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…

MOTIVATION

Keywords

- Many floating keywords
  - Data-mining, Knowledge discovery, Machine Learning, Artificial Intelligence…
- Comes from territory, perspectives, types of problems, researchers, etc
- We are going to focus on substance, not labeling.
- I am just going to call it "Machine Learning"
  - You can call it whatever you want

AI in CS

Statistics

Database in CS

Management

Industrial Engineering

…….

Abundance of Data

- Data are being collec Network Data rywhere

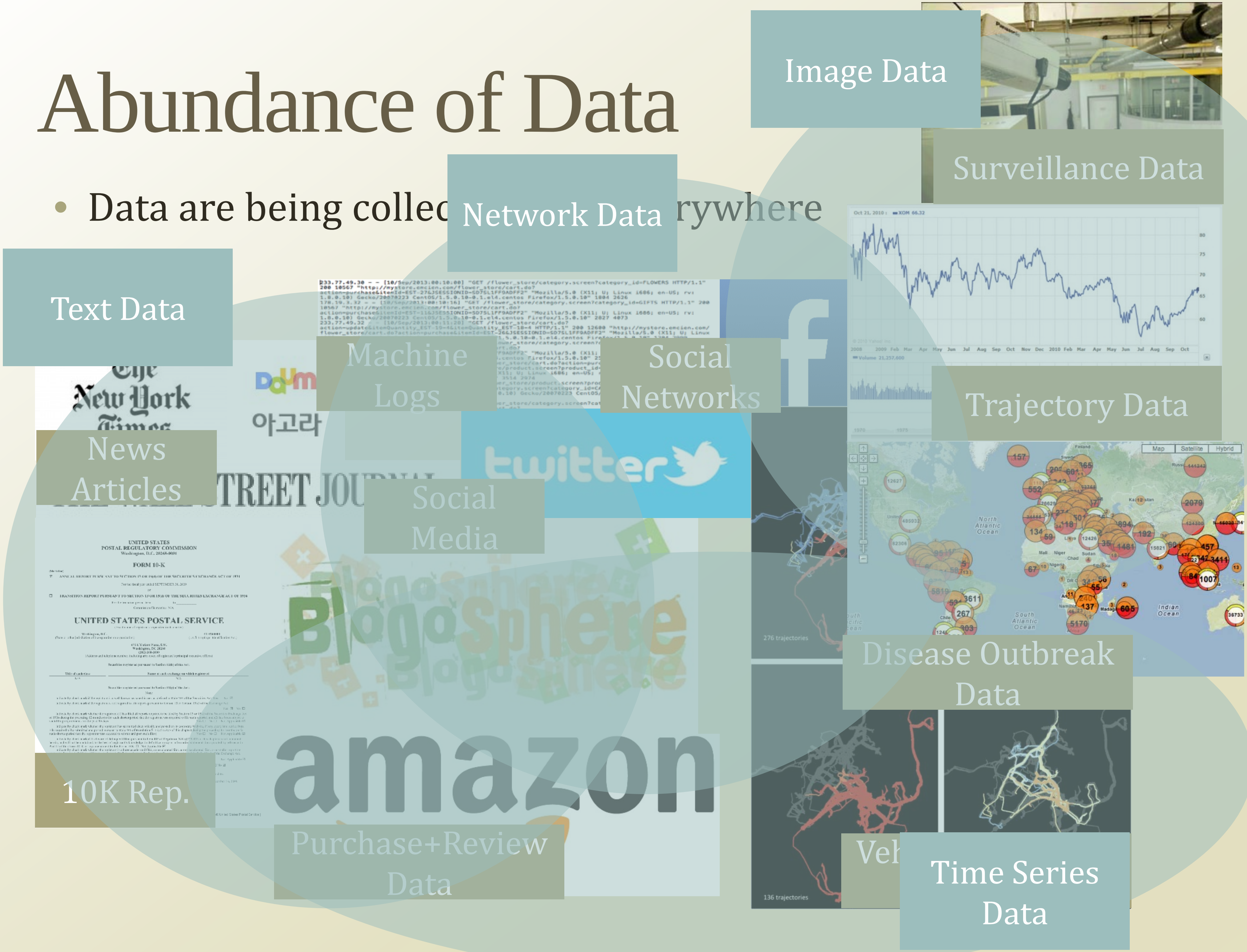


Image Data

Surveillance Data

Network Data

Text Data

Machine Logs

Social Networks

Trajectory Data

News Articles

Social Media

Disease Outbreak Data

10K Rep.

Purchase+Review Data

Veh

Time Series Data

```
233.77.49.30 - - [10/Sep/2013:00:10:00] "GET /flower_store/category.screen?category_id=FLOWERS HTTP/1.1"
200 10567 "http://mystore.emcien.com/flower_store/cart.do?
action=purchase&itemId=EST-27&JSESSIONID=SD7SL1FF9ADFF2" "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:
1.8.0.10) Gecko/20070223 CentOS/1.5.0.10-0.1.el4.centos Firefox/1.5.0.10" 1804 2626
178.19.3.32 - - [10/Sep/2013:00:10:16] "GET /flower_store/category.screen?category_id=GIFTS HTTP/1.1" 200
10567 "http://mystore.emcien.com/flower_store/cart.do?
action=purchase&itemId=EST-11&JSESSIONID=SD7SL1FF9ADFF2" "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:
1.8.0.10) Gecko/20070223 CentOS/1.5.0.10-0.1.el4.centos Firefox/1.5.0.10" 2827 4073
233.77.49.32 - - [10/Sep/2013:00:11:20] "GET /flower_store/cart.do?
action=update&itemQuantity_EST-10-4&itemQuantity_EST-10-4 HTTP/1.1" 200 12600 "http://mystore.emcien.com/
flower_store/cart.do?action=purchase&itemId=EST-26&JSESSIONID=SD7SL1FF9ADFF2" "Mozilla/5.0 (X11; U; Linux
```

