



# Introduction to Statistical Machine Learning; MLE and MAP

STAT5241 Section 2

Statistical Machine Learning

Xiaofei Shi



# Basic administrative details:

- All up-to-date information is on Courseworks
- Lectures are intended to be self-contained. For supplementary readings:
  - Pattern Recognition and Machine Learning, Christopher Bishop.
  - Machine Learning: A probabilistic perspective, Kevin Murphy.
  - The Elements of Statistical Learning: Data Mining, Inference and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman.
  - Machine Learning, Tom Mitchell.



## Basic administrative details:

- Instructor: Xiaofei Shi [xs2427@columbia.edu](mailto:xs2427@columbia.edu)

Office hours: Thu 12:30 - 1:30 pm

- TAs: Ling Chen [lc3521@columbia.edu](mailto:lc3521@columbia.edu)

Jaesung Son [js4638@columbia.edu](mailto:js4638@columbia.edu)

Zhanhao Zhang [zz2760@columbia.edu](mailto:zz2760@columbia.edu)

Office hours: Ling Chen Mon 8:00 pm - 9:30 pm

Jaesung Son Wed 2:00 pm - 3:30 pm





# Tentative Evaluation Plan

Course grade = 40% Homework + 40% Project + 20% Participation

- (40%) Homework: 4 homework in total;
- (40%) Project
- (20%) Final: See school schedule;



# What is machine learning:

- The study of computer algorithms that improve automatically through experience



# What is machine learning:

- The study of computer algorithms that improve automatically through experience
- 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  $E$ .”

  - experience  $E$ : training data
  - task  $T$ : improve decision in prediction, classification, clustering etc
  - performance measure  $P$ : loss function



# What is machine learning:

- The study of computer algorithms that improve automatically through experience

- 



# What is machine learning:

- The study of computer algorithms that improve automatically through experience

- 



- What is its relationship with AI, Data Science, Data Mining and Statistics?





# While there is overlap, there are differences

- Statistics: the goal is the understanding of the data at hand
- Artificial Intelligence: the goal is to build an intelligent agent
- Data Mining: the goal is to extract patterns from large-scale data
- Data Science: the science encompassing collection, analysis, and interpretation of data



# Learning tasks:

ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTC  
GATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACG  
CTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATTCGATAAATC  
GGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGC  
AATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCA  
ATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGAT  
AACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTG  
AGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGA  
GCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGA  
GATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGAT  
AGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCT  
GAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCAATT  
CGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAAC  
GCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAACGCTGAGCAATTC  
CTGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCA  
AATTCGGATATCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAACGCTGAGCAACGCTGAGCA  
ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGAT  
AGCATTTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATCGGATAACGCTGAGCA  
AATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCA  
ATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAGCAATTCGAT  
AGCAATTCGATAACGCTGAGCAATTCGGATAGCAATTCGATAGCAATTCGATAGCAATTCGATAG  
GCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAGCAATTCGATAGCAATTCGATAGCA  
GATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAGCAATTCGATAGCAATTCGATAG  
CTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGATAGCAATTCGATAGCAATTCGATAG  
TGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAAC  
TTCGATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTC  
GATAGCAATTCGATAACGCTGAGCAATTCGGATAACGCTGAGCAATTCGATAGCAATTCGATAAC  
GCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGATATCGATAGCA  
ATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATTCGGAT  
AACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATA

Which part is the gene?

Supervised and  
unsupervised learning (can  
also use active learning)

# Learning tasks:

Predict stock price

Help making trading decisions!

