COMP7180: Quantitative Methods for DAAI

(Credits from Prof. Andrew Ng)

(Credits from HKBU)

Course Instructors: Dr. Yang Liu and Dr. Bo Han

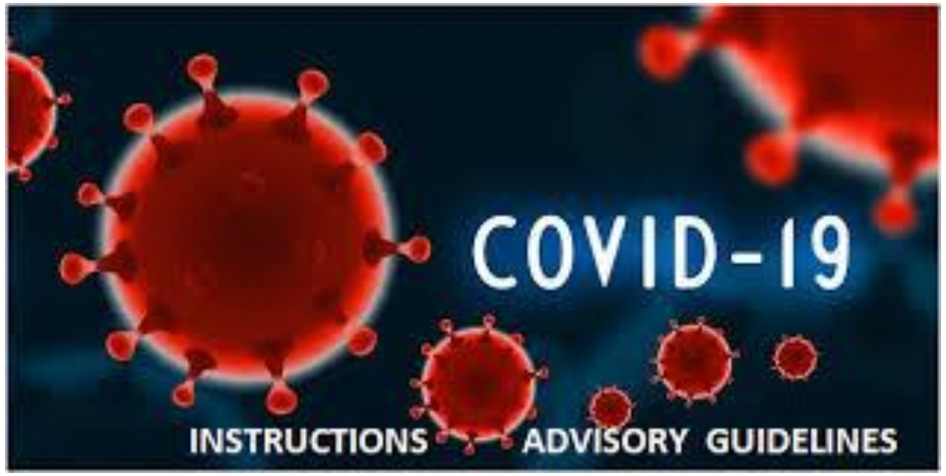
Teaching Assistant: Mr. Minhao Li

Course Contents

- Continuous and Discrete Random Variables (Week 7)
- Conditional Probability and Independence (Week 8)
- Maximum Likelihood Estimation (Week 9) Our Focus
- Mathematical Optimization (Week 10)
- Convex and Non-Convex Optimization (Week 11)
- Course Review (Week 12)

Review: Exercise

Consider following table of counts that are obtained from an observed sample of individuals both males and females, who had taken Covid-19. X, Y, Z are random variables. Z represents the gender, X represents whether individuals recover from Covid-19, and Y represents whether individuals have been treated.



Z = 1 Male; Z = 0 Female;
X = 1 Recovery (康复）； X = 0 No Recovery ;
Y = 1 Treatment（治疗）； Y = 0 No Treatment;

	Z=1, X=1	Z=1, X=0	Z=0, X=1	Z=0, X=0
Y=1	245	105	315	735
Y=0	630	420	70	280

245, 105, 315, 735, 630,420,70, 280 rpresents the number of individuals corresponding to different values of X, Y and Z.

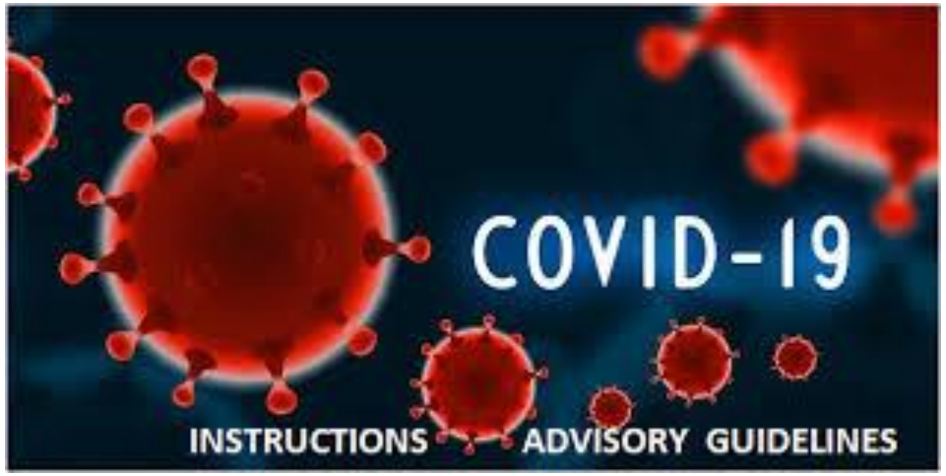
Review: Exercise

Please compute that

$P(X=1|Y=1,Z=1)$ $\quad$ $P(X=1|Y=0,Z=1)$

$P(X=1|Y=1,Z=0)$ $\quad$ $P(X=1|Y=0,Z=0)$

$P(X=1|Y=1)$ $\quad$ $P(X=1|Y=0)$



	Z=1, X=1	Z=1, X=0	Z=0, X=1	Z=0, X=0
Y=1	245	105	315	735
Y=0	630	420	70	280

Review: Exercise

Solution:

$P(X=1|Y=1,Z=1)=P(X=1,Y=1,Z=1)/P(Y=1,Z=1)=P(X=1,Y=1,Z=1)/(P(X=1,Y=1,Z=1)+P(X=0,Y=1,Z=1))=245/(245+105) = 0.7$

$P(X=1|Y=0,Z=1)=P(X=1,Y=0,Z=1)/P(Y=0,Z=1)=P(X=1,Y=0,Z=1)/(P(X=1,Y=0,Z=1)+P(X=0,Y=0,Z=1))= 630/(630+420) = 0.6$

$P(X=1|Y=1,Z=1)$ means the the recovery probability for individuals who are male and have been treated.

$P(X=1|Y=0,Z=1)$ means the the recovery probability for individuals who are male and have not been treated.

Review: Exercise

Solution:

$P(X=1|Y=1,Z=0)=P(X=1,Y=1,Z=0)/P(Y=1,Z=0)=P(X=1,Y=1,Z=0)/(P(X=1,Y=1,Z=0)+P(X=0,Y=1,Z=0))=315/(315+735) = 0.3$

$P(X=1|Y=0,Z=0)=P(X=1,Y=0,Z=0)/P(Y=0,Z=0)=P(X=1,Y=0,Z=0)/(P(X=1,Y=0,Z=0)+P(X=0,Y=0,Z=0))=70/(70+280) = 0.2$

$P(X=1|Y=1,Z=0)$ means the the recovery probability for individuals who are female and have been treated.

$P(X=1|Y=0,Z=0)$ means the the recovery probability for individuals who are female and have not been treated.

Review: Exercise

Solution:

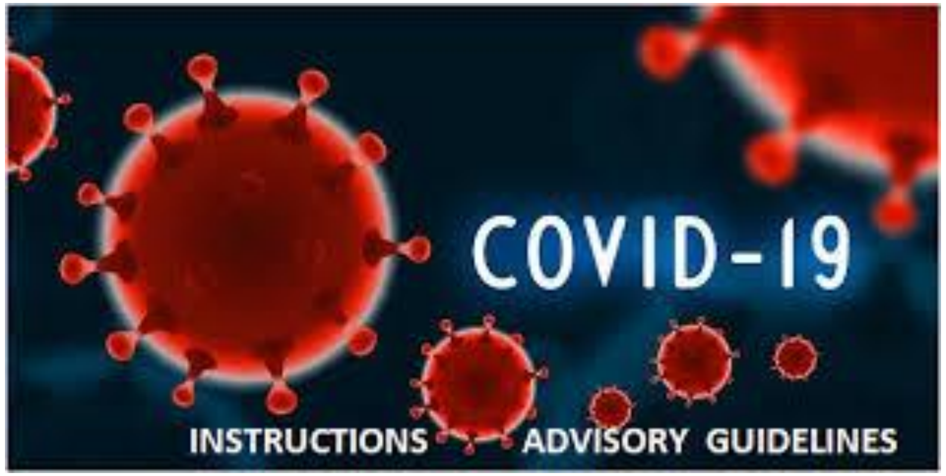
$P(X=1|Y=1)=P(X=1,Y=1)/P(Y=1)=(245+315)/(245+105+315+735) = 0.4$

$P(X=1|Y=0)=P(X=1,Y=0)/P(Y=0)=(630+70)/(630+420+70+280)=0.5$

$P(X=1|Y=1)$ means the the recovery probability for individuals who have been treated.

$P(X=1|Y=0)$ means the the recovery probability for individuals who have not been treated.

Review: Exercise



$P(X=1|Y=1,Z=1)=0.7>P(X=1|Y=0,Z=1)=0.6,$
$P(X=1|Y=1,Z=0)=0.3>P(X=1|Y=0,Z=0)=0.2.$

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated.

- The recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

Review: Exercise

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated.

- The recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

- Does this mean that the treatment can make postive affects?

That is: is the recovery probability for individuals who have been treated larger than the recovery probability for individuals who have not been treated?

Review: Exercise

To answer this question, we need to compute

$$P(X=1|Y=1) \text{ and } P(X=1|Y=0)$$

Then, we need to compare them.

We discover that $P(X=1|Y=1)=0.4 < P(X=1|Y=0)=0.5$.