Examples of Machine Learning Applications

- Machine Learning is everywhere…

Document Classification

Sports
Science
News

Stock Market Prediction

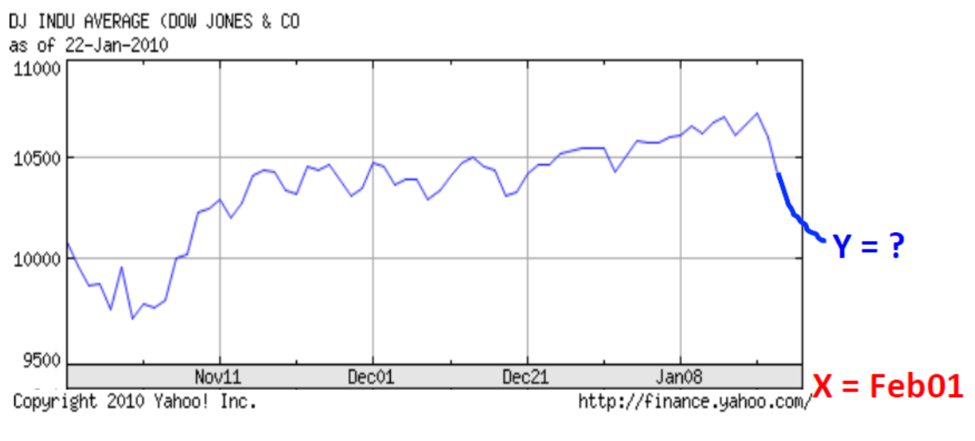
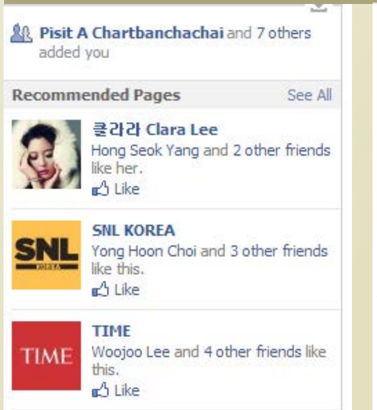


Plate Num. Recognition

SNS Recommendation

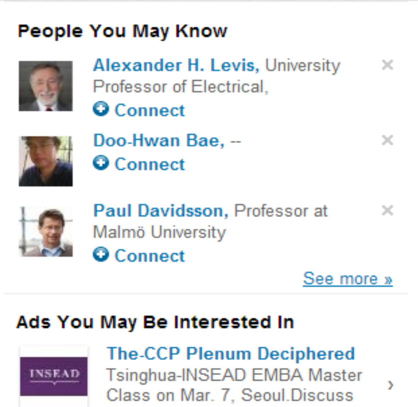

Pisit A Chartbanchachai and 7 others added you

Recommended Pages See All

클라라 Clara Lee
Hong Seok Yang and 2 other friends like her.
Like

SNL KOREA
Yong Hoon Choi and 3 other friends like this.
Like

TIME
Woojoo Lee and 4 other friends like this.
Like


People You May Know

Alexander H. Levis, University Professor of Electrical,
Connect

Doo-Hwan Bae, --
Connect

Paul Davidsson, Professor at Malmö University
Connect

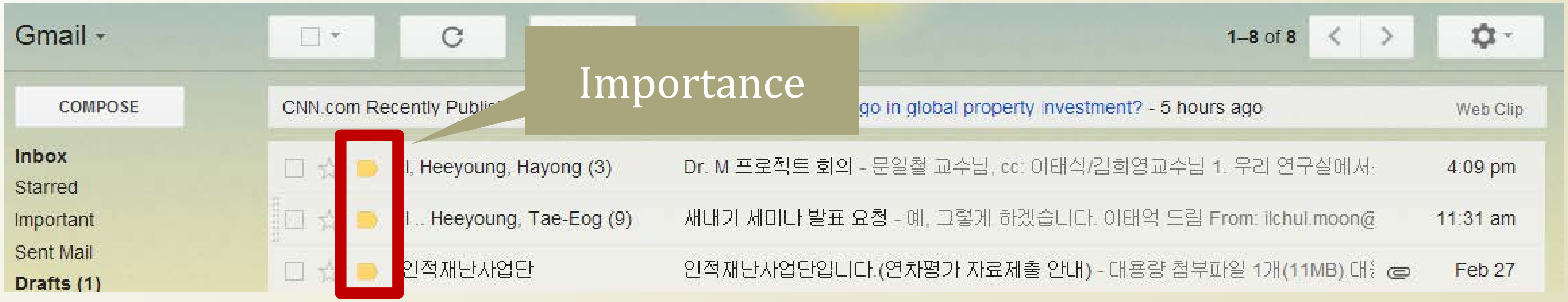
See more »

Ads You May Be Interested In

The-CCP Plenum Deciphered
Tsinghua-INSEAD EMBA Master Class on Mar. 7, Seoul.Discuss

Helicopter Control

Spam Filtering and more



CNN.com Recently Publi... go in global property investment? - 5 hours ago

l, Heeyoung, Hayong (3)	Dr. M 프로젝트 회의 - 문일철 교수님, cc: 이태식/김희영교수님 1. 우리 연구실에서	4:09 pm
l .. Heeyoung, Tae-Eog (9)	새내기 세미나 발표 요청 - 예, 그렇게 하겠습니다. 이태억 드림 From: ilchul.moon@	11:31 am
인적재난사업단	인적재난사업단입니다.(연차평가 자료제출 안내) - 대용량 첨부파일 1개(11MB) 대	Feb 27

