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…

WARMING UP
A SHORT EPISODE

Thumbtack Question

- There is a gambling site with a game of flipping a thumbtack
  - Nail is up, and you betted on nail's up you get your money in double
  - Same to the nail's down

- A billionaire wants to enter the gambling
  - With scientific and engineering supports
    - He is paying you a big chunk of money
  - He asks you
    - I have a thumbtack, if I flip it, what's the probability that it will fall with the nail's up?
  - Your response?

Number of Trials

$$\hat{\theta} = \frac{a_H}{a_H + a_T}$$

- You report your proof to the billionaire
  - From the observations of your trials, and from the MLE perspective, and by assuming the binomial distribution assumption……..
  - $\boldsymbol{\theta}$ is 0.6
  - So, you are more likely to win a bet if you choose the ***head***
- He says okay.
  - Billionaire
    - While you were calculating, I was flipping more times.
    - It turns out that we have 30 heads and 20 tails.
    - Does this change anything?
  - Your response
    - No, nothing's changed. Same. 0.6
  - Billionaire
    - Then, I was just spending time for nothing????
- You say no
  - Your additional trials are valuable to …….

MLE perspective

Bayes

Wait!!!

A student whose name is Bayes raised his hand

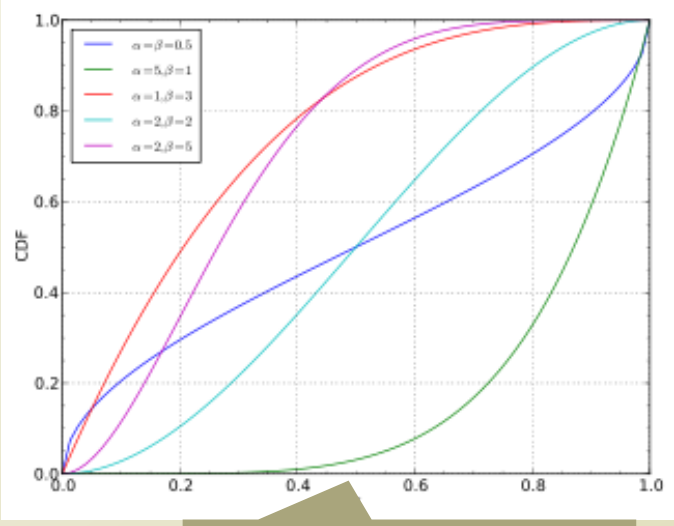
Incorporating Prior Knowledge

- Bayes says
  - Wait. Billionaire.
  - Is it really true that the thumbtack has 60% chance of head?
  - Don't you think it is 50 vs 50?
- Billionaire says
  - Well. I thought so...
  - But, how to merge the previous knowledge to my trials?
- Bayes says
  - So, I give you this theorem!
  - $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
  - You already dealt with $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
  - $P(\theta)$ is the part of the prior knowledge
- Your response is
  - Then, $P(\theta|D)$ is the conclusion influenced by the data and the prior knowledge?
- Bayes says
  - Yes, and it will be our future prior knowledge!

$$\boldsymbol{P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}}$$

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

More Formula from Bayes Viewpoint



Nice match to the range!

- $P(\theta|D) \propto P(D|\theta)P(\theta)$
  - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
  - $P(\theta) = ????$ → 이전 실험에 대해서 알아야해.
- We need to represent the prior knowledge well
  - So, the multiply goes smooth and does not complicate the formula
- Bayes says
  - Why not use the Beta distribution?
  - $P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \Gamma(\alpha) = (\alpha-1)!$
- Your response is
  - Wow convenient!
  - $P(\theta|D) \propto P(D|\theta)P(\theta) \propto \theta^{a_H}(1-\theta)^{a_T}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
    $= \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$ → $P(D|\theta)$의 해석과 동일.
  - Also, you notice one interesting face from the above formula…

$P(\theta)$: 이전경험없음 = const
$\therefore P(\theta) \sim \theta$

Beta : 특정범위에서의 [0,1] 2 confine
↓
param α, β 만족

Maximum a Posteriori Estimation

- Billionaire says
  - Hey! Stop! I am here!
  - So, you are talking about the formula
  - I want the most probable and more approximate θ
- Your response is
  - We are there.
  - Previously in MLE, we found θ from $\hat{\theta} = argmax_\theta P(D|\theta)$
    - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
    - $\hat{\theta} = \frac{a_H}{a_H+a_T}$
  - Now in MAP, we find θ from $\hat{\theta} = argmax_\theta P(\theta|D)$
    - $P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$
    - $\hat{\theta} = \frac{a_H+\alpha-1}{a_H+\alpha+a_T+\beta-2}$
  - The calculation is same because anyhow it is the maximization

