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

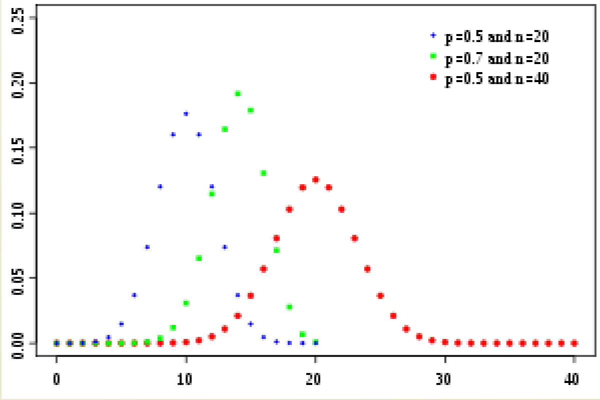
- There is a gambling site with a game of flipping a thumbtack
  - Nail is up, and you betted on nail's up you get your money in double
  - Same to the nail's down

- 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?

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……

Binomial Distribution



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

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

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

Makes order insensitive

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

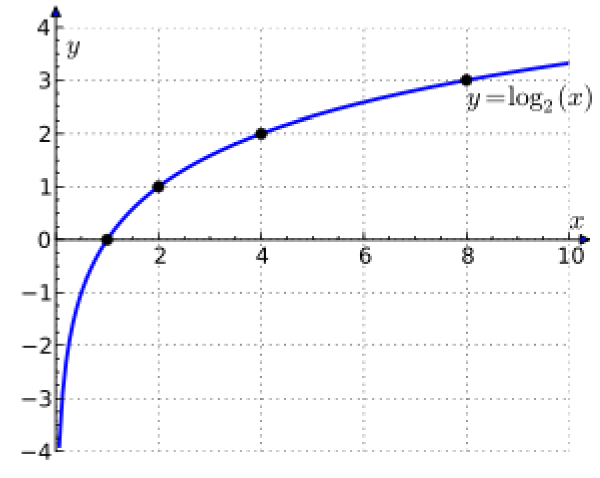
Maximum Likelihood Estimation

- $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
- 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 θ
- 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 $\boldsymbol{\theta}$
    - What's the condition to **make $\boldsymbol{\theta}$ *most plausible?***

- One candidate is the **Maximum Likelihood Estimation (MLE) of $\boldsymbol{\theta}$**
  - Choose θ that maximizes the probability of observed data

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

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
- $\hat{\theta} = argmax_\theta lnP(D|\theta) = argmax_\theta \ln\{\theta^{a_H}(1-\theta)^{a_T}\}$
  $= argmax_\theta\{a_H\, ln\theta + a_T \ln(1-\theta)\}$
- 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 θ is $\frac{a_H}{a_T + a_H}$, the θ becomes the best candidate from the MLE perspective
- $\hat{\theta} = \frac{a_H}{a_H + a_T}$

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
  - Your additional trials reduce the error of our estimation
  - Right now, we have $\hat{\theta} = \frac{a_H}{a_H+a_T}$, N $= a_H + a_T$
  - Let's say θ^* 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
  - This is Probably Approximate Correct (PAC) learning
    - Probably? (0.01% case)
    - Approximately? ($\varepsilon = 0.1$)