Importance

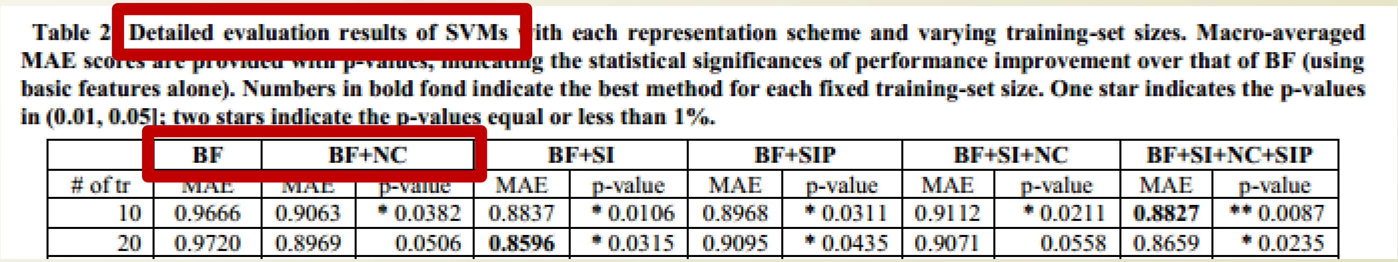


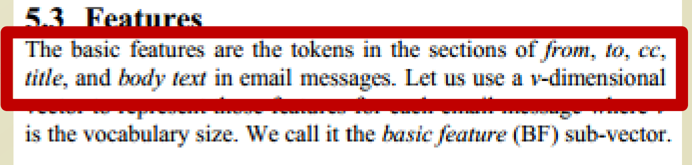
Table 2 Detailed evaluation results of SVMs with each representation scheme and varying training-set sizes. Macro-averaged MAE scores are provided with p-values, indicating the statistical significances of performance improvement over that of BF (using basic features alone). Numbers in bold fond indicate the best method for each fixed training-set size. One star indicates the p-values in (0.01, 0.05]; two stars indicate the p-values equal or less than 1%.

	BF	BF+NC		BF+SI		BF+SIP		BF+SI+NC		BF+SI+NC+SIP	
# of tr	MAE	MAE	p-value	MAE	p-value	MAE	p-value	MAE	p-value	MAE	p-value
10	0.9666	0.9063	* 0.0382	0.8837	* 0.0106	0.8968	* 0.0311	0.9112	* 0.0211	**0.8827**	** 0.0087
20	0.9720	0.8969	0.0506	**0.8596**	* 0.0315	0.9095	* 0.0435	0.9071	0.0558	0.8659	* 0.0235

SVM?

- Spam filter
- More?
  - Importance vs. Urgency
- How to predict an important email?
  - Social networks
  - Contents
- Shinjae Yoo, Yiming Yang, Frank Lin, and Il-Chul Moon, Mining Social Networks for Personalized Email Prioritization, ACM SIGKDD Conference, Paris, France, Jun, 28, 2009

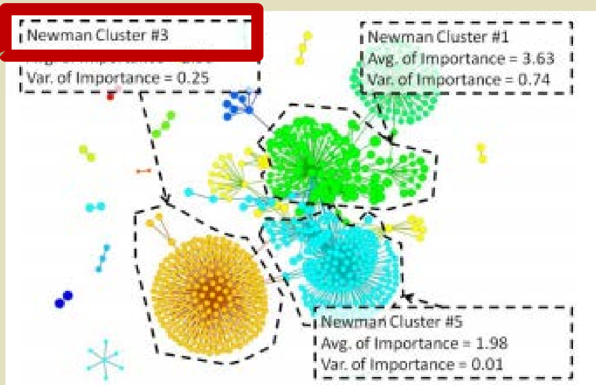
Features



5.3 Features

The basic features are the tokens in the sections of *from*, *to*, *cc*, *title*, and *body text* in email messages. Let us use a *v*-dimensional vector to represent these features for each email message, where *v* is the vocabulary size. We call it the *basic feature* (BF) sub-vector.

Clusters? Is this a machine learning technique?

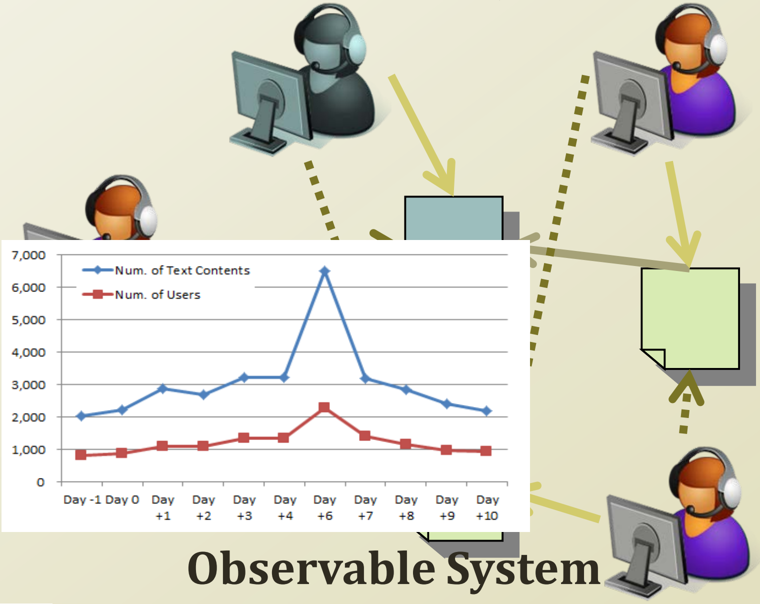

Newman Cluster #3
Avg. of Importance = [illegible]
Var. of Importance = 0.25
Newman Cluster #1
Avg. of Importance = 3.63
Var. of Importance = 0.74
Newman Cluster #5
Avg. of Importance = 1.98
Var. of Importance = 0.01