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…

$p(D)$	Normalizing Con
$p(D\|\theta)$	Likelihood
$p(\theta)$	prior
$p(\theta\|D)$	posterior

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

BASICS

What we just saw is…

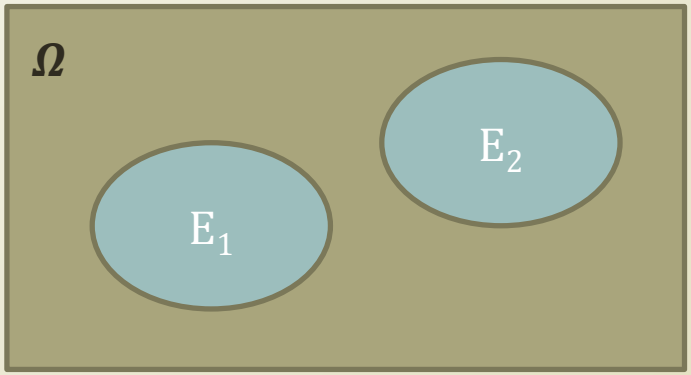
Bayes says

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

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

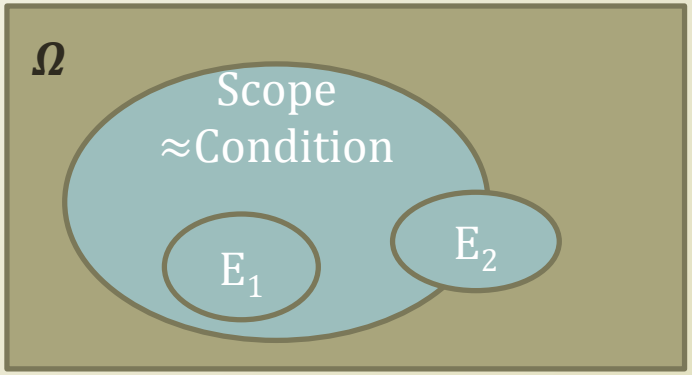
$$\mathrm{P}(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i) \ \textit{when a sequence of mutually exclusive}$$

  - Subsequent characteristics

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

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

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,….
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - The conditional probability of A given B
- Some handy formula

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

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

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

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

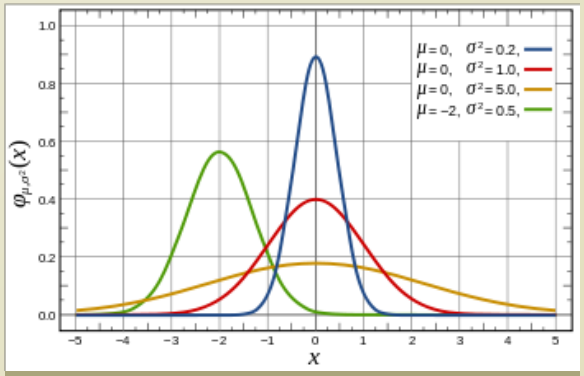
Nice to see that we can recover the target event by adding the whole conditional probs and priors

Probability Distribution

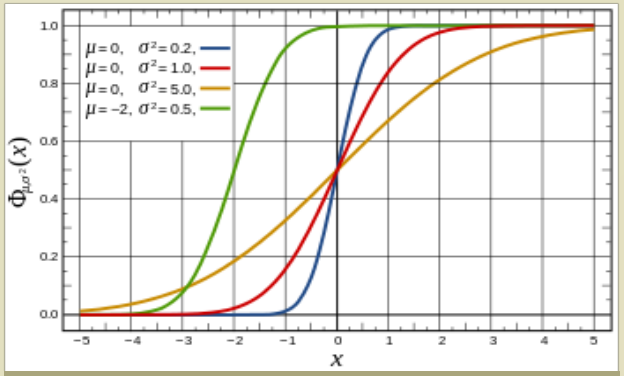
- Probability distribution assigns
  - A probability to a subset of the potential events of a random trial, experiment, survey, etc.
- A function mapping an event to a probability
  - Because we call it a probability, the probability should keep its own characteristics (or axioms)
  - An event can be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