Market Summary > Nasdaq Composite  
INDEXNASDAQ: .IXIC

12,888.28 +18.28 (0.14%) ↑

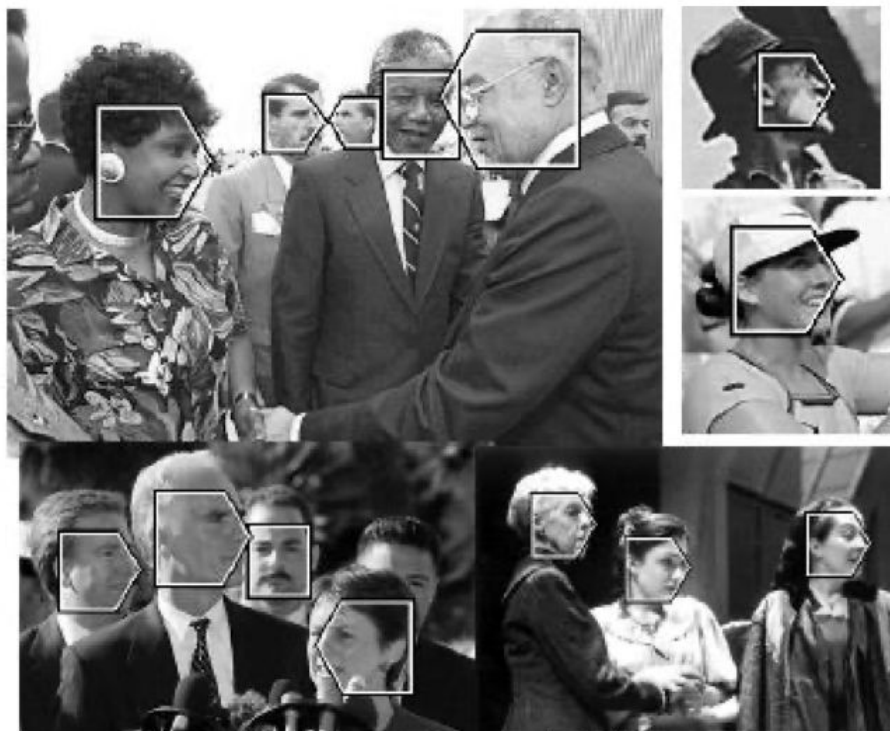
Dec 31, 5:15 PM EST · Disclaimer

1 day 5 days 1 month 6 months YTD 1 year 5 years Max



Supervised  
learning

# Learning tasks:



# Learning tasks:

## NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.





















So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these beliefs with a confidence score. NELL has high confidence in 3,938,530 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or [@cmunell on Twitter](#), browse and download its [knowledge base](#), read more about our [technical approach](#), or join the [discussion group](#).



semi supervised learning

## Recently-Learned Facts

Refresh

Instance	iteration	date learned	confidence
<a href="#">glass_window_restoration</a> is a <a href="#">household item</a>	1069	03-aug-2017	97.5  
<a href="#">bracelets_curb</a> is a kind of <a href="#">clothing</a>	1069	03-aug-2017	90.9  
<a href="#">hillsborough_lista_d_attesa_crea_un_gruppo_meetup</a> is a <a href="#">visualizable thing</a>	1069	03-aug-2017	99.1  
<a href="#">parison Levitra Viagra Cialis</a> is a <a href="#">drug</a>	1069	03-aug-2017	97.7  
<a href="#">the_democratic_daily</a> is a <a href="#">newspaper</a>	1069	03-aug-2017	100.0  
<a href="#">barcelona_international_airport</a> is an airport <a href="#">in the city barcelona</a>	1073	22-aug-2017	100.0  
<a href="#">john003</a> has brother <a href="#">james</a>	1073	22-aug-2017	100.0  
<a href="#">omaha_world_herald</a> is a newspaper <a href="#">in the city new york</a>	1073	22-aug-2017	93.8  
<a href="#">abc</a> is a company <a href="#">headquartered in the city new york</a>	1073	22-aug-2017	100.0  
<a href="#">arachnids001</a> is an arthropod <a href="#">as well as mites</a> also is	1073	22-aug-2017	93.8  



# Learning tasks:

- Speech recognition, Natural language processing
- Computer vision
- Web forensics
- Medical outcomes analysis Robotics
- Sensor networks
- Social networks
- ...
- Many, many more...



# Data

- Observations:
  - fully observed
  - partially observed: censored data, hidden states, etc...
  - designed experiments
  - actively collected data



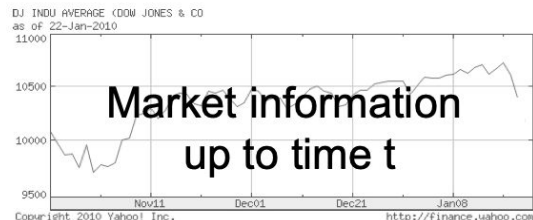
# Task

- Supervised learning, or also called prediction:
  - Regression - given input, estimate output
  - Classification - given input, estimate category
- Unsupervised learning:
  - Data only contains inputs, but no “supervision” in data as to the descriptive outputs
    - density estimation
    - clustering
    - dimensionality reduction



# Supervised learning:

Feature Space  $\mathcal{X}$



Label Space  $\mathcal{Y}$

→

“Sports”  
“News”  
“Science”  
...