Opinion Mining and more

PCC Cheonan
Sank on Mar 26, 2010

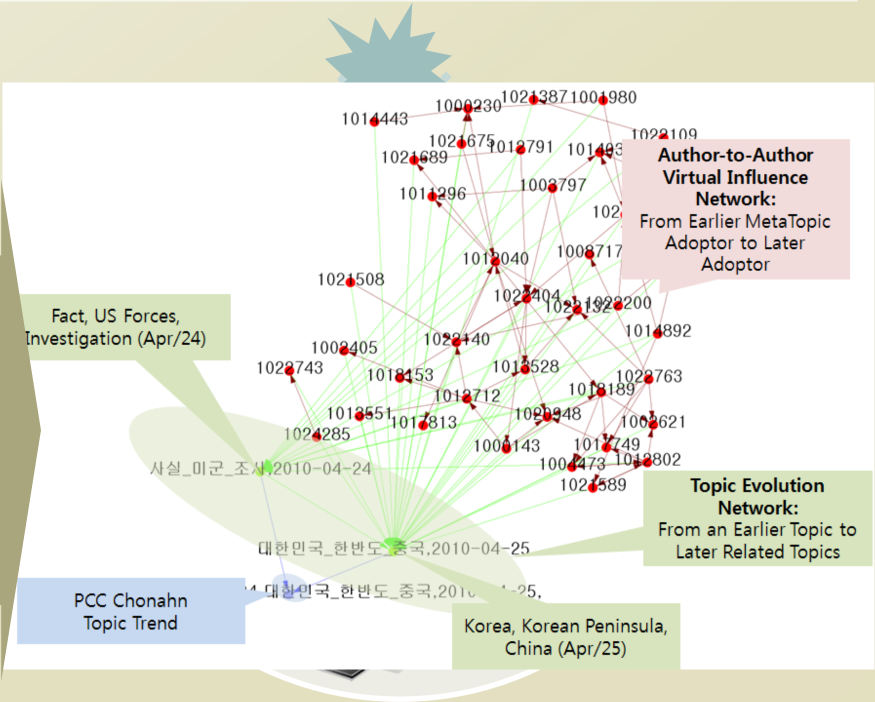
Finding out consensus of the population

- Mining population's perception of the event
- Mining key opinion buried in a data chunk
- Estimating future polarity of the population
- Strategy to maintain the unity of the population



Observable System

Data-mining on SNS and Texts



Author-to-Author Virtual Influence Network: From Earlier MetaTopic Adoptor to Later Adoptor

Fact, US Forces, Investigation (Apr/24)

Topic Evolution Network: From an Earlier Topic to Later Related Topics

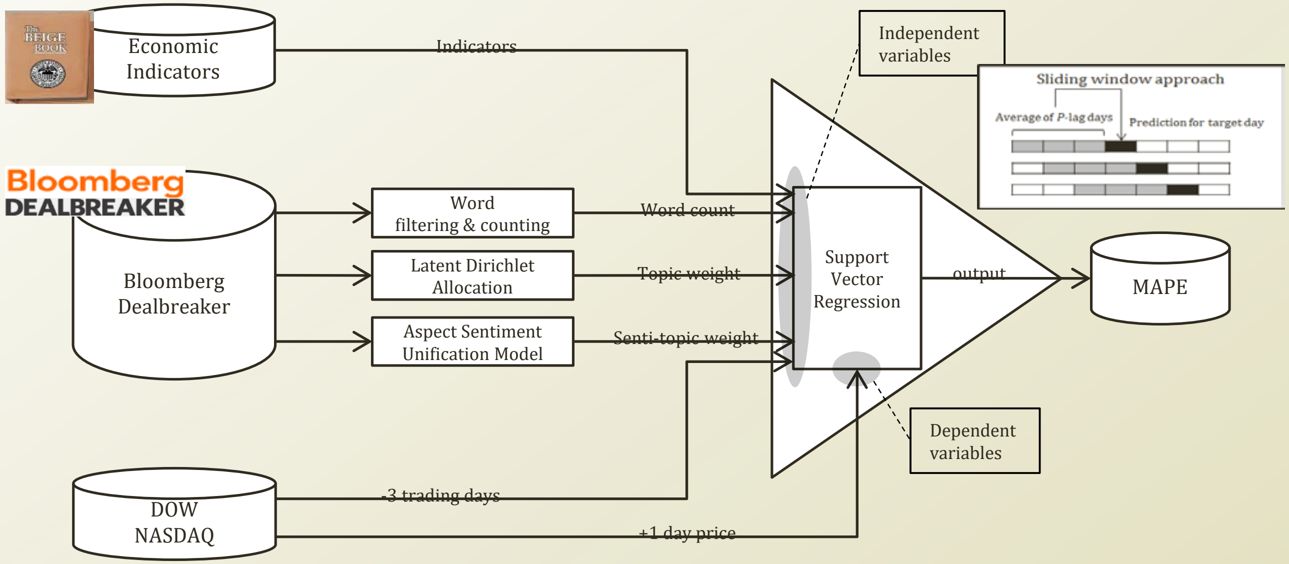
PCC Chonahn Topic Trend

Korea, Korean Peninsula, China (Apr/25)

Implicit System

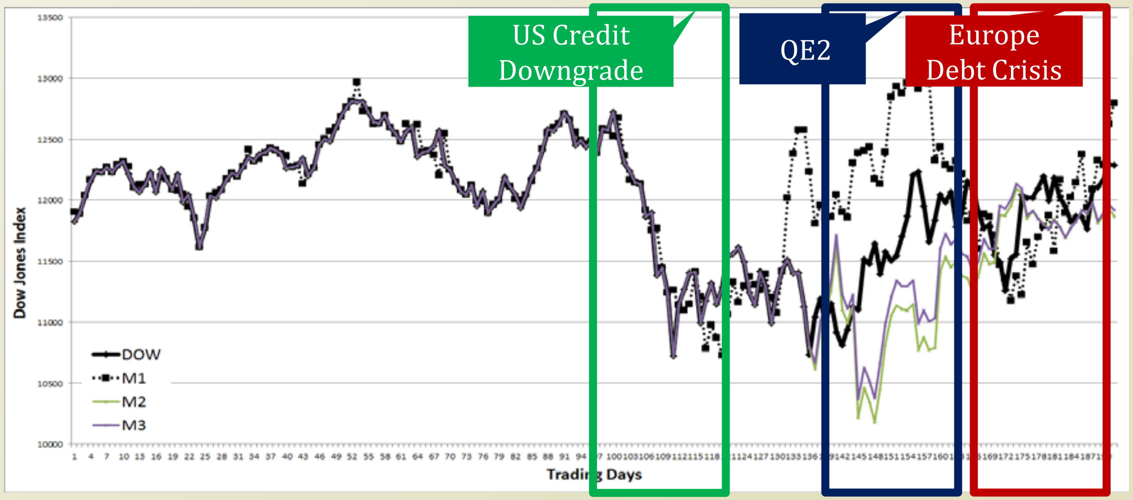
W.S. Lee, J.Y. Bak, A.H. Oh, and I.-C. Moon. Media Analysis to Estimate Stock Market Movement. Information Fusion. Major Revision