- That is: the recovery probability for individuals who have been treated is smaller than the recovery probability for individuals who have not been treated.

Review: Exercise

It seems that

$P(X=1|Y=1,Z=1)>P(X=1|Y=0,Z=1)$,
$P(X=1|Y=1,Z=0)>P(X=1|Y=0,Z=0)$
and

$P(X=1|Y=1)<P(X=1|Y=0)$ are conflict.

- The recovery probability for individuals who are male and have been treated is larger than the recovery probability for individuals who are male and have not been treated;

  and
- the recovery probability for individuals who are female and have been treated is larger than the recovery probability for individuals who are female and have not been treated.

But

- the recovery probability for individuals who have been treated is smaller than the recovery probability for individuals who have not been treated

Why does this happen?

It is called Simpson's paradox.

Review: Exercise

Note that

$P(Z=1|Y=1)P(X=1|Y=1,Z=1)+P(Z=0|Y=1)P(X=1|Y=1,Z=0)=P(X=1|Y=1)$

$P(Z=1|Y=0)P(X=1|Y=0,Z=1)+P(Z=0|Y=0)P(X=1|Y=0,Z=0)=P(X=1|Y=0)$

Although $P(X=1|Y=1,Z=1)>P(X=1|Y=0,Z=1)$, $P(X=1|Y=1,Z=0)>P(X=1|Y=0,Z=0)$, the conditional probabilities $P(Z=1|Y=1)$, $P(Z=0|Y=1)$, $P(Z=1|Y=0)$ and $P(Z=0|Y=0)$ can affect the values of $P(X=1|Y=1)$ and $P(X=1|Y=0)$.

That is the basic reason why the Simpson's paradox happens.

Review: Exercise

Detailly,

let $u = P(Z=1|Y=1)$ and $v = P(Z = 1|Y = 0)$, then if we hope that

$$P(X=1|Y=1)>P(X=1|Y=0)$$

We need the following inequality:

$$0.7u + 0.3(1 - u) > 0.6v + 0.2(1 - v).$$

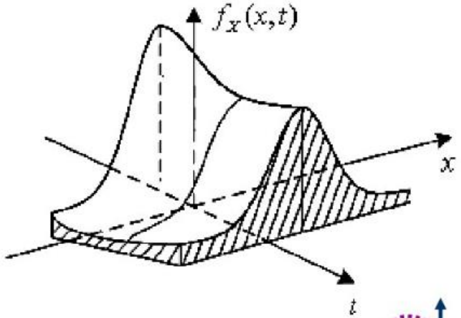
Whether the inequality can success depends on the values of u and v.

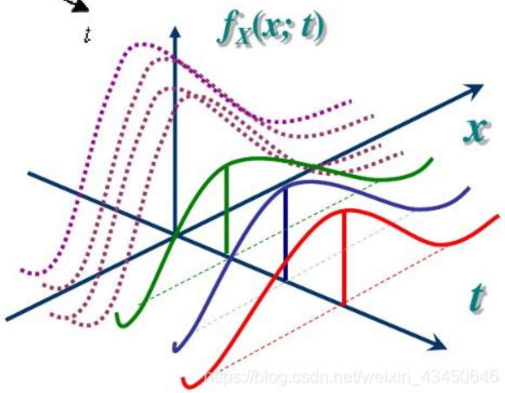
$$0.7u+0.3(1-u)>0.6v+0.2(1-v) \text{ if and only if } v-u<0.25.$$

But, $u=P(Z=1|Y=1)=0.25$ and $v =P(Z=1|Y=0)=0.75$. So $v-u=0.5>0.25$. That is the reason why $P(X=1|Y=1)<P(X=1|Y=0)$.

Review An Important Concept

We say n random variables X1, X2,...,Xn are Independent and identically distributed (i.i.d.), if





1. X1,X2,...,Xn are Independent with each other;

$$P(X1,X2,...,Xn) = P(X1)P(X2)...P(Xn);$$

2. X1,X2,...,Xn have same probability distribution;

$$P(X1)=P(X2)=...=P(Xn)$$

Review An Important Concept

We say n data x1, x2,...,xn are Independent and identically distributed (i.i.d.), if x1,x2,...,xn are drawn from n i.i.d. random variables X1, X2,...,Xn.

Therefore, if X1, X2,..., Xn are discrete random variables, then

P(X1=x1,X2=x2,...,Xn=xn) = P(X1=x1)P(X2=x2)...P(Xn=xn)= $\prod_{i=1}^{n} P(Xi = xi)$

We will give an example to help understant this concept.

An Important Concept: Example

Tossing a coin (n times). If the possibility to appear the head H is μ, then possibility to appear the tail (T) is $1-\mu$.

For each i=1,....,n,

Xi is the random variable that the i-th tossing a coin.

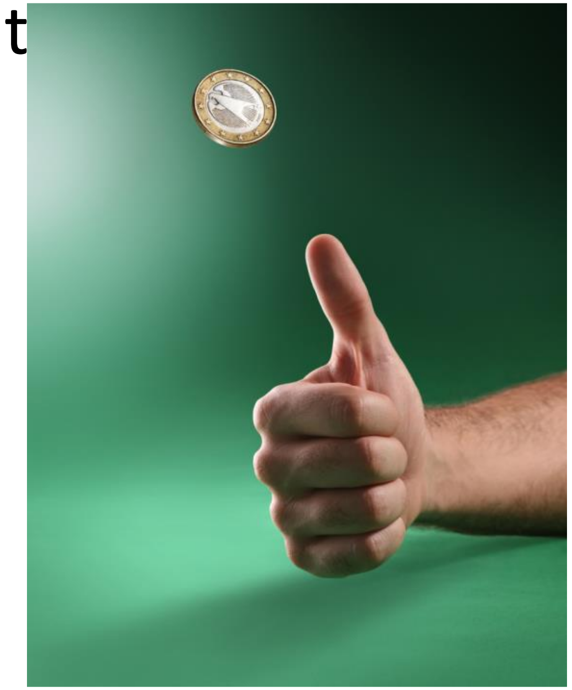
Let Xi(H) = μ and Xi(T) = $1-\mu$.

For each i=1,...,n, if the i-th tossing a coin will not be affected by other attempts, then random variables X1, X2,...,Xn are Independent and identically distributed (i.i.d.).

An Important Concept: Example

Tossing a coin (n times). If the possibility to appear the head H is μ, then possibility to appear the tail (T) is $1 - \mu$.

Let xi be the outcome of i-th tossing. That is x_i is the out



of the random variable Xi, then

$x_1, x_2, . . ., x_n$ are Independent and identically distributed (i.i.d.)

Maximum Likelihood (ML) Estimation

- To start our new topic, we introduce a simple question.

Tossing a coin. If the possibility to appear the head is μ, then fliping a coin is a Bernoulli Distribution with parameter μ.

Bernoulli(μ)

P(X=1) = μ

P(X=0) = 1-μ

X is the random variable:

X=1 means the head appears; X=0 means the tail appears.

Maximum Likelihood (ML) Estimation

Suppose that $x_1, x_2, \ldots, x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials (for example, coin flipping), each with success probability μ.

Question: assume μ is unknown, can we estimate μ by given data $x_1, x_2, \ldots, x_n$? **For which μ is $x_1, x_2, \ldots x_n$ most likely?**

Maximum Likelihood (ML) Estimation

Question: assume μ is unknown, can we estimate μ by given data $x_1, x_2, \ldots, x_n$? **For which μ is $x_1, x_2, \ldots x_n$ most likely?**

To address this issue, we introduce

Maximum Likelihood (ML) Estimation

Maximum Likelihood (ML) Estimation

- In many artificial intelligence and machine learning applications, the objective is **to estimate the model parameters from the given data**.

- For example, given a distribution class P(X;α), where α is a parameter from a parameter space. Now given data (x1,x2,...,xn) which are drawn from an unknown distribution P(X;α_0), we want to ask that how to select a suitable parameter α_0 by given data (x1,x2,...,xn)?

- The **Maximum Likelihood Estimation (MLE)** is one of the most widely used methods of estimating the parameters of a model.