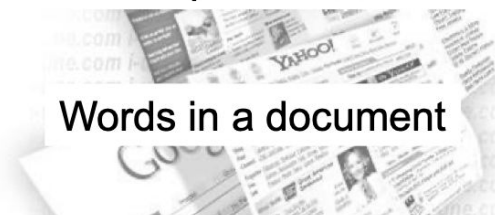
→

Share Price  
“\$ 24.50”

**Task:** Given  $X \in \mathcal{X}$ , predict  $Y \in \mathcal{Y}$ .

# Supervised learning:

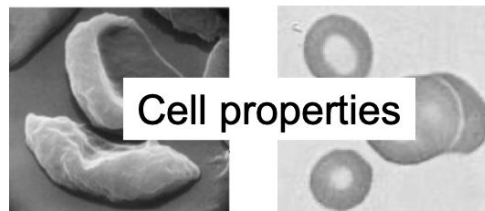
Feature Space  $\mathcal{X}$



Label Space  $\mathcal{Y}$



"Sports"  
"News"  
"Science"  
...



"Anemic cell"  
"Healthy cell"

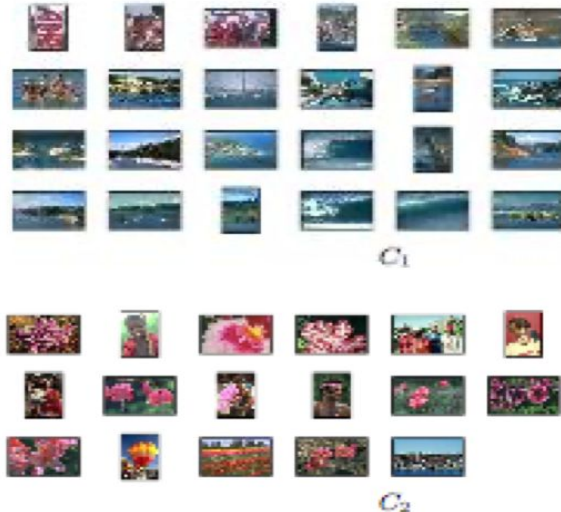
**Task:** Given  $X \in \mathcal{X}$ , predict  $Y \in \mathcal{Y}$ .



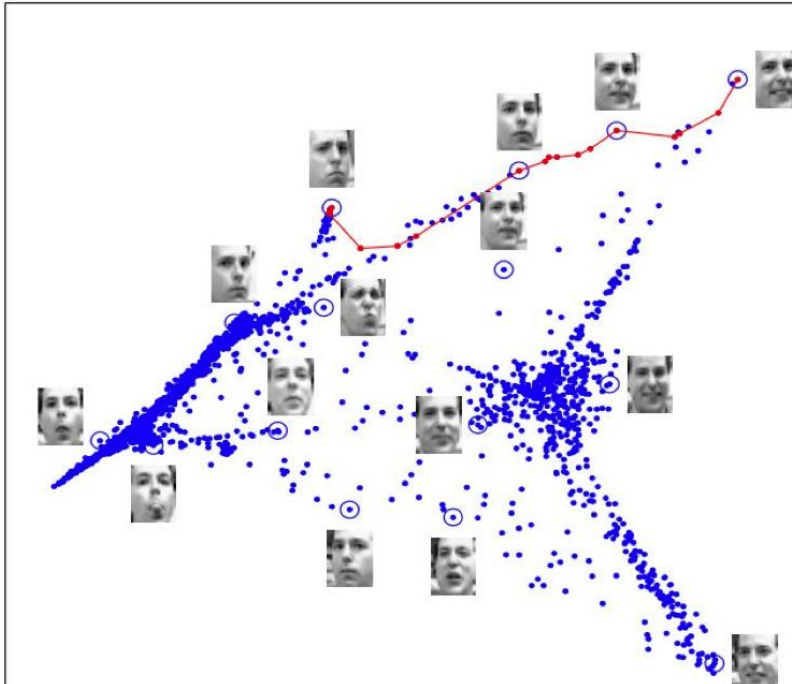
# Unsupervised learning:

Group similar things e.g. images

[Goldberger et al.]