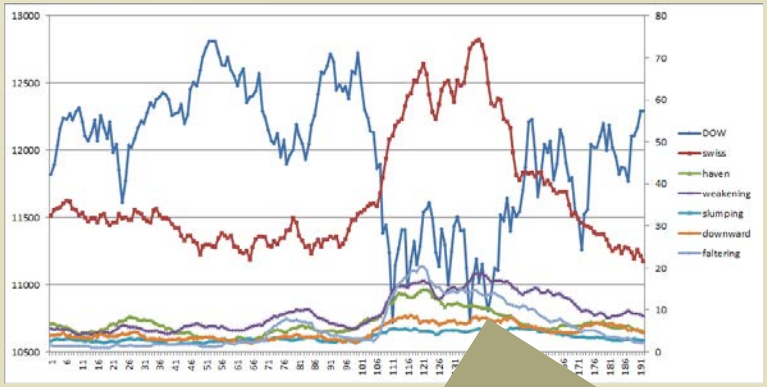
Stock Market Prediction and more



High Coefficients on Prediction

TopicWeight 26	TopicWeight 1	TopicWeight 7
-0.609	0.520	0.508
notes	obama	jun
moodys	republican	pence
swaps	republicans	na
treasuries	congress	swiss
versus	senate	chg
ratings	bill	francs
auction	barack	spa
default	lawmakers	fullyear
strategist	administration	nv
franc	democrats	dividend
twoyear	taxes	firstquarter
samp	white	ks
currencies	workers	paris
yen	democrat	reporting
swiss	obamas	tech





Heavy negative correlation between "*swiss*" and DJIA

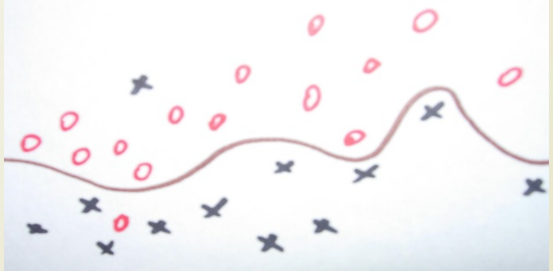
Types of Machine Learning

Machine Learning

.....

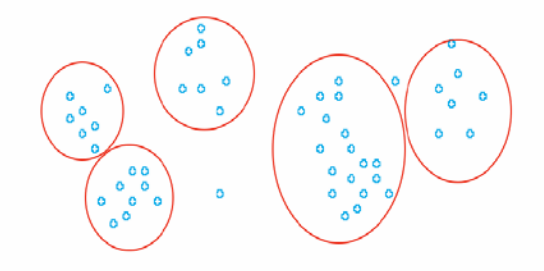
Supervised Learning

You know the true answers of some of instances



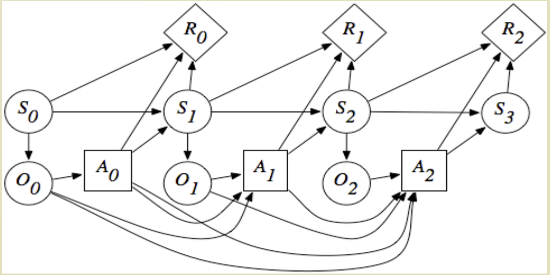
Unsupervised Learning

You do not know the true answers of instances



Reinforcement Learning

You do know the objective, but you do not know how to achieve



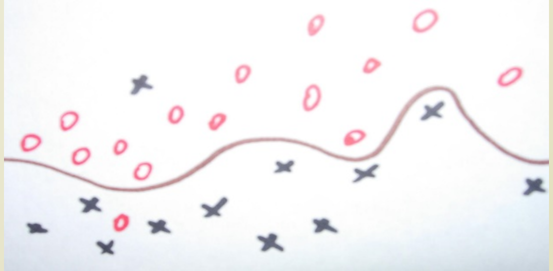
- ***You*** can
  - Machine learning
  - Dataset provider
  - Machine learning users
  - etc

- Various classifications by different professors
  - Purpose, data types, etc
- Other learning classifications also exist

Supervised Learning

Supervised Learning

You know the true answers of some of instances

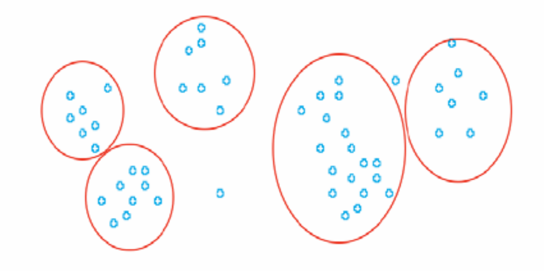


- **You know the true value, and you can provide examples of the true value.**
- Cases, such as
    - Spam filtering
    - Automatic grading
    - Automatic categorization
- Classification or Regression of
    - Hit or Miss: Something has **either disease or not.**
    - Ranking: Someone received **either A+, B, C, or F**.
    - Types: An article is **either positive or negative**.
    - Value prediction: The price of this artifact is **X**.
- Methodologies
    - Classification: estimating a discrete dependent value from observations
    - Regression: estimating a (continuous) dependent value from observations