Credicts from Dr. Liu Yang

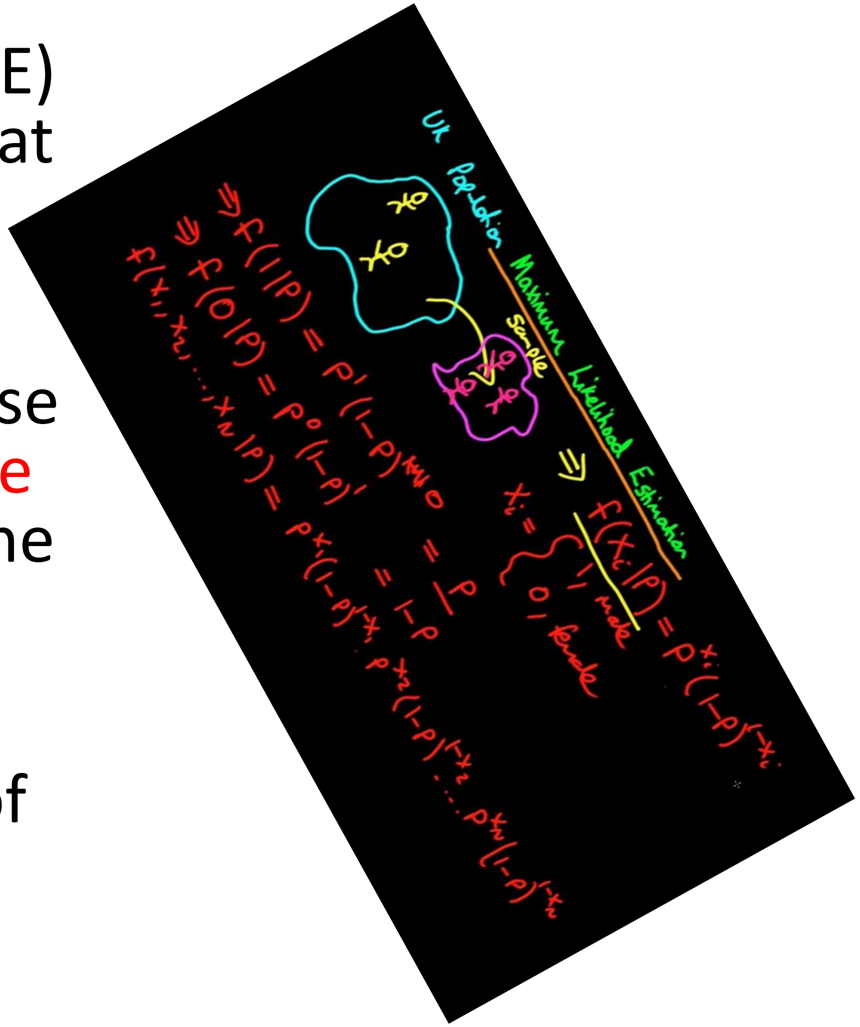
Maximum Likelihood (ML) Estimation

- The method of Maximum Likelihood Estimation (MLE) selects the set of values of the model parameters that maximizes the likelihood function.

- In other words, the basic principle of MLE is to choose values that “explain” the data best by maximizing the probability of the data we've seen as a function of the parameters.

- The **Maximum Likelihood Estimation (MLE)** is one of the most widely used methods of estimating the parameters of a model.

- It answers the question: What values of parameters would make the observations most probable ?



Maximum Likelihood (ML) Estimation

- A distribution class $P(X;\alpha)$, where α is from a parameter space Δ.
- For each α from the space Δ, $P(X;\alpha)$ corresponds to a distribution.
- We have data $S=(x1,x2,\ldots,xn)$, which are drawn from an unknown distribution $P(X)$, Independent and identically distributed (i.i.d.)

Problem: what is the optimal parameter α^* selected from the parameter space Δ, such that the seleced distribution $P(X;\alpha^*)$ is the most possible distribution sampling data S?

Maximum Likelihood (ML) Estimation: Example

- See the question introduced in the begining.
- Suppose that $x_1, x_2, . . ., x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials, each with success probability μ.

- The parameter ɑ is μ.
- The parameter space Δ is $\{\mu: 0<\mu<1\}$.
- Data $S=(x_1, x_2, . . ., x_n)$
- Distribution class $P(X;\mu)$ is: to each μ,
  $P(X=1;\mu) = \mu$; $P(X=0;\mu) = 1-\mu$
  So $P(X=x_i;\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$

Maximum Likelihood (ML) Estimation

The seleced distribution $P(X;\alpha^*)$ is the most possible distribution sampling data S=(x1,x2,...,xn), i.i.d.

Understanding above sentence, we can formulate it as follows:

$$\underset{\alpha\in\Delta}{\operatorname{argmax}} P(x1, x2, \ldots, xn; \alpha)$$

here we assume $P(X;\alpha)$ is a discrete distribution.

- $P(x1, x2, \ldots, xn; \alpha)$ means the largest probability for $P(X;\alpha)$ that S is observed.

Maximum Likelihood (ML) Estimation

Because (x1,…,xn), are Independent and identically distributed,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} P(x1, x2, \ldots, xn; \alpha)$$

is equal to

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

How to address the equation?

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha)$$

Step 1. Because above issue is difficult to compute, we take a small trick (Take log function).

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}}\ \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Step 2. Using the property of log function:

$$\log \prod_{i=1}^{n} P(X = xi; \alpha) = \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

Therefore,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Step 3. We need to optimize

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \textstyle\sum_{i=1}^{n} \log P(X = xi; \alpha) \quad (1)$$

and obtain the optimal solution.

The solution of Eq. 1 is called Maximum Likelihood Estimation.

Maximum Likelihood (ML) Estimation

If the distribution class consists of continuous distributions, that is P(X;α) is a continuous distribution with respect to all α∈ Δ.

Then the Maximum Likelihood Estimation is

$$\underset{\alpha\in\Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_X(xi;\alpha) \quad (2)$$

where $p_X(x;\alpha)$ is the density function of P(X;α).

Maximum Likelihood (ML) Estimation

How to obtain the solution of

$$\underset{\alpha\in\Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha) \ ?$$

- This is related to optimization problem.
- Generally, there are no unviersal approaches to give soultions to all Maximum Likelihood (ML) Estimation.
- The approaches are case by case.

Maximum Likelihood (ML) Estimation

In this class, we introduce a common used approach.

This approach is based on a simple theorem:

If 1) a function f(x1,x2,...,xd) is differentiable,
2) $x^* = (x1^*, x2^*, \ldots, xd^*)$ is the maximum point of f, then

$$\frac{\partial f}{\partial xi}(x1^*, x2^*, \ldots, xd^*) = 0.$$

Maximum Likelihood (ML) Estimation

Using this theorem, if $\sum_{i=1}^{n} \log P(X = xi; \alpha)$ is differentiable, then

Let $\alpha=(\alpha 1,\alpha 2,\ldots,\alpha d)$,

$$\frac{\partial \sum_{i=1}^{n} \log P(X=xi;\alpha)}{\partial \alpha i} = 0, \quad \text{for } i=1,\ldots,d$$

Then, addressing above equations.

Check that you've found a maximum rather than a minimum or saddle-point, and be careful if α belongs to Δ.

Exercises: MLE for Binomial Distribution

- Now we address our orignal issue.

Suppose that $x_1, x_2, \ldots, x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials (for example, coin flipping), each with success probability μ.

- $P(X=1;\mu) = \mu$
- $P(X=0;\mu) = 1-\mu$

So $P(X=x_i;\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$

MLE: For which μ is $x_1, x_2, \ldots x_n$ most likely?

Exercises: MLE for Binomial Distribution

Maximum Likelihood (ML) Estimation：

$$\underset{0\le\mu\le 1}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = x_i; \mu)$$

$$\sum_{i=1}^{n} \log P(X = x_i; \mu) = \sum_{i=1}^{n} \log \mu^{x_i} (1-\mu)^{1-x_i}$$

$$= \sum_{i=1}^{n} x_i \log \mu + \sum_{i=1}^{n} (1 - x_i) \log(1-\mu)$$

Credicts from Dr. Liu Yang

Exercises: MLE for Binomial Distribution

Derivation of MLE

$$\frac{\partial \sum_{\mathrm{i=1}}^{\mathrm{n}} \log \mathrm{P}(\mathrm{X=x_i};\mu)}{\partial \mu},$$

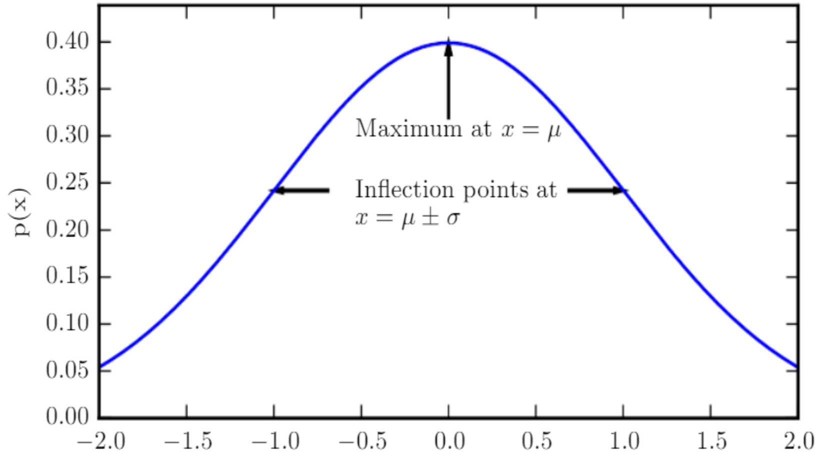
- $\frac{\partial \sum_{\mathrm{i=1}}^{\mathrm{n}} \log \mathrm{P}(\mathrm{X=x_i};\mu)}{\partial \mu} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\mathrm{x_i}}{\mu} - \sum_{i=1}^{n} \frac{(1-\mathrm{x_i})}{1-\mu} = 0.$
- We have that $\mu = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x_i} / \mathrm{n}$

Credicts from Dr. Liu Yang

Exercises: MLE for Gaussian Distribution

Suppose you have x1,x2,...,xn (i.i.d) $N(\mu, \sigma^2)$

$$\sqrt{\frac{1}{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(\mathrm{x}-\mu)^2)$$



- But you don't know μ (you do know σ^2)

MLE: For which μ is x1, x2, ..., xn most likely?