# Unsupervised learning:



Facial recognition  
dimensionality reduction

[Saul & Roweis '03]



# Task

- Supervised learning:  
Given a set of features and labels learn a model that will predict a label to a new feature set
- Unsupervised learning  
Discover patterns in data
- Reasoning under uncertainty  
Determine a model of the world either from samples or as you go along
- Active learning  
Select not only model but also which examples to use



# Algorithm



- Model-based methods
  - probabilistic model of the data
  - parametric models
  - non-parametric models
- Model-free methods

# Model-based algorithm

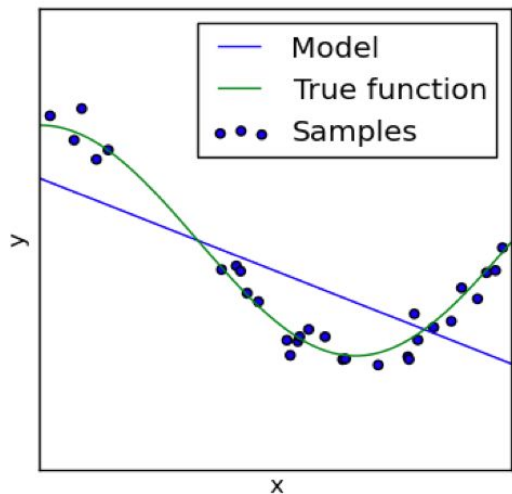


- Learning: from data to model
  - build a model to summarize or to generate the data
  - get estimation on model parameters
  - thus know how to generate future data
- Inference: from model to data
  - Given the model, how can we answer questions relevant to us



# Parametric model

- Fixed size model that the number of parameters does not grow with the data
- More data  $\rightarrow$  better fit of the model

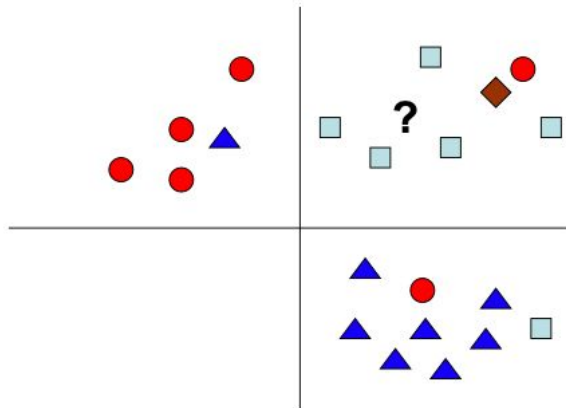


Fitting a simple line (2 params)  
to a bunch of one-dim. samples

Model: data = point on line + noise

# Non-parametric model

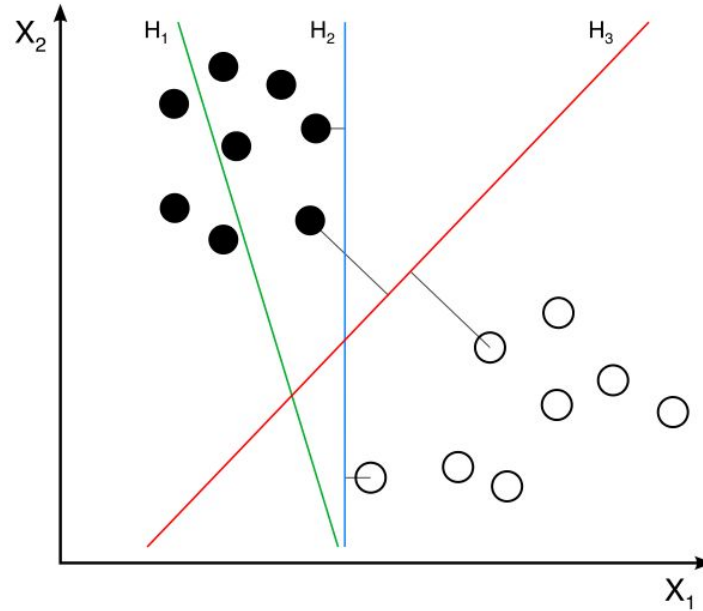
- The number of parameters grows with the data
- More data → a more complex model