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

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



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

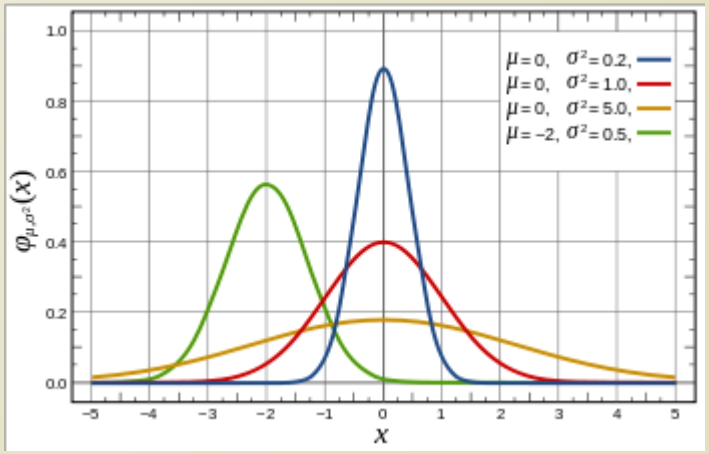


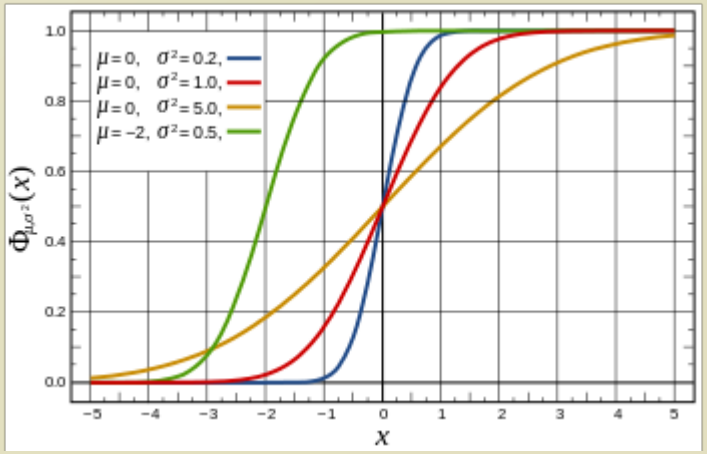
Cumulative Distribution Function
(CDF) =$\int_{-\infty}^{x} f(x)\,dx$

Normal Distribution

- Very commonly observed distribution
  - Continuous numerical value

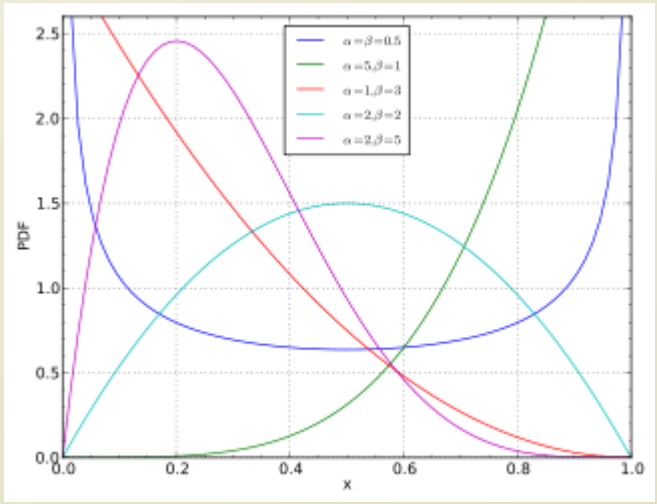
- $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Notation: $N(\mu, \sigma^2)$
- Mean: μ
- Variance: σ^2