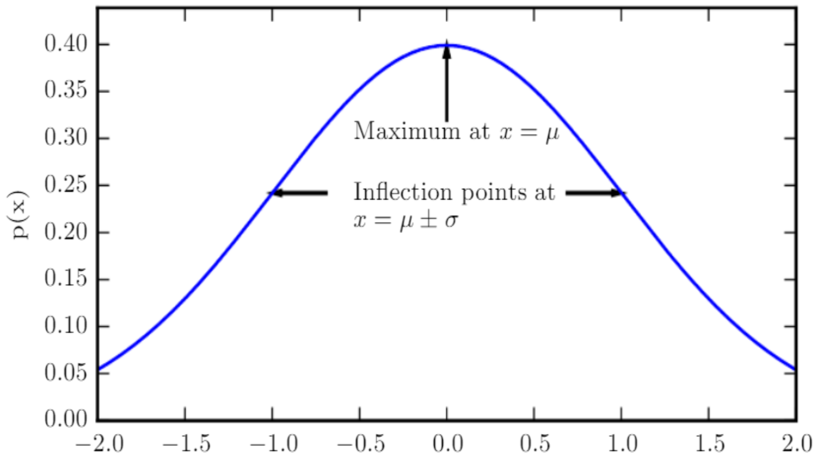
Credicts from Dr. Liu Yang

Exercises: MLE for Gaussian Distribution

Compute the MLE $\underset{\mu \in R}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_X(\text{xi}; \mu)$

$$\underset{\mu \in R}{\arg\max} \frac{1}{\sqrt{2\pi}\ \sigma} \sum_{i=1}^{n} -\frac{(\text{xi} - \mu)^2}{2\sigma^2}$$

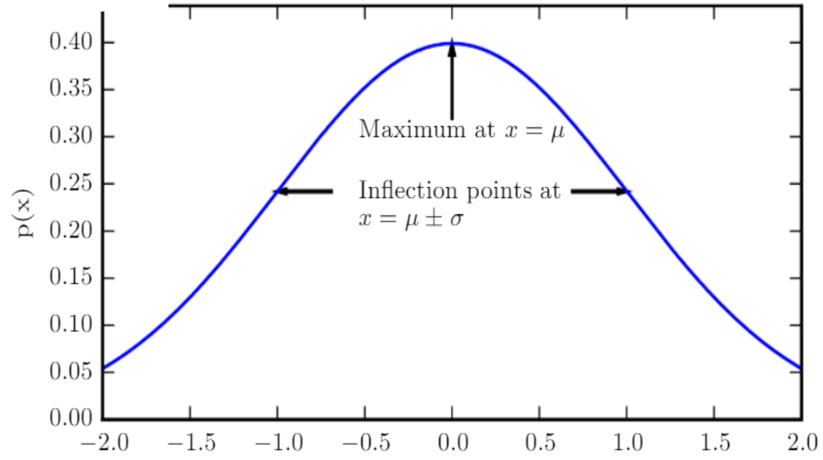
$$= \underset{\mu \in R}{\arg\min} \sum_{i=1}^{n} (\text{xi} - \mu)^2$$



Exercises: MLE for Gaussian Distribution

Derivation the equation $\arg\min_{\mu \in R} \sum_{i=1}^{n} (\text{xi} - \mu)^2$

$$\frac{\mathrm{d} \sum_{\mathrm{i}=1}^{n} (\mathrm{xi} - \mu)^2}{\mathrm{d}\mu} = 2 \sum_{\mathrm{i}=1}^{n} (\mathrm{xi} - \mu) = 0$$



So the solution is

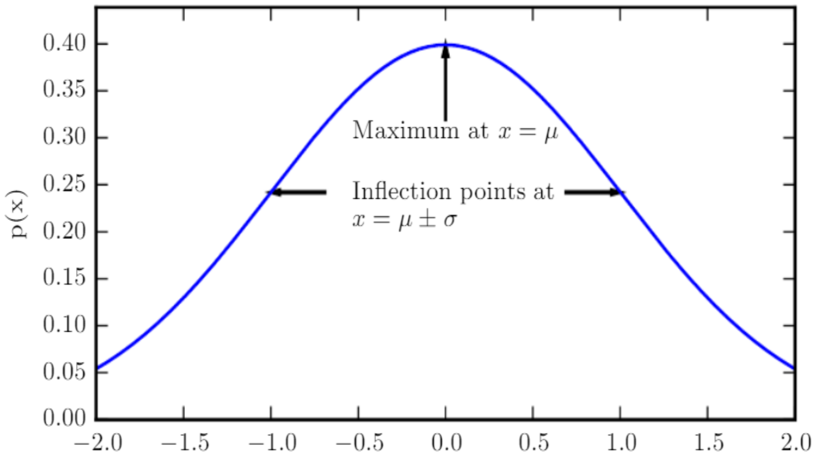
$$\mu = \frac{\sum_{i=1}^{n} \mathrm{xi}}{n}$$

Credicts from Dr. Liu Yang

Exercises: MLE for Gaussian Distribution

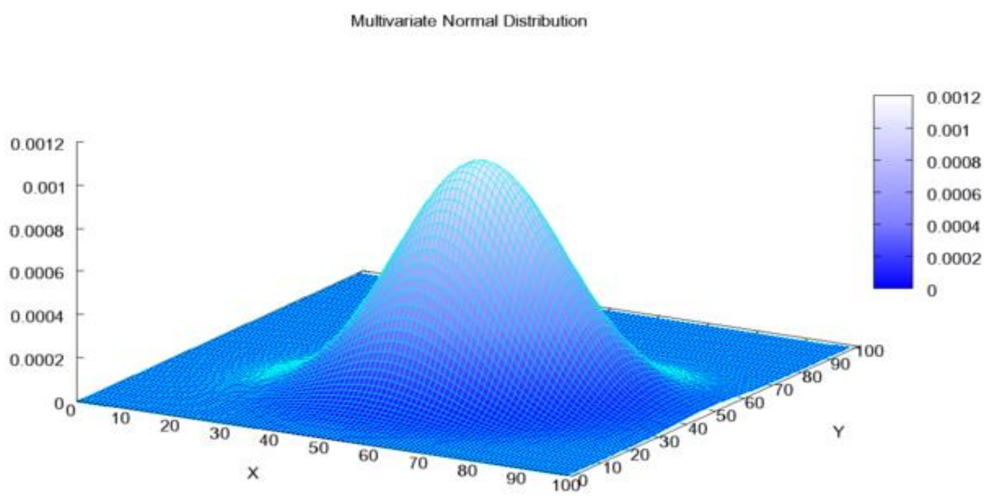
- In conclusion, the best estimate of the mean of a gaussian distribution is the mean of the sample!

$$\mu = \frac{\sum_{i=1}^{n} \mathrm{xi}}{n}$$



Exercises: MLE for high-dimensional Gaussian Distribution

- Given a 2 $\times$ 2 positive semi-definite matrix Σ and a 2 $\times$ 1 vector μ, a three dimensional normal distribution $N(\mu, \Sigma)$ can be represented as follows: the density function of this distribution is



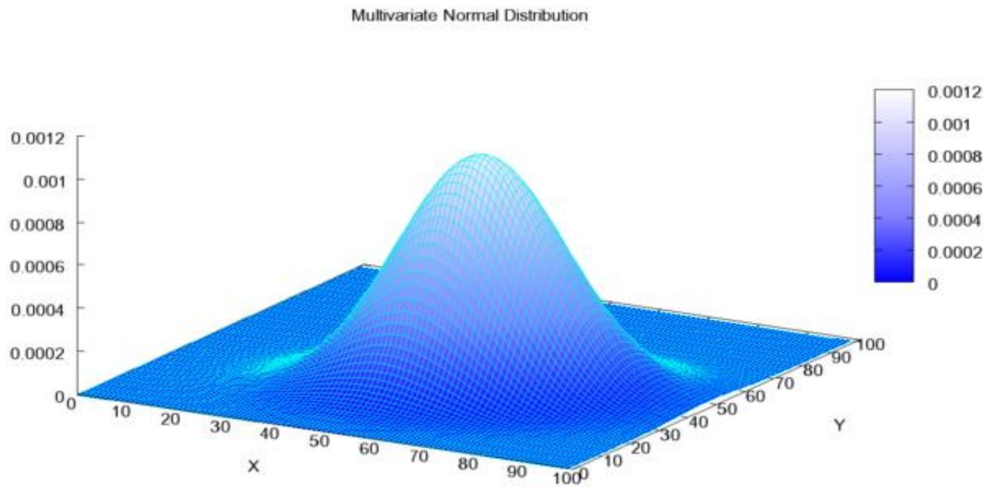
$$p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}$$

where $|\Sigma|$ is the determinant of Σ and $\mathbf{x}=(x, y)^{\mathbf{T}}$.

