Motivation and Basics

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

- Motivate the study on
  - Machine learning, AI, Datamining….
  - Why? What?
  - Overview of the field
- Short questions and answers on a story
  - What consists of machine learning?
  - MLE
  - MAP
- Some basics
  - Probability
  - Distribution
  - And some rules…

BASICS

- 확률/param 추정방법
MLE : prior knowledge X. Data 많아지면 - frequentist
MAP : " O. - Bayesian

What we just saw is…

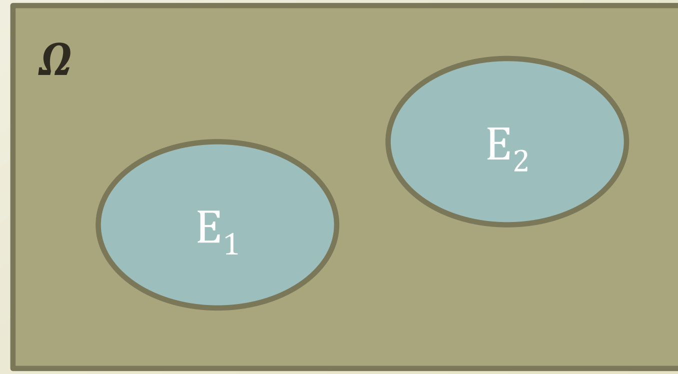
Bayes says

Why not use the Beta distribution?

← From the knowledge of probability, distribution, and statistics

- A struggle
  - Billionaire
    - To earn money by analyzing a small dataset out of huge possibilities
  - You
    - To give the billionaire the best probable and approximate answers from the small dataset
  - Bayes
    - To convince you that the prior knowledge can be incorporated to the answers
- Eventually
  - Trying to find out the nature of the thumbtack game
  - The key is the probability of the thumbtack outcome, either head or tail
- Underlying knowledge to solve the problem
  - Probability
  - Distribution
  - Some mathematical tricks
- To go further, you need to know these

Probability



- Philosophically, Either of the two
  - Objectivists assign numbers to describe states of events, i.e. counting
  - Subjectivists assign numbers by your own belief to events, i.e. betting
- Mathematically
  - A function with the below characteristics

'확률'이라는 f'n. with 인자 E (event)

$$P(E) \in R \qquad P(E) \geq 0 \qquad P(\Omega) = 1$$

event Cont.

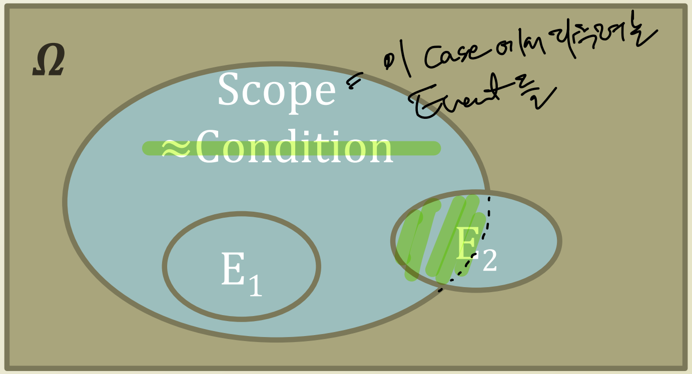
$$P(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i) \text{ when a sequence of mutually exclusive}$$

  - Subsequent characteristics

$$\text{if } A \subseteq B \text{ then } P(A) \leq P(B) \qquad P(\emptyset) = 0 \qquad 0 \leq P(E) \leq 1$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad P(E^C) = 1 - P(E)$$

Conditional Probability



- We often do not handle the universe, Ω
- Somehow, we always make conditions
  - Assuming that the parameters are X, Y, Z,.....
  - Assuming that the events in the scope of X, Y, Z,....
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$ → Scope

  A가 일어날 확률 | B가 참일 때
  - The conditional probability of A given B
- Some handy formula

$$Posterior = \frac{Likelihood \times Prior\ Knowledge}{Normalizing\ Constant}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

$$P(A) = \sum_n P(A|B_n)P(B_n)$$

Nice to see that we can switch the condition and the target event

Nice to see that we can recover the target event by adding the whole conditional probs and priors

Probability Distribution

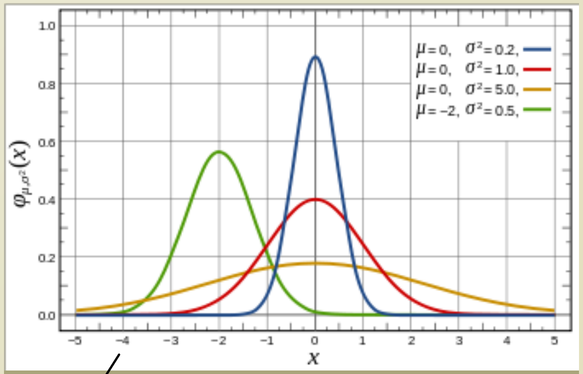
"Mapping".