Unsupervised Learning

Unsupervised Learning

You do not know the true answers of instances



- **You don't know the true value, and you cannot provide examples of the true value.**
- Cases, such as
  - Discovering clusters
  - Discovering latent factors
  - Discovering graph structures
- Clustering or filtering or completing of
  - Finding **the representative topic words from text data**
  - Finding **the latent image from facial data**
  - Completing the incomplete **matrix of product-review scores**
  - Filtering the **noise from the trajectory data**
- Methodologies
  - Clustering: estimating sets and affiliations of instances to the sets
  - Filtering: estimating underlying and fundamental signals from the mixture of signals and noises

WARMING UP
A SHORT EPISODE

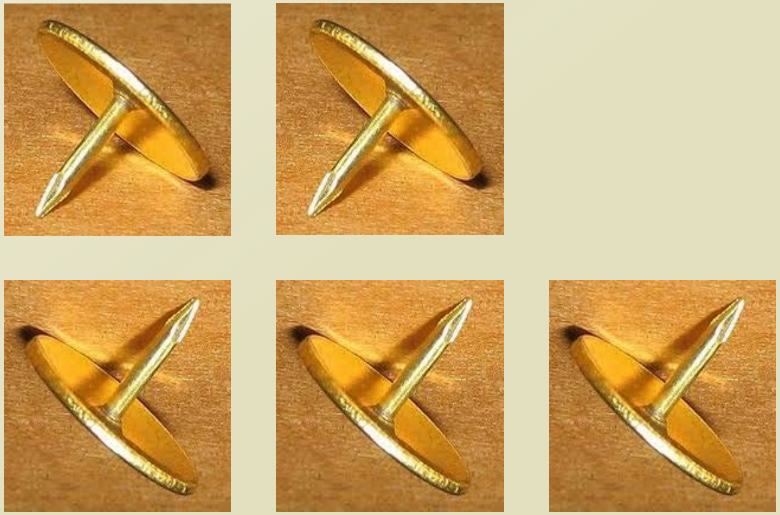
Thumbtack Question

- 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

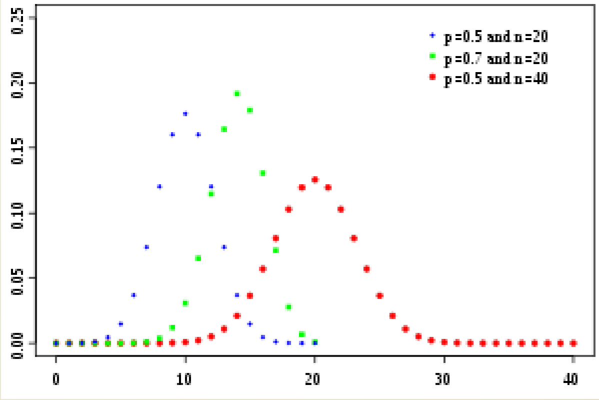
- A billionaire wants to enter the gambling
  - With scientific and engineering supports
    - He is paying you a big chunk of money
  - He asks you
    - I have a thumbtack, if I flip it, what's the probability that it will fall with the nail's up?
  - Your response?

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)=θ^3(1- θ)^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***

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

Makes order insensitive

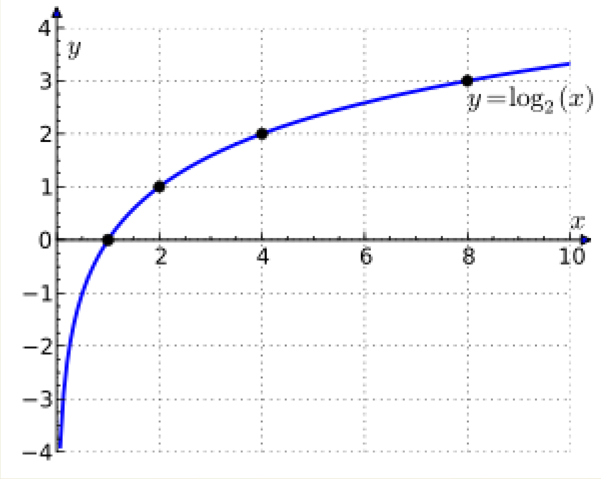
$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

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

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

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, ε > 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? ($\varepsilon = 0.1$)

Probably Approximately Correct Learning

Old, but good book,
So called classic

- The PAC learning theory from the machine learning community,
  - The learner
    - Receives sample data
    - Select a generalization function, hypothesis
      - From a certain class of possible functions
    - Such that the selected function
      - Has low generalization error
      - With high probability
- Definition of machine learning
  - Many views exist
  - Tom Mitchell
    - A computer program is said to
      - learn from experience E
      - With respect to some class of tasks T
      - And performance measure P, if its performance at tasks in T, as measured by P, improves with experience E
- Is PAC Learning machine learning?

Machine Learning

Given a dataset, we select a **generalized function** that provides **most probable and most approximate answers**

Wait!!!

A student whose name is Bayes raised his hand

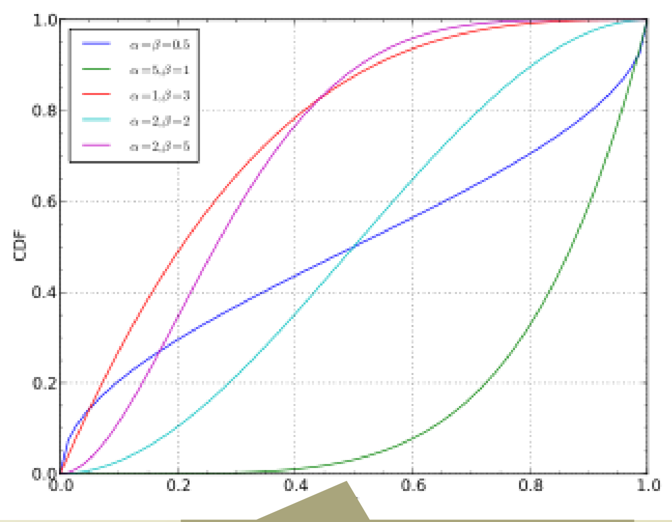
Incorporating Prior Knowledge

- 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(θ) 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!