Exercises: MLE for high-dimensional Gaussian Distribution

- If μ=(a1,a2) and Σ is a diagonal matrix with eigenvalues λ1, λ2 (λ1 > 0, λ2 > 0),

$$\Sigma = \begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix}$$



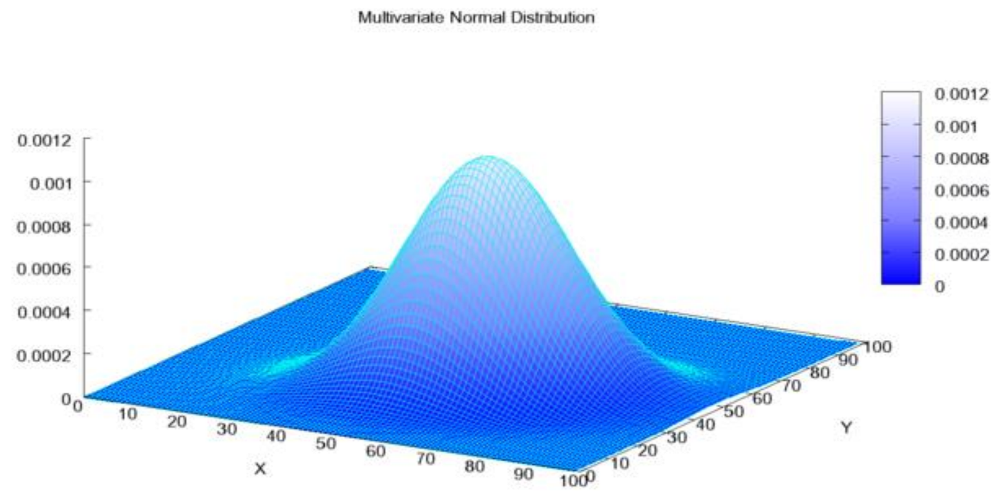
then $f(x, y)$ can be writeen as:

$$p_{XY}(x, y; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^2 \lambda 1 * \lambda 2}} e^{-\frac{1}{2\lambda 1}(x-a1)^2 - \frac{1}{2\lambda 2}(y-\mathrm{a}2)^2}$$

Exercises: MLE for high-dimensional Gaussian Distribution

- If $\Sigma = \begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix}$

and we have n data (x1,y1),...,(xn,yn) sampled from a two-dimensional Gaussian Distribution *N(μ, Σ)*, i.i.d., calculate μ by the maximum likelihood estimation method.



Exercises: MLE for high-dimensional Gaussian Distribution

- Maximum Likelihood (ML) Estimation：

$$\underset{\mu}{\operatorname{argmax}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \log p_{XY}(\mathrm{xi}, \mathrm{yi}; \mu, \Sigma)$$

It is equal to

$$\underset{\mathrm{a1,a2}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{1}{2\lambda 1}(\mathrm{xi} - \mathrm{a1})^2 + \frac{1}{2\lambda 2}(\mathrm{yi} - \mathrm{a2})^2\right)$$

Exercises: MLE for high-dimensional Gaussian Distribution

Derivation the equation $\mathrm{G}(\mathrm{a1}, \mathrm{a2}) = \sum_{i=1}^{n} (\frac{1}{2\lambda 1}(\mathrm{xi} - \mathrm{a1})^2 + \frac{1}{2\lambda 2}(\mathrm{yi} - \mathrm{a2})^2)$

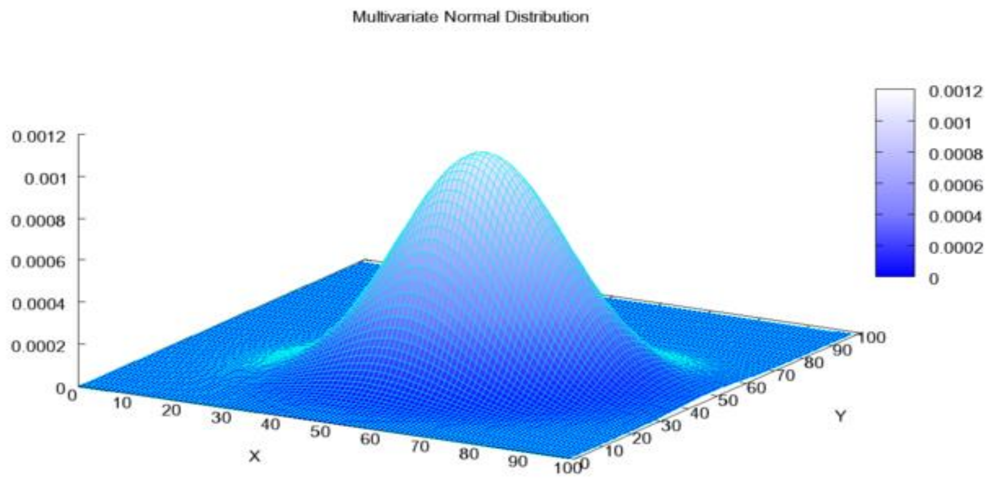
$$\frac{\partial \mathrm{G}}{\partial \mathrm{a1}} = \sum_{i=1}^{n} \frac{\mathrm{a1} - \mathrm{xi}}{\lambda 1} = 0, \qquad \frac{\partial \mathrm{G}}{\partial \mathrm{a2}} = \sum_{i=1}^{n} \frac{\mathrm{a2} - \mathrm{yi}}{\lambda 2} = 0$$

So a1 = $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{xi}$, a2 = $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{yi}$

Exercises: MLE for high-dimensional Gaussian Distribution

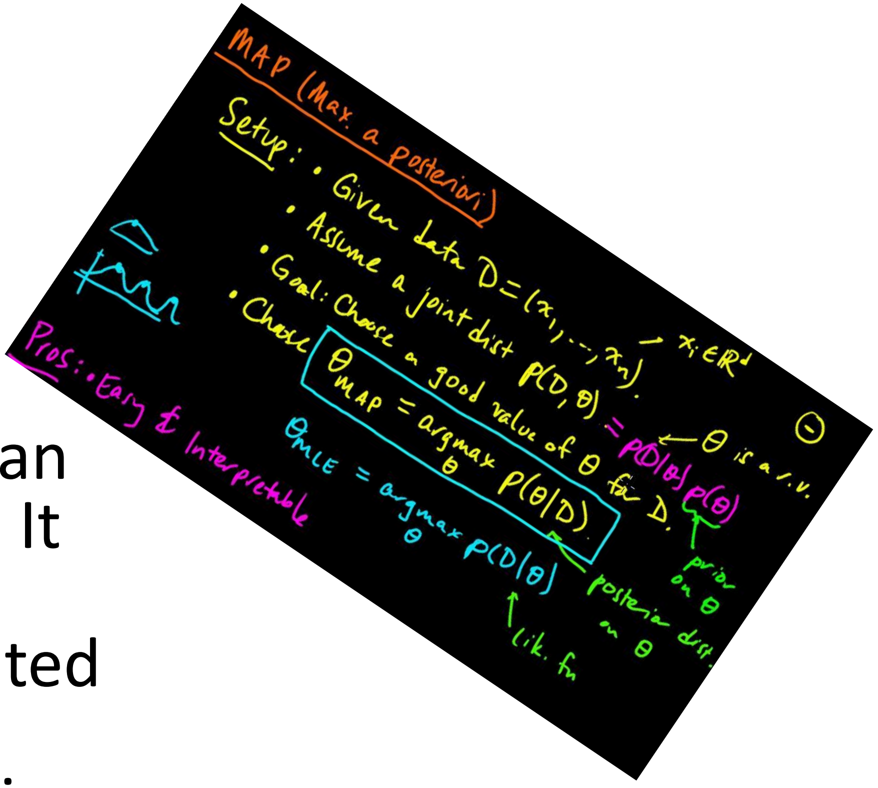
- In conclusion, the best estimate of the mean of a two-dimensional gaussian distribution is the mean of the sample!