- What is the class of the ?  
Input
- Can use the other points (k nearest neighbors) but the number of points to search scales with the input data

# Discriminative model

- Find the best line that separates black points from white points
- No generative assumption





# Common topics

- Mathematical framework:  
Well defined concepts based on explicit assumptions
- Representation:  
How to encode/decode text? Images?
- Model selection  
Which model should we use? How complex should it be?
- Use of prior knowledge  
How do we take our beliefs into consideration? How much can we assume?



# Theoretical foundation: probability

- In order to translate our task into formal mathematical problem, we need the language of

Probability : the study of uncertainty



# A brief introduction to probability

- Random variables  
refer to an event whose status is unknown:
  - A = “the stock price of google is going to increase by 0.1% tomorrow”: binary
  - A = “the app you use for food delivery” : discrete
  - A = “the chance of snow in NYC tomorrow” : continuous
- The set of all possible outcomes
  - All of the possible outcomes a random variable can take





# Probability

A variety of useful facts can be derived from just three axioms:

1.  $0 \leq P(A) \leq 1$
2.  $P(\text{true}) = 1, P(\text{false}) = 0$
3.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



# Joint probability

$P(A, B)$

If we assume independence, then  $P(A, B) = P(A) P(B)$

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0





# Joint probability

$P(A, B)$

If we assume independence, then  $P(A, B) = P(A) P(B)$

$P[\text{snow tomorrow}] = \frac{1}{2}$

$P[\text{snow today}] = \frac{1}{2}$

$P[\text{snow today and tomorrow}] = ?$

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0



# Joint probability

$P(A, B)$

If we assume independence, then  $P(A, B) = P(A) P(B)$

$P[\text{snow tomorrow}] = \frac{1}{2}$

$P[\text{snow today}] = \frac{1}{2}$

$P[\text{snow today and tomorrow}] = \frac{3}{8}$

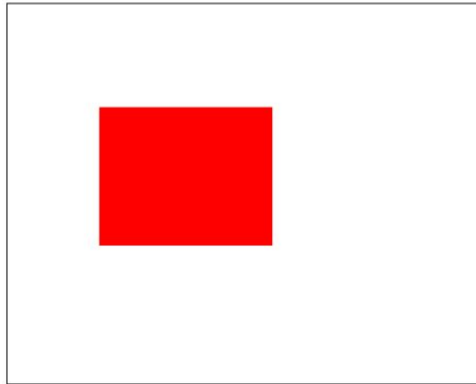
Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0



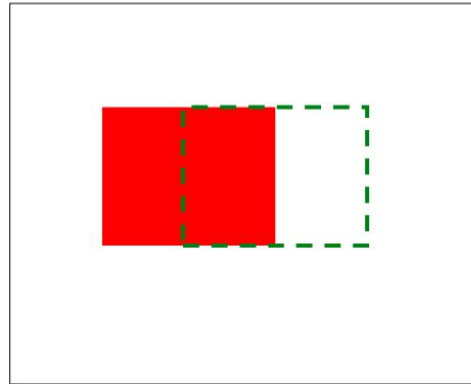
# Conditional probability

$P(A | B)$ : The fraction of cases where  $A$  is true if  $B$  is true

$P(A = 0.2)$



$P(A|B = 0.5)$





# Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- $P[\text{snow tomorrow}] = \frac{1}{2}$   
 $P[\text{snow tomorrow} \mid \text{snow today}] = \frac{3}{4}$   
 $P[\text{no snow tomorrow} \mid \text{snow today}] = \frac{1}{4}$

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0



# Chain rule

- The joint probability can be calculated in terms of conditional probability:

$$P(A,B) = P(A|B) P(B)$$

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning

# Bayes rule

- Derive from chain rule
- One of the most important rules for this class

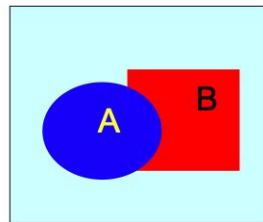
Often it would be useful to derive the rule a bit further:

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

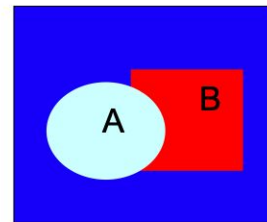
This results from:  
 $P(B) = \sum_A P(B,A)$



$P(B,A=1)$



$P(B,A=0)$



# An example

- Suppose you have a coin, if I flip it, what's the probability it will fall with the head up?
- You might want to flip the coin several times



# An example

- Suppose you have a coin, if I flip it, what's the probability it will fall with the head up?
- You might want to flip the coin several times

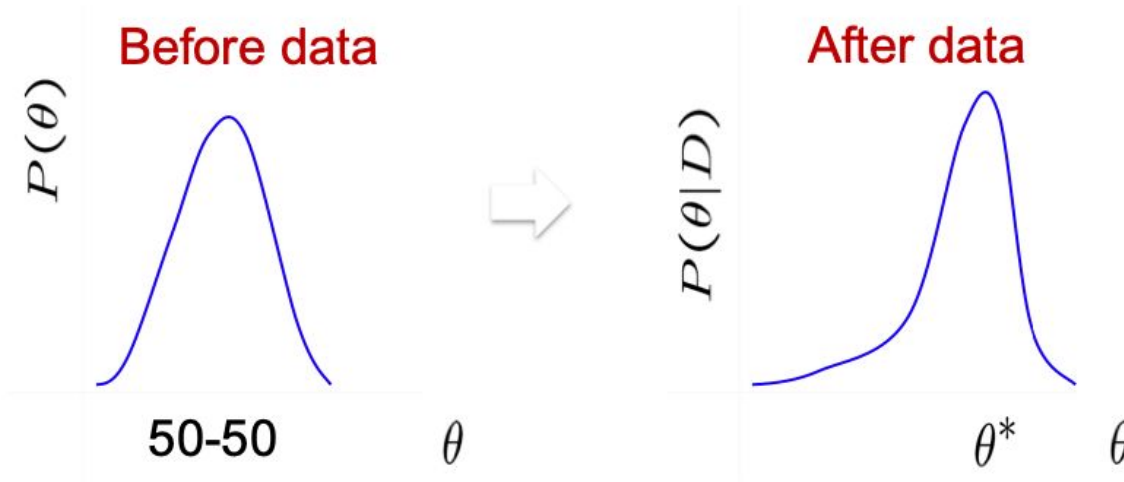


- The probability is  $\frac{3}{5}$  because frequency of heads in all flips
- Would you bet money on this estimation?



# What about your prior knowledge?

- Rather than estimating a single parameter, we obtain a distribution over possible values of this parameter





# Bayesian learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

posterior      likelihood      prior



# Prior distribution

- Beliefs in an event in the absence of any other information
- Source of prior:
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer's approach)
- Uninformative priors:
  - Uniform distribution
  - inappropriate distribution
- Conjugate priors:
  - closed-form representation of posteriors

# Conjugate prior

## Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$



# Conjugate prior

## Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

$$P(\mathcal{D} | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Then posterior is Beta distribution