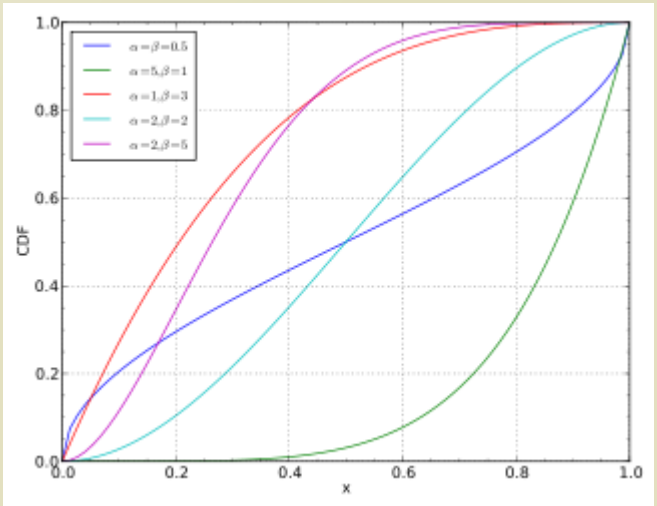




Beta Distribution

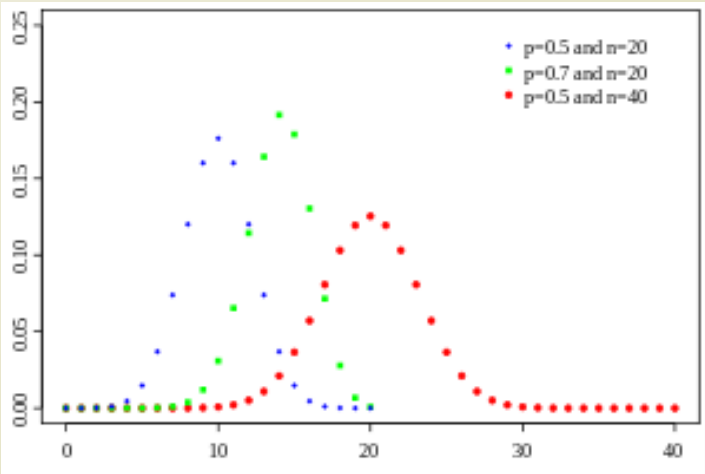
- Supports a closed interval
  - Continuous numerical value
  - [0,1]
  - Very nice characteristic
  - Why?
    - Matches up the characteristics of probs
- $f(\theta;\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$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



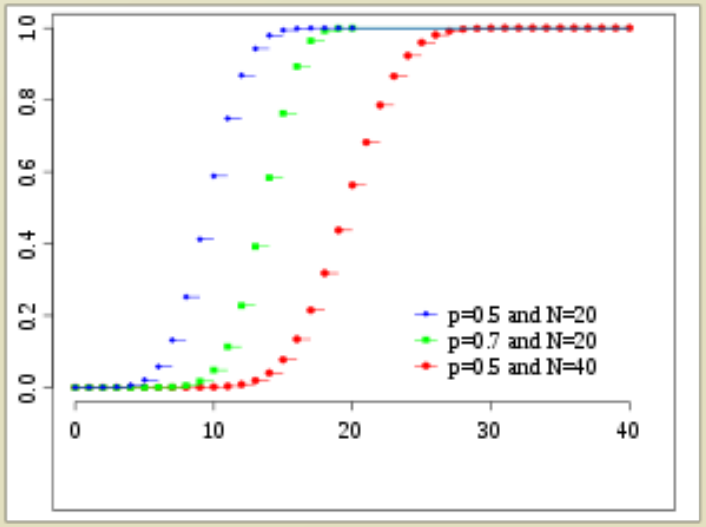


Binomial Distribution

- 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

- 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: Mult(P), $P =< p_1, \dots, p_k >$
- Mean: $\mathrm{E}(x_i) = np_i$
- Variance: $\mathrm{Var}(x_i) = np_i(1 - p_i)$

Conclusion from Anecdote

- Billionaire says
  - Wait you and Bayes!
  - Who is right? The numbers are different!
- Bayes says
  - Not really… if you give us enough money to replicate the game!
- You say
  - Yes! If a_H and a_T become big, α and β becomes nothing…
- Billionaire says
  - Enough talking
  - Still, α and β are important if I don't give you more trials
  - Who decides α and β?
- Bayes and you say
  - Well… maybe grad students? =)

→ 관측이 많아지면 MAP ≠ MLE

MLE

$$\hat{\theta} = \frac{a_H}{a_H + a_T}$$

Quiz 1-2: $\frac{9+6}{10+10} = \frac{15}{20}$

MAP

$$\hat{\theta} = \frac{a_H + \alpha - 1}{a_H + \alpha + a_T + \beta - 2}$$

IV 영향하는 a_H, a_T Point.
= 사전정보 영향 ↓.