In fact, it also holds for high-dimensional gaussian distribution.



Maximum A Posteriori (MAP) Estimation

- In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.



- The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to the method of maximum likelihood (ML) estimation, but employs an augmented optimization objective which incorporates a prior distribution (that quantifies the additional information available through prior knowledge of a related event) over the quantity one wants to estimate.

Maximum A Posteriori (MAP) Estimation

- Consider a distribution class P(X;ɑ).
- MAP regards the parameter ɑ as a random variable.
- Therefore, we can rewrite P(X;ɑ) as the conditional distribution P(X|ɑ).
- In MAP, we aim to estimate the parameter ɑ, given the data x1,...,xn i.i.d. from an unknown distribution $P(X;\alpha^*)$:

$$\underset{\alpha}{\operatorname{argmax}} \; P(\alpha|x1,x2,...,xn)$$

.

Maximum A Posteriori (MAP) Estimation

- In MAP, we mainly use Bayesian rule:

$$P(\alpha|x1,x2,\ldots,xn) = (P(x1,x2,\ldots,xn|\alpha)\ P(\alpha))/P(x1,x2,\ldots,xn)$$

- Note that when the data x1,x2,...,xn are given, P(x1,x2,...,xn) is a constant.

- Therefore,

$$\underset{\alpha}{\operatorname{argmax}}\ P(\alpha|x1,x2,\ldots,xn) = \underset{\alpha}{\operatorname{argmax}}\ P(x1,x2,\ldots,xn|\alpha)\ P(\alpha)$$

Maximum A Posteriori (MAP) Estimation

- $\underset{\alpha}{\operatorname{argmax}}\ P(x1,x2,\ldots,xn|\alpha)\ P(\alpha)$

is called maximum a poserior estimation.

Because x1,...,xn are Independent and identically distributed, then

$$\underset{\alpha}{\operatorname{argmax}}\ P(x1,x2,\ldots,xn|\alpha)\ P(\alpha) = \underset{\alpha}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = x\mathrm{i}|\alpha)P(\alpha)$$

- Compared with ML estimation, MAP estimation has a prior distribution P(α).
- MAP estimation: find the most likely parameter settings under the posterior.

Exercises: Maximum A Posteriori (MAP) Estimation

- Suppose we need to determine if a patient has a rare disease, given a laboratory test of that patient.

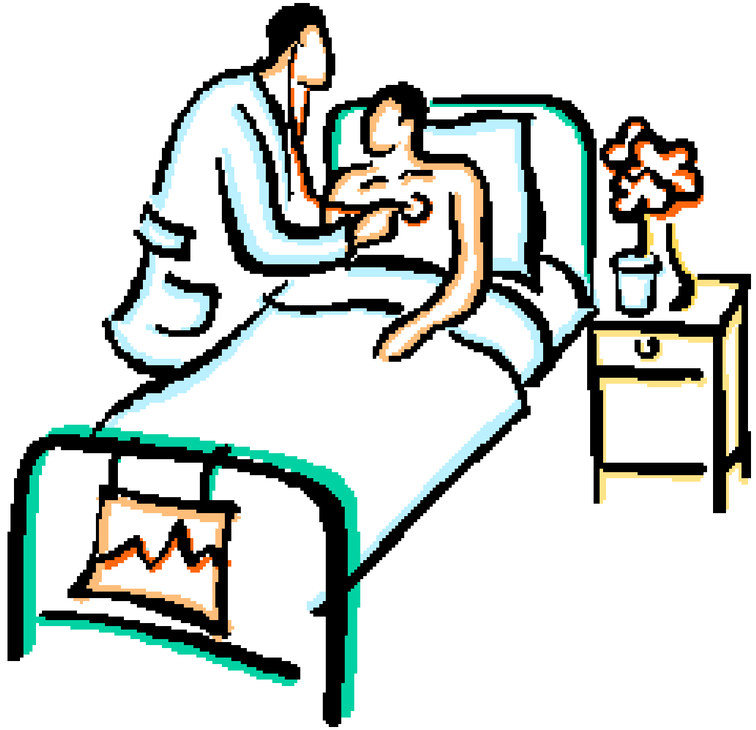
- We consider a set of two random variables: $\alpha=1$ (disease) and $\alpha=-1$ (no disease).

- X is the random variable related to the laboratory test. X=1 means the positive in the laboratory test, and X=0 means the negative in the laboratory test.

Credicts from Book Foundations of Machine Learning

Exercises: Maximum A Posteriori (MAP) Estimation

- Suppose that the disease is rare, say $P(\alpha=1)=0.005$
- The laboratory is relatively accurate: $P(X=1|\alpha=1) = 0.98$, $P(X=0|\alpha=0) = 0.95$.



- If the test is positive, what should be the diagnosis?

In other words, we have a data x which is equal to 1.

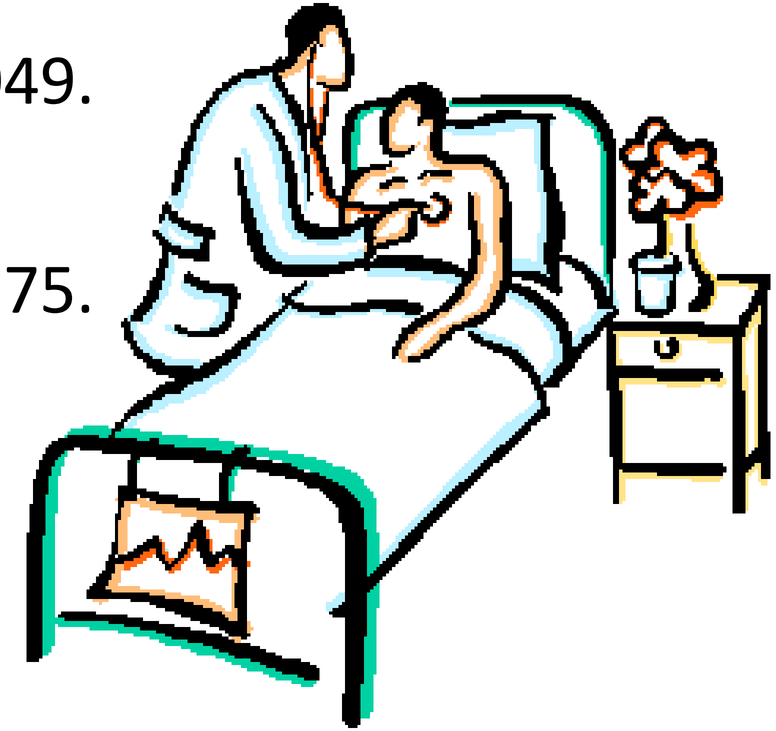
Credicts from Book Foundations of Machine Learning

Exercises: Maximum A Posteriori (MAP) Estimation

- Using maximum a poserior estimation

$$\underset{\alpha}{\operatorname{argmax}}\ P(X=1|\alpha)\,P(\alpha)$$

- If $\alpha = 1$, then $P(X=1|\alpha=1)\,P(\alpha=1) = 0.98*0.005 = 0.0049$.
- If $\alpha = 0$, then $P(X=1|\alpha=0)\,P(\alpha=0) = 0.05*0.995 = 0.4975$.



Because $P(X=1|\alpha=0)\,P(\alpha=0) > P(X=1|\alpha=1)\,P(\alpha=1)$, we obtain that the solution is

$$0 = \underset{\alpha}{\operatorname{argmax}}\ P(X=1|\alpha)\,P(\alpha)$$

Credicts from Book Foundations of Machine Learning

Exercises: Maximum A Posteriori (MAP) Estimation

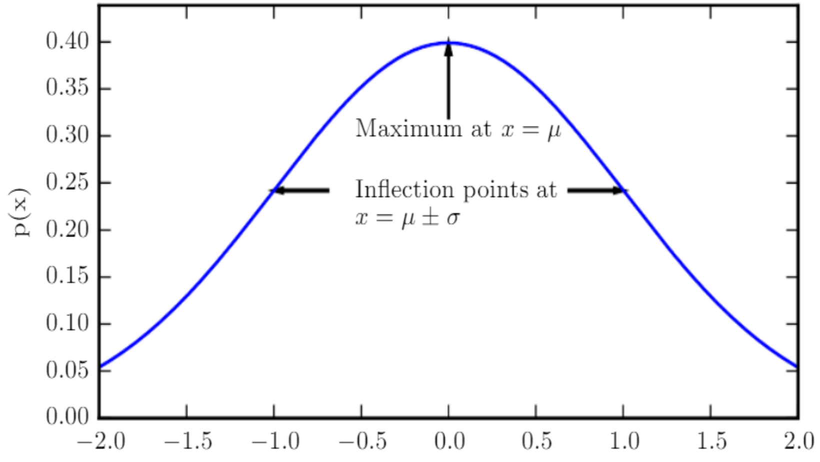
Thus, in this case, the MAP prediction is no disease: according to the MAP solution,with the values indicated, a patient with a positive test result is nonetheless more likely not to have the disease!