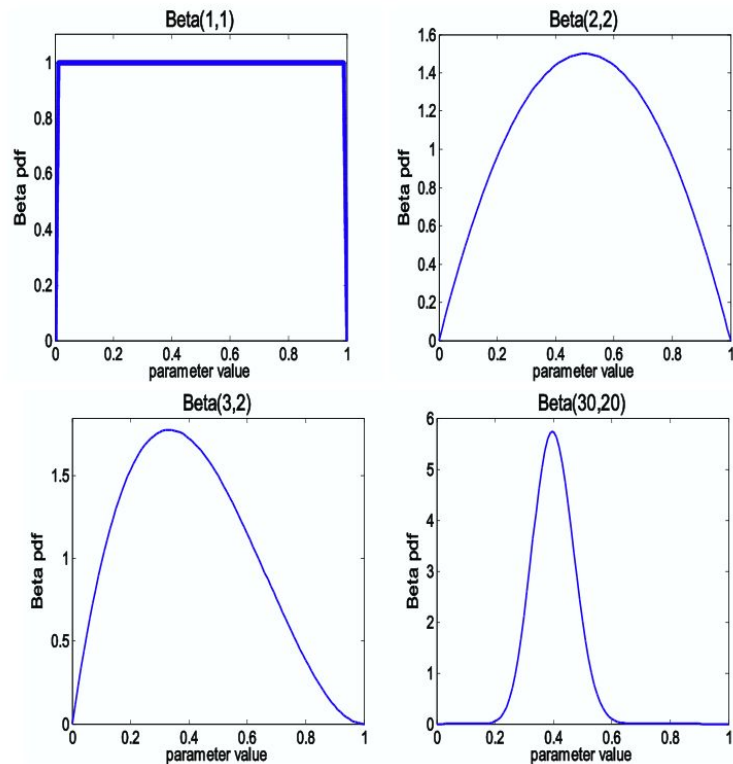
$$P(\theta | D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



# Conjugate prior

Beta prior:

$Beta(\beta_H, \beta_T)$  More concentrated as values of  $\beta_H, \beta_T$  increase

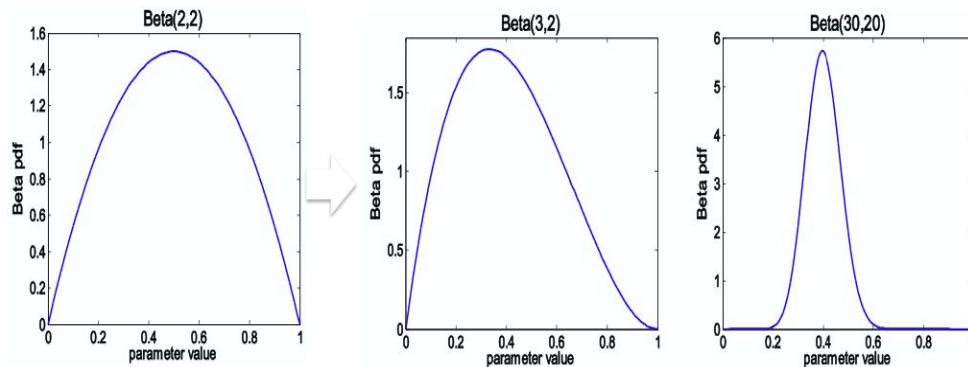


# Conjugate prior

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

$$P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

Posterior:



As  $n = \alpha_H + \alpha_T$   
increases

As we get more samples, effect of prior is “washed out”



# Conjugate prior:

- Gaussian prior + Gaussian sample distribution → Gaussian posterior
- Beta prior + Bernoulli sample distribution → Beta posterior
- Gamma prior + exponential sample distribution → Gamma posterior
- Dirichlet prior + multinomial sample distribution → Dirichlet posterior







# Posterior distribution

- The approach seen so far is what is known as a Bayesian approach
- Prior information encoded as a distribution over possible values of parameter
- Using the Bayes rule, we can get an updated posterior distribution over parameters



# Maximum likelihood principle (MLE)

Data likelihood:  $P(D | M) = q^{n_1} (1 - q)^{n_2}$

We would like to find:  $\arg \max_q q^{n_1} (1 - q)^{n_2}$

Or more generally,  $\hat{P}(\text{dataset} | M) = \hat{P}(x_1 \wedge x_2 \dots \wedge x_n | M) = \prod_{k=1}^n \hat{P}(x_k | M)$

- Our goal is to determine the values for the parameters in M
- We can do this by maximizing the probability of generating the observed samples



## An example: Coin flips



# MLE v.s. MAP

- Maximum Likelihood estimation (MLE):

Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

- Maximum a posteriori (MAP) estimation:

Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta|D) \\ &= \arg \max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$



# MAP: Coin flips



# References

- Tom Mitchell: Machine Learning, Chapter 6
- Kevin Murphy: Machine Learning: A probabilistic perspective, Chapter 1, 2, 5, 6
- Ziv Bar-Joseph, Tom Mitchell, Pradeep Ravikumar and Aarti Singh: CMU 10-701