# 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

압정

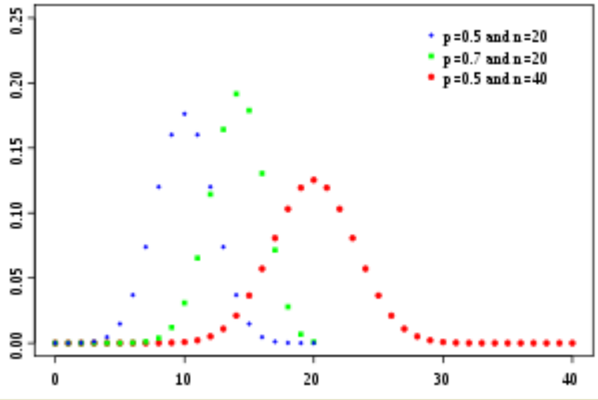
- There is a gambling site with a game of flipping a thumbtack
  - Nail is up, and you betted on nail's up you get your money in double
  - Same to the nail's down

- 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? 압정으로 하는 도박에서 돈을 벌 수 있을까?
  - Your response?

# Experience from trials

- My response is
  - Please flip it a few times
- Billionaire tried for five times
  - The nail's up case is three out of five trials
- My response is
  - You should invest
    - 3/5 to nail's up case
    - 2/5 to nail's down case
- The billionaire asks why?
- Then,
  - You answer……

HHTTT




# Binomial Distribution

- Binomial distribution is
  - The ***discrete probability distribution*** (↔ Cont.)
    - Of the number of successes in a sequence of ***n independent yes/no experiments***, and each success has the probability of $\boldsymbol{\theta}$
  - Also called a Bernoulli experiment
- Flips are i.i.d
  - Independent events
  - Identically distributed according to binomial distribution
- $P(H) = \theta$, $P(T)=1-\theta$
- $P(HHTHT)= \theta\theta(1-\theta)\theta(1-\theta)=\theta^3(1-\theta)^2$
- Let's say
  - D as Data = H,H,T,H,T
    - $n=5$
    - $k=a_H=3$
    - $p=\theta$
  - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$

***n*** and ***p*** are given as parameters, and the value is calculated by varying ***k***

$$f(k;n,p) = P(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

Makes order insensitive

# Maximum Likelihood Estimation

- $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
- Data: We have observed the sequence data of D with $a_H$ and $a_T$
- Our hypothesis
  - The gambling result of thumbtack follows the binomial distribution of $\theta$
- How to make our hypothesis strong?
  - Finding out a better distribution of the observation
    - Can be done, but you need more rational.
  - Finding out the best candidate of $\boldsymbol{\theta}$ → 어떤 θ일 때 해야 Data 잘 표현 & Hypothesis 강화?
    - What's the condition to ***make θ most plausible?***
- One candidate is the **Maximum Likelihood Estimation (MLE) of *θ***
  - Choose θ that maximizes the probability of observed data

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

θ hat

KAIST

# Maximum likelihood estimation

언제 Max 하는지 알 수 있음 (derivative)

This article is about the statistical techniques. For computer data storage, see Partial response maximum likelihood.

In statistics, **maximum likelihood estimation** (**MLE**) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.[1] The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.[2][3][4]

If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the linear regression model.[5] Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

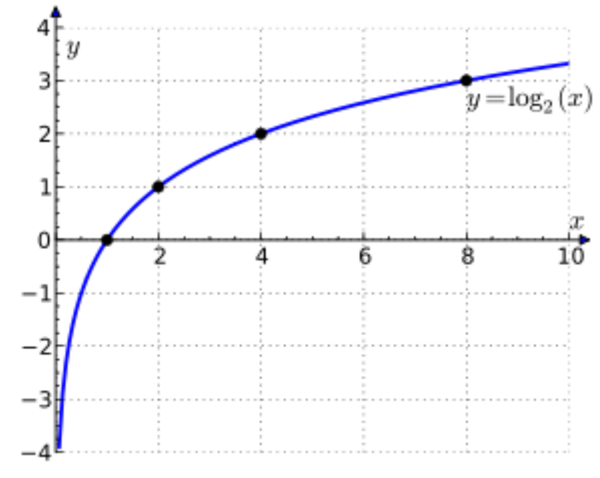
From the vantage point of Bayesian inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is a special case of an extremum estimator, with the objective function being the likelihood.

Contents

## Principles

From a statistical standpoint, a given set of observations are a random sample from an unknown population. The goal of maximum likelihood estimation is to make inferences about the population that is most likely to have generated the sample,[6] specifically the joint probability distribution of the random variables $\{y_1, y_2, \ldots\}$, not necessarily independent and identically distributed. Associated with each probability distribution is a unique vector $\theta = [\theta_1,\, \theta_2,\, \ldots,\, \theta_k]^{\mathsf{T}}$ of parameters that index the probability distribution within a parametric family $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$, where $\Theta$ is called the parameter space, a finite-dimensional subset of Euclidean space. Evaluating the joint density at the observed data sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ gives a real-valued function,

# MLE Calculation




- $\hat{\theta} = argmax_{\theta} P(D|\theta) = argmax_{\theta} \theta^{a_H}(1-\theta)^{a_T}$
- This is going nowhere, so you use a trick
  - Using the log function : monotonic 증가 ∴ 일대일대응 → (P 최대 = log 최대) 되는 θ 같다
- $\hat{\theta} = argmax_{\theta} lnP(D|\theta) = argmax_{\theta} \ln\{\theta^{a_H}(1-\theta)^{a_T}\}$
  $= argmax_{\theta}\{a_H\, ln\theta + a_T \ln(1-\theta)\}$
- Then, this is a maximization problem, so you use a derivative that is set to zero
  - $\frac{d}{d\theta}(a_H\, ln\theta + a_T \ln(1-\theta)) = 0$
  - $\frac{a_H}{\theta} - \frac{a_T}{1-\theta} = 0$
  - $\theta = \frac{a_H}{a_T + a_H}$
  - When $\theta$ is $\frac{a_H}{a_T+a_H}$, the $\theta$ becomes the best candidate from the MLE perspective
- $\hat{\theta} = \frac{a_H}{a_H + a_T}$ → (bin & MLE & Optimization) 셋 모두 고려한 결과였던 것.

# 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 …….

# Simple Error Bound

- Your response 추론을 한것 ($\hat{\theta}$) ∴ Error가 있다 → N과 관계.
  - Your additional trials reduce the error of our estimation $N \propto (\text{error bound})^{-1}$
  - Right now, we have $\hat{\theta} = \frac{a_H}{a_H + a_T}$, $N = a_H + a_T$
  - Let's say $\theta^*$ is the true parameter of the thumbtack flipping for any error, $\varepsilon > 0$
  - We have a simple upper bound on the probability provided by Hoeffding's inequality
  - $P(|\hat{\theta} - \theta^*| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}$

Coming from a friend in the math. dept.

- Billionaire asks you
  - Can you calculate the required number of trials, N?
    - To obtain $\varepsilon = 0.1$ with 0.01% case
- Now, your professor jumps in and says 대략 정답.
  - This is Probably Approximate Correct (PAC) learning
    - Probably? (0.01% case) 에러로 이 확률안에서 Correct한 결과가 있다.
    - Approximately? ($\varepsilon = 0.1$)