- $P(\theta|D) \propto P(D|\theta)P(\theta)$
  - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
  - $P(\theta) =????$
- 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!
  - $P(\theta|D) \propto P(D|\theta)P(\theta) \propto \theta^{a_H}(1-\theta)^{a_T}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
    $= \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$
  - Also, you notice one interesting face from the above formula…

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

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? =)

MLE

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

MAP

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

BASICS

What we just saw is…

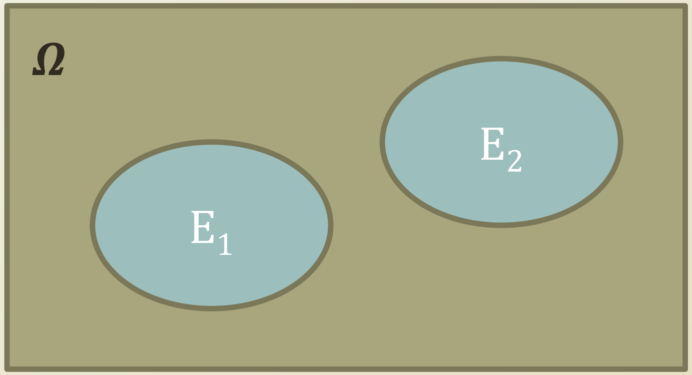
Bayes says

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

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

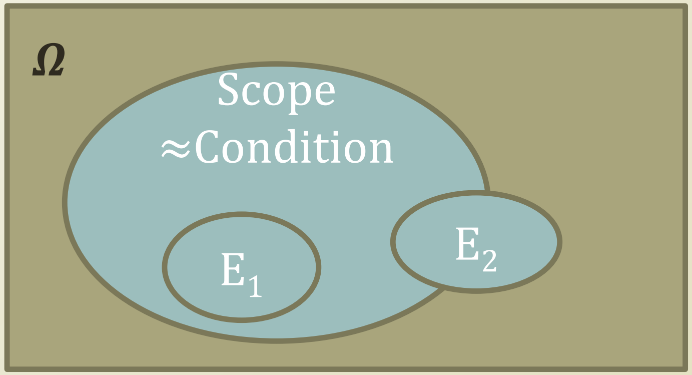
$$\mathrm{P}(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i) \ \textit{when a sequence of mutually exclusive}$$

  - Subsequent characteristics

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

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

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

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

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

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

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

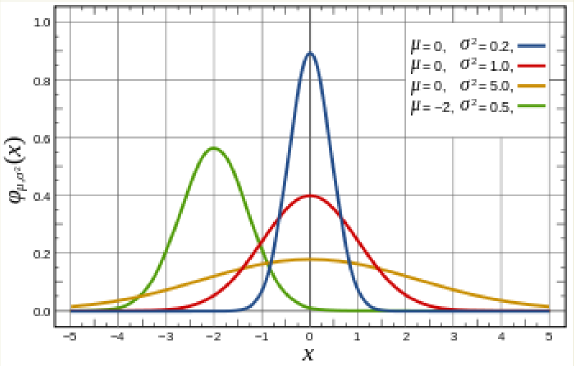
Nice to see that we can recover the target event by adding the whole conditional probs and priors

Probability Distribution

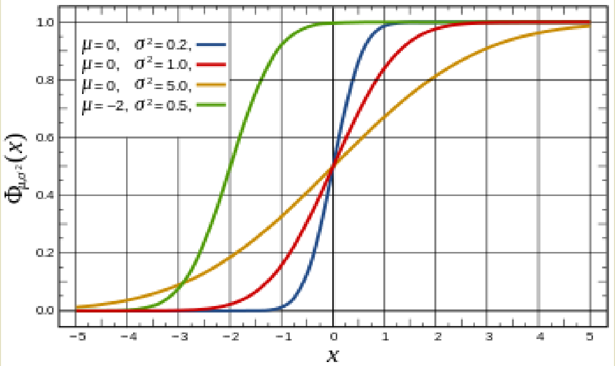
- Probability distribution assigns
  - A probability to a subset of the potential events of a random trial, experiment, survey, etc.
- A function mapping an event to a probability
  - Because we call it a probability, the probability should keep its own characteristics (or axioms)
  - An event can be
    - 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)=$f(x)$

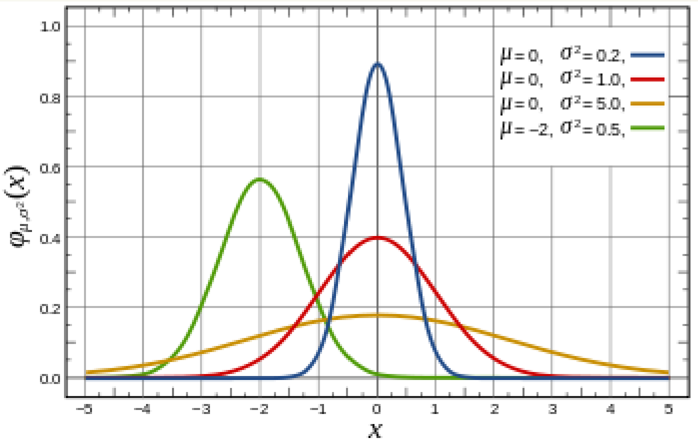


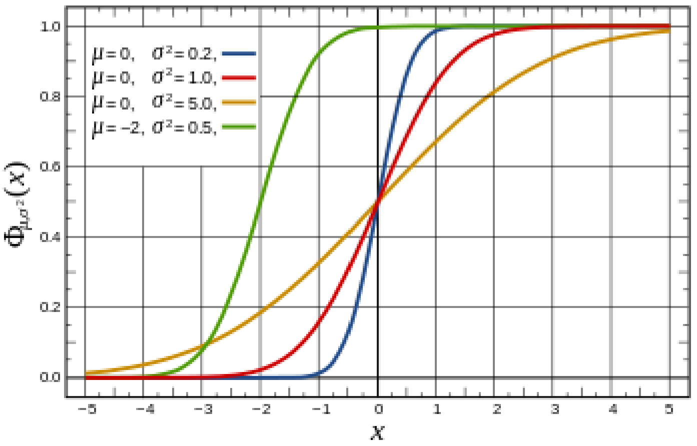
Cumulative Distribution Function
(CDF) = $\int_{-\infty}^{x} f(x)\,dx$