- Probability distribution ~~assigns~~
  - A probability to a subset of the potential events of a random trial, experiment, survey, etc.
- A function mapping an event to a probability
  - Because we call it a probability, the probability should keep its own characteristics (or axioms)
  - An event can be
    - A continuous numeric value from surveys, trials, experiments…
    - A discrete categorical value from surveys, trials, experiments…
- For example,

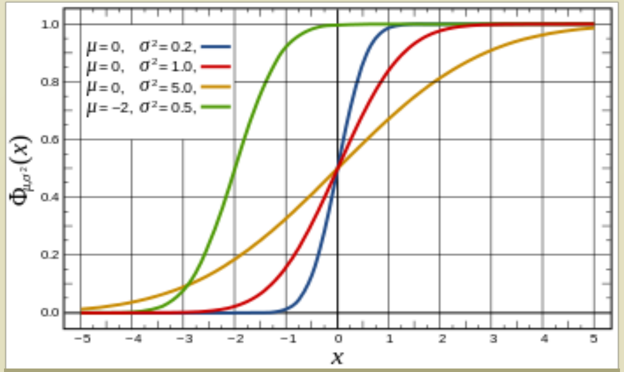
$$f(x) = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}$$

Event

f: a probability distribution function
x: a continuous value
f(x): assigned probs



Probability Density Function (PDF)=$f(x)$

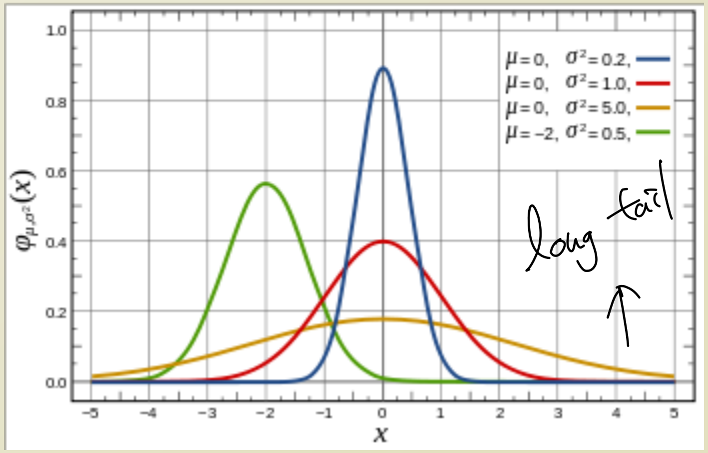


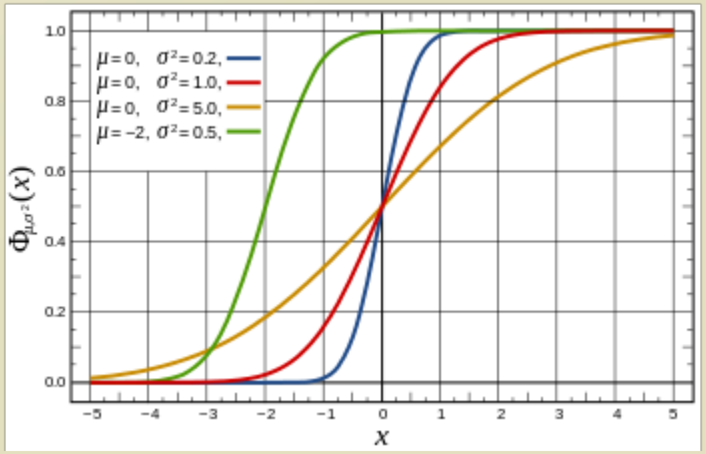
Cumulative Distribution Function (CDF) $=\int_{-\infty}^{x} f(x)\,dx$

Normal Distribution

- Very commonly observed distribution
  - Continuous numerical value

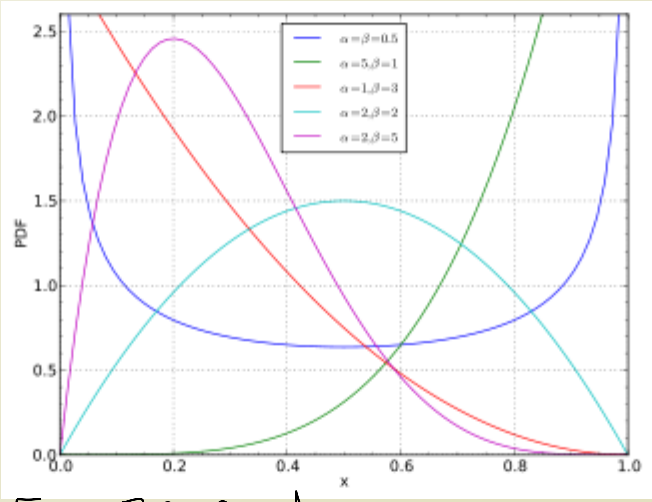
- $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Notation: $N(\mu,\sigma^2)$
- Mean: μ
- Variance: σ^2