Exercises: MAP for Gaussian Distribution

Suppose you have x1,x2,...,xn (i.i.d) $N(\mu,\sigma^2)$ with density

$$p(x|\mu) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$p(\mu) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\mu-\mu_0)^2\right)$$



MAP: For which μ is?

Exercises: MAP for Gaussian Distribution

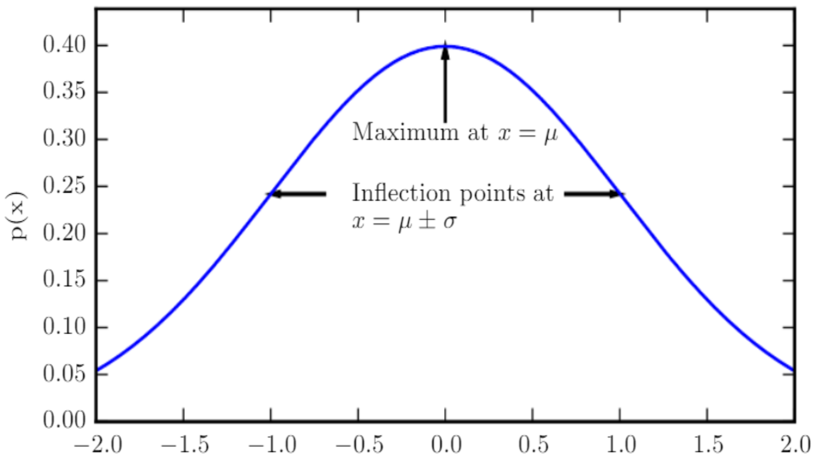
Using maximum a poserior estimation $\underset{\mu}{\operatorname{argmax}} \prod_{\mathrm{i=1}}^{\mathrm{n}} p(\mathrm{xi}|\, \mu)\mathrm{p}(\mu)$

$$= \underset{\mu}{\operatorname{argmax}} \prod_{\mathrm{i=1}}^{\mathrm{n}} \exp(-\frac{1}{2\sigma^2}(\mathrm{xi}-\mu)^2)\exp(-\frac{1}{2\sigma^2}(\mu-\mu_0)^2)$$

$$= \underset{\mu}{\operatorname{argmax}} \log(\prod_{\mathrm{i=1}}^{\mathrm{n}} \exp(-\frac{1}{2\sigma^2}(\mathrm{xi}-\mu)^2)\exp(-\frac{1}{2\sigma^2}(\mu-\mu_0)^2))$$

$$= \underset{\mu}{\operatorname{argmax}} -\sum_{i=1}^{n}(\mathrm{xi}-\mu)^2-(\mu-\mu_0)^2$$

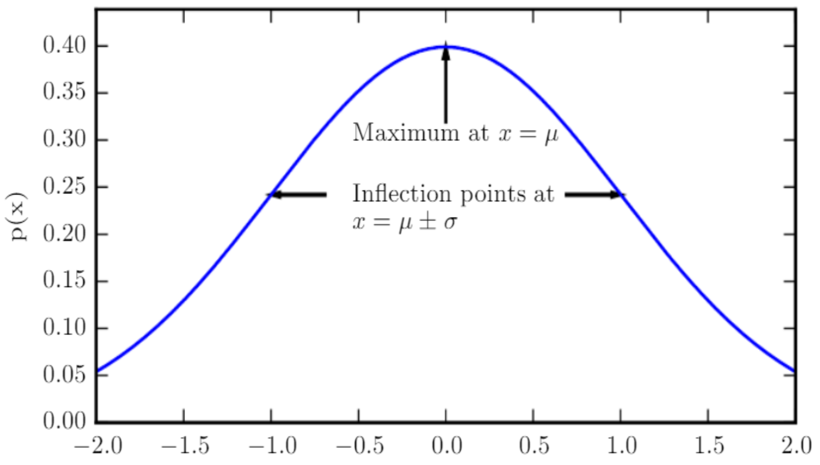
$$= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n}(\mathrm{xi}-\mu)^2+(\mu-\mu_0)^2$$



Exercises: MAP for Gaussian Distribution

Addressing this optimization problem:

$$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathrm{xi} - \mu)^2 + (\mu - \mu_0)^2$$



Derivation

$$\frac{d\left(\sum_{i=1}^{n}(\mathrm{xi}-\mu)^2+(\mu-\mu_0)^2\right)}{d\mu} = 2\sum_{i=1}^{n}(\mu - \mathrm{xi}) + 2(\mu - \mu_0) = 0$$

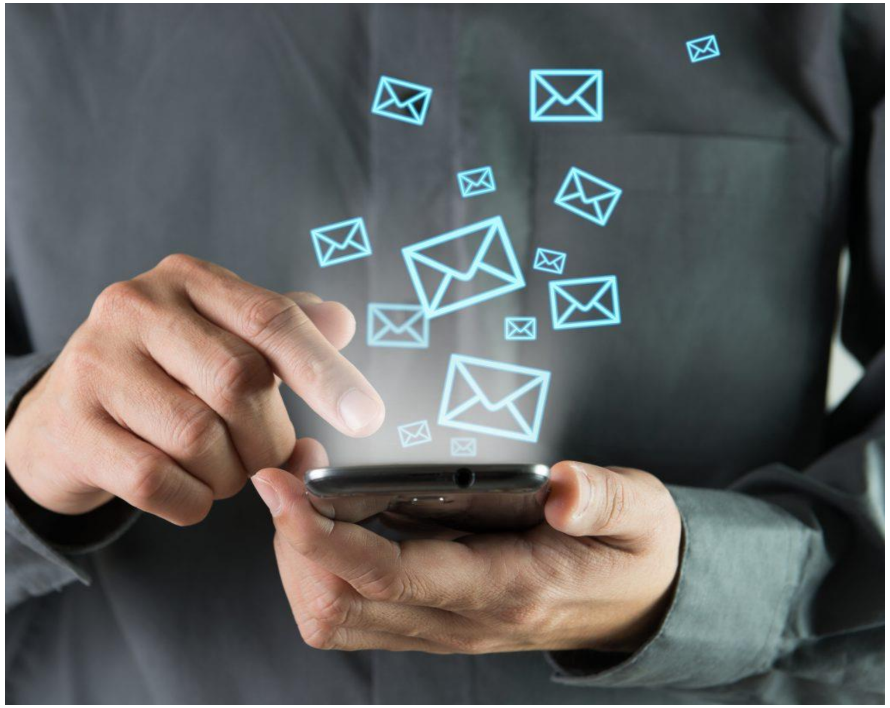
- The solution is

$$\mu = \frac{\sum_{i=1}^{n} \mathrm{xi} + \mu_0}{n+1}$$

Exercises: Maximum A Posteriori (MAP) Estimation

- Imagine you sent a message α to your friend that is either 1 or 0 with probability p and 1−p, respectively.

- Unfortunately that message gets corrupted by Gaussian noise N with zero mean and unit variance. Then what your friend would receive is a message X given by $X=\alpha+N$.

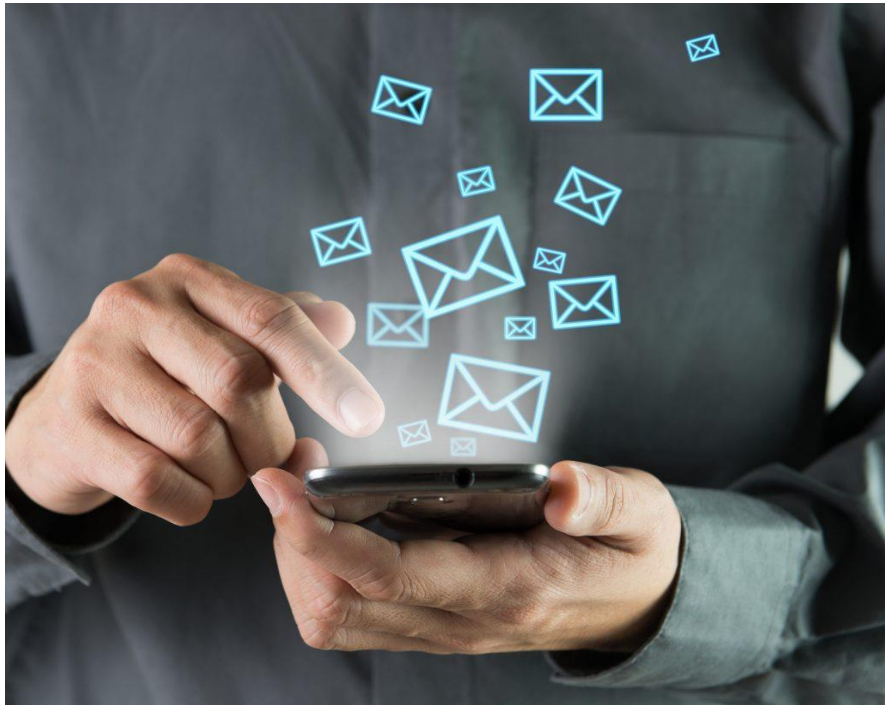


Exercises: Maximum A Posteriori (MAP) Estimation

- Given that what your friend observed was that X takes a particular value x, that is X=x, he wants to know which was, probably, the value of α that you sent to him.

