log(1 - \theta)$$

$$\ell'(\theta) = \sum_{i=1}^{n} x_i/\theta - (1 - x_i)/(1 - \theta) = 0$$

$$\frac{1}{\theta} \underbrace{\sum_{i=1}^{n} x_i}_{s} = \frac{1}{1 - \theta} \underbrace{\sum_{i=1}^{n} (1 - x_i)}_{N-s}$$

$$s(1 - \theta) = (N - s)\theta$$

$$\theta^{MLE} = \frac{s}{N}$$

# Homework: Parameter Estimation

- Unbiased: $\mathbb{E}[\theta^{MLE}] = \theta$
- Variance goes to 0: $\mathbb{V}[\theta^{MLE}] = \theta(1-\theta)/N$
- Consistent: $\mathbb{P}\{|\theta^{MLE} - \theta| \geq \epsilon\} \xrightarrow{n \to \infty} \infty$
- Normality: $\sqrt{N}(\theta^{MLE} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$

# Homework 1

## Extend to Binominal Distribution

The probability of getting exactly $s$ heads in $N$ independent Bernoulli trials of tossing a coin is a Binomial distribution.

$$X \sim Bin(s|N,\theta) \Rightarrow P(X=s) = C_N^s \theta^s (1-\theta)^{N-s}$$

If, after taking the experiments $n$ times, we get the data $D = \{s_1, s_2, \ldots, s_n\}$, then what is the sensible value of $\theta$? (Hint: using MLE)

# Homework 2:

## Extend to Categorical Distribution

$X \in \{1, 2, \ldots, C\} = [\![C]\!]$ with probability $\mathbb{P}(X = k) = \theta_k$ written as $Cat(\theta_1, \ldots, \theta_k)$

$$\sum_{k=1}^{C} \theta_k = 1, \theta_k \geq 0, \forall k$$

- A dice: $\theta_k =$ probability that a throw returns $k$
- Multi-class classification: $\mathbb{P}(y = k|x) = \theta_k(x)$

Exercise: Given data $D = (x_1, x_2, \ldots, x_n) \in [\![C]\!]^n$, find MLE estimate of $\theta_k, k \in [\![C]\!]$. Hint: use Lagrange multiplier method

# Thank you

**Email me**
**trongld@vnu.edu.vn**