Normal Distribution

- Very commonly observed distribution
  - Continuous numerical value

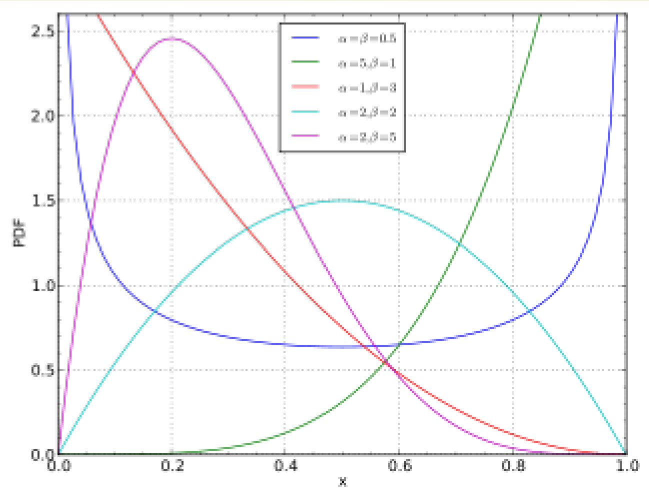
- $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Notation: $N(\mu, \sigma^2)$
- Mean: μ
- Variance: σ^2

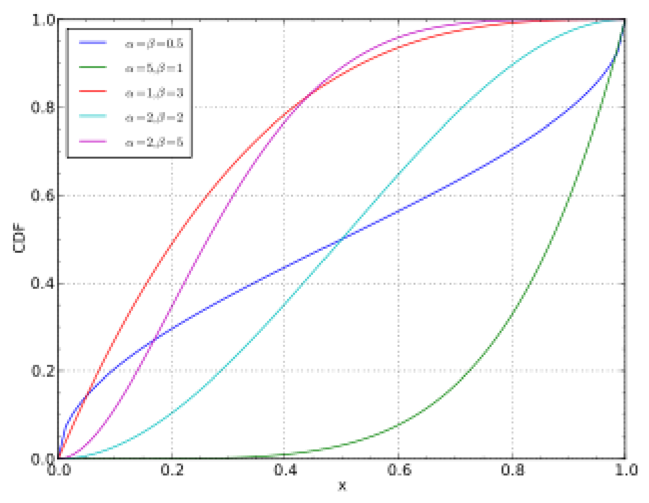




Beta Distribution

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

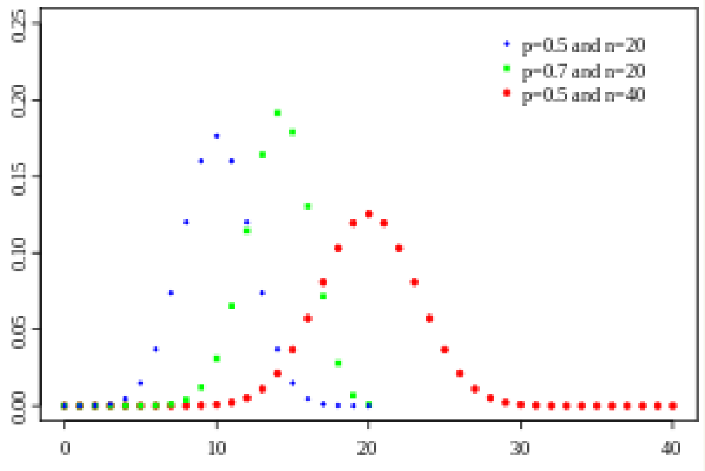


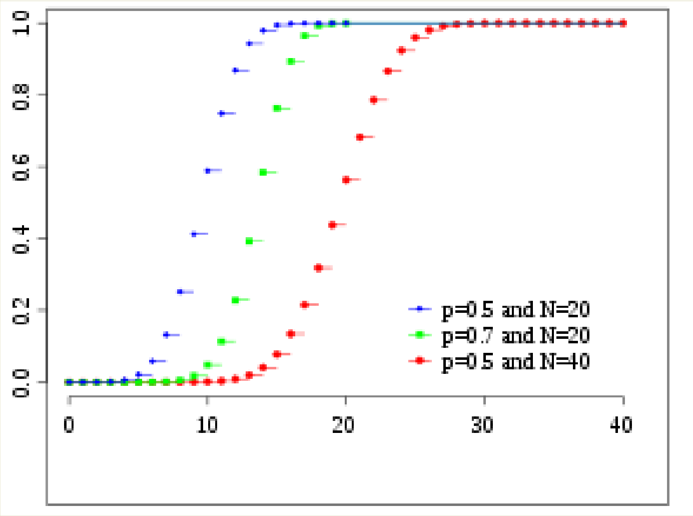


Binomial Distribution

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

Acknowledgement

- This slideset is greatly influenced
  - By Prof. Carlos Guestrin at CMU
  - By Prof. Eric P. Xing at CMU
- Some images are copied from the lecture slides made
  - By Prof. Carlos Guestrin at CMU
  - By Prof. Eric P. Xing at CMU

Further Readings

- Bishop Chapter 1 and 2