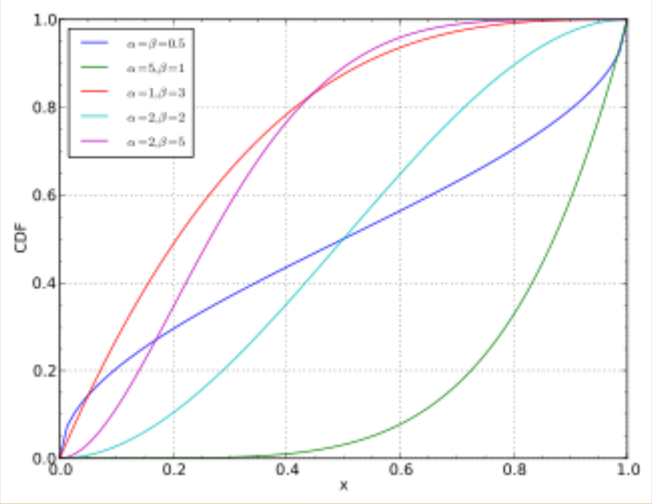


Beta Distribution

- Supports a closed interval
  - Continuous numerical value
  - [0,1]
  - Very nice characteristic
  - Why?
    - Matches up the characteristics of probs
- $f(\theta;\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$
  $\Gamma(\alpha)=(\alpha-1)!, \alpha\in N^{+}$
- Notation: $\text{Beta}(\alpha,\beta)$
- Mean: $\frac{\alpha}{\alpha+\beta}$
- Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

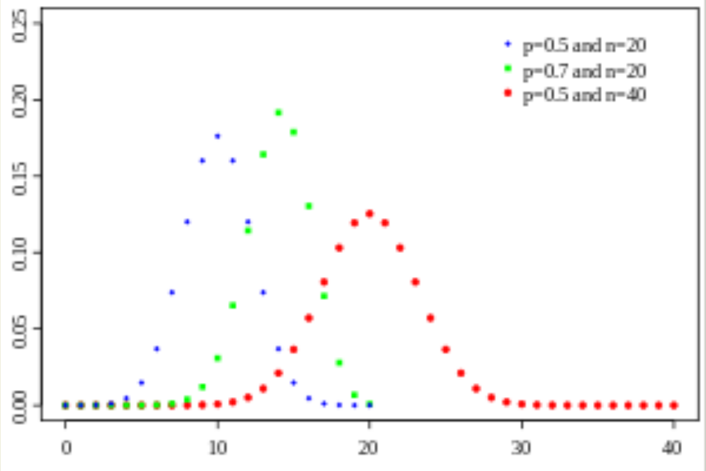


[0,1] 로 Confine
→ 범위가 정해질때 많이 쓰는 Dist

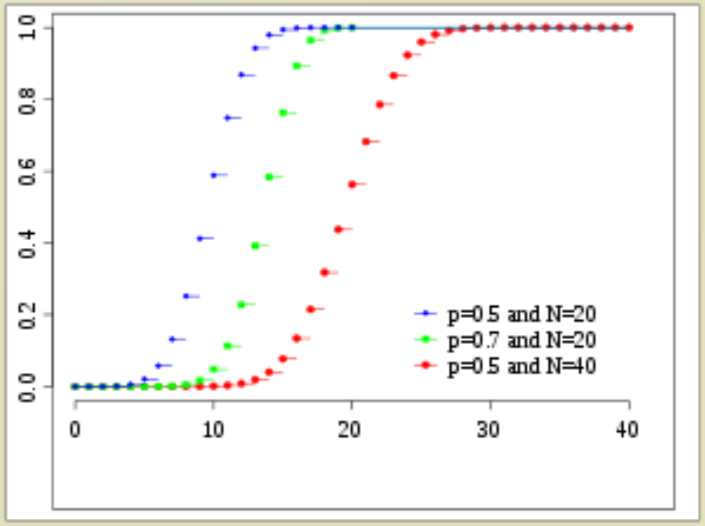


Binomial Distribution

- Simplest distribution for discrete values
  - Bernoulli trial, yes or no
  - 0 or 1
  - Selection, switch….



- $f(\theta; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Notation: $\mathrm{B}(n, p)$
- Mean: np
- Variance: $\mathrm{np}(1-\mathrm{p})$



Multinomial Distribution

still discrete

- The generalization of the binomial distribution
  - Beyond yes/no
  - Choose A, B, C, D, E,….,Z
  - Word selection, cluster selection…..

text mining할 때 많이 이용.

- $f(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n!}{x_1! \dots x_k!} {p_1}^{x_1} \dots {p_k}^{x_k}$
- Notation: $\mathrm{Mult}(P), P =< p_1, \dots, p_k >$
- Mean: $\mathrm{E}(x_i) = np_i$
- Variance: $\mathrm{Var}(x_i) = np_i(1 - p_i)$

Multinomial Distribution

- The generalization of the binomial distribution
  - Beyond yes/no
  - Choose A, B, C, D, E,….,Z
  - Word selection, cluster selection…..

- $f(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n!}{x_1! \dots x_k!} {p_1}^{x_1} \dots {p_k}^{x_k}$
- Notation: $\mathrm{Mult}(P), P =< p_1, \dots, p_k >$
- Mean: $\mathrm{E}(x_i) = np_i$
- Variance: $\mathrm{Var}(x_i) = np_i(1 - p_i)$