By MAP, we sholud compute

$$\underset{\alpha}{\operatorname{argmax}}\ p_X(x|\alpha)\, P(\alpha)$$



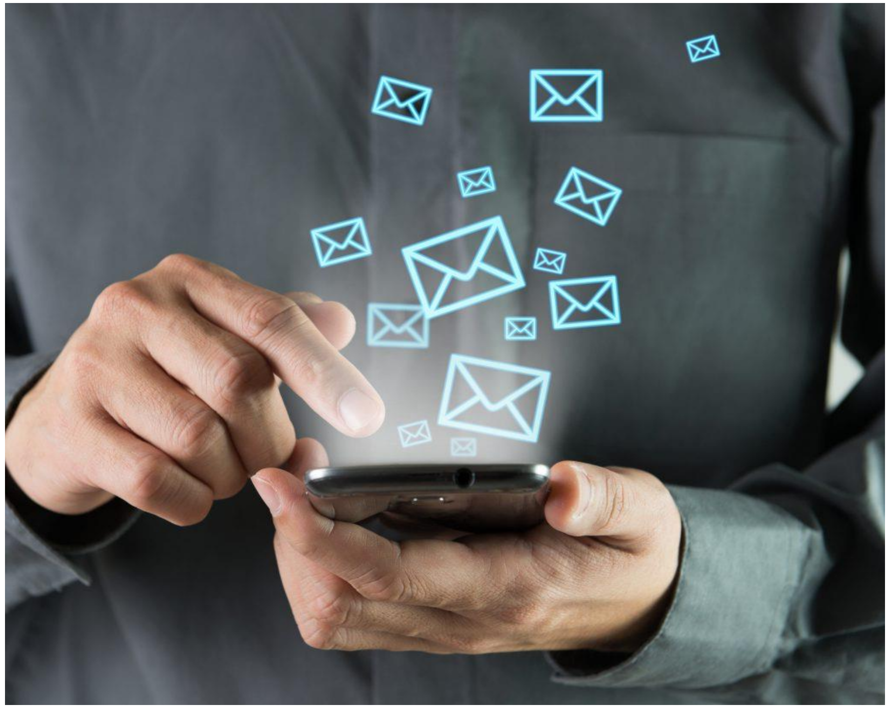
Exercises: Maximum A Posteriori (MAP) Estimation

- p_{X}(x|a=1)

= p_{N}(x-1)= $\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(\mathrm{x}-1)^2)$

- p_{X}(x|a=0)

= p_{N}(x)= $\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(\mathrm{x})^2)$

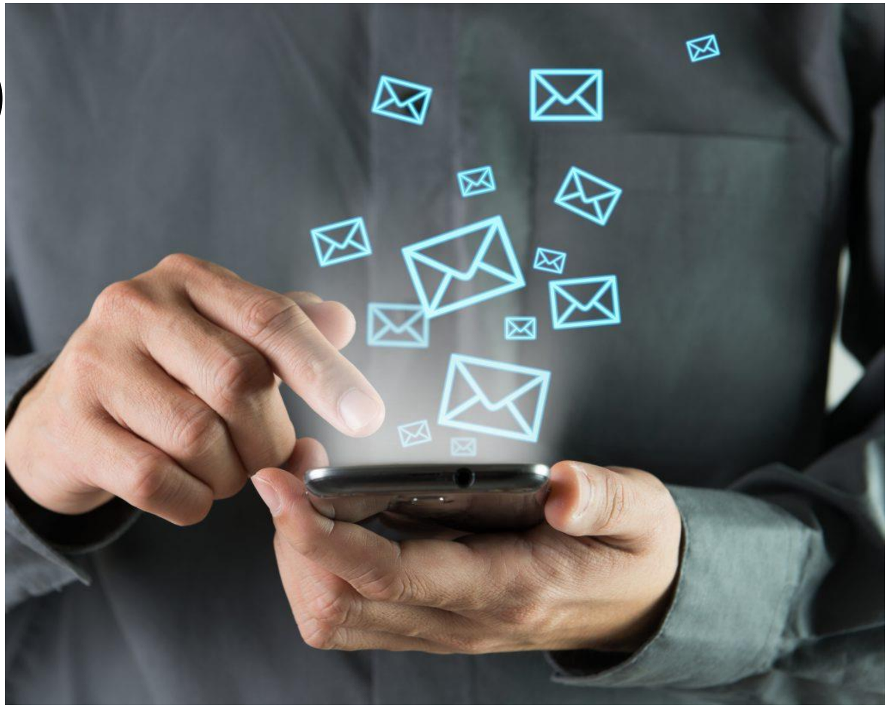


Exercises: Maximum A Posteriori (MAP) Estimation

Therefore,

$$p_X(x|a=1)\ P(a=1)=p\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(x-1)^2)$$

$$p_X(x|a=0)\ P(a=0)=(1-p)\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(x)^2)$$



Then, $p_X(x|a=1)\ P(a=1) > p_X(x|a=0)\ P(a=0)$

if and only if

$$x>0.5+\log(1-p)-\log(p)$$

Exercises: Maximum A Posteriori (MAP) Estimation

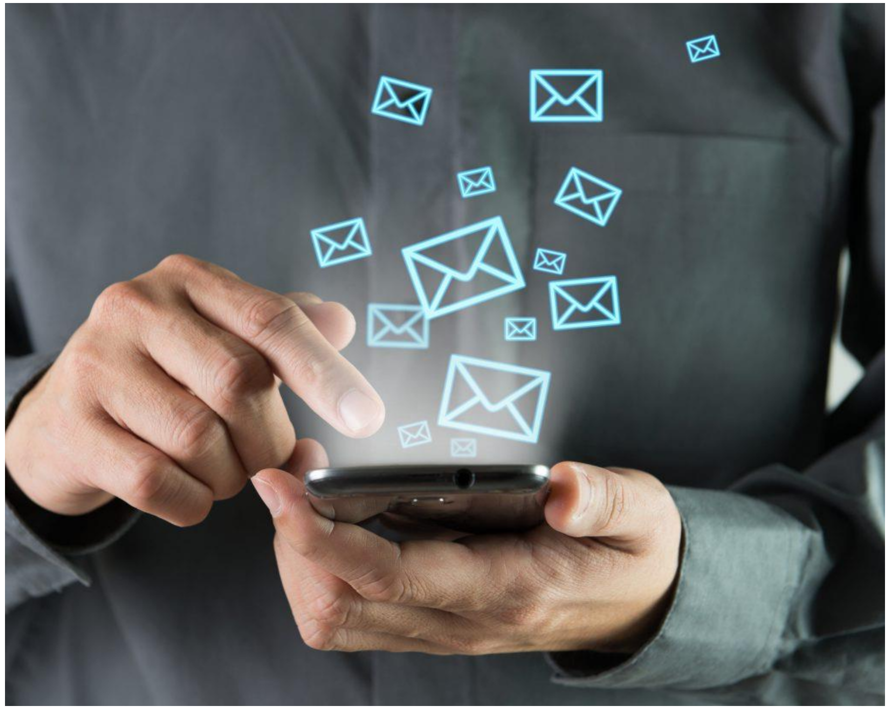
Therefore,

$$\underset{\alpha}{\text{argmax}}\; p_X(x|\alpha)\, P(\alpha) = 1,$$

if x>0.5+log(1−p)-log(p)

$$\underset{\alpha}{\text{argmax}}\; p_X(x|\alpha)\, P(\alpha) = 0,$$

if x<0.5+log(1−p)-log(p)



Exercises: Maximum A Posteriori (MAP) Estimation

Therefore,

$$p_X(\text{x}|\alpha\text{=1})\ \text{P}(\alpha\text{=1})\text{=p}\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(\text{x}-1)^2)$$

$$p_X(\text{x}|\alpha\text{=0})\ \text{P}(\alpha\text{=0})\text{=(1-p)}\sqrt{\frac{1}{2\pi}}\exp(-\frac{1}{2}(\text{x})^2)$$

Because

$$\underset{\alpha}{\text{argmax}}\ p_X(\text{x}|\alpha)\ \text{P}(\alpha) = \underset{\alpha}{\text{argmax}}\ \log p_X(\text{x}|\alpha) + \log \text{P}(\alpha)$$

Thank You!