UVA CS 4501: Machine Learning

Lecture 12: MLE

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

Sample space and Events

- O : Sample Space,
  - result of an experiment / set of all outcomes
  - If you toss a coin twice O = {HH,HT,TH,TT}
- Event: a subset of O
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and O

From Events to Random Variable

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - O = all possible students (sample space)
  - What are events (subset of sample space)
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking_Yes = … who works hard
  - Very cumbersome

  - Need "functions" that maps from O to an attribute space T.
  - P(H = YES) = P({student ϵ O : H(student) = YES})

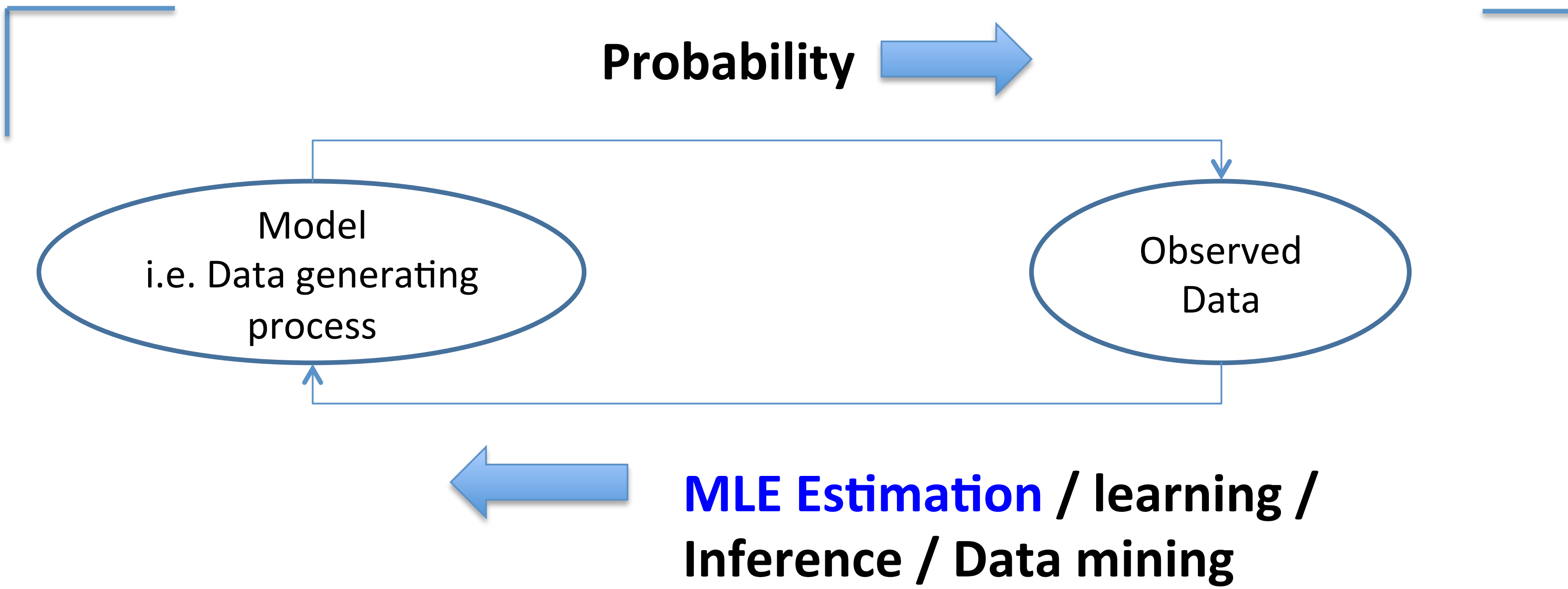
If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule
- 2. Marginal probability
  - Use the total law of probability
- 3. Conditional probability
  - Use the Bayes Rule

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

The Big Picture



Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Maximum Likelihood Estimation

A general Statement

Consider a sample set $T=(X_1 \dots X_n)$ which is drawn from a probability distribution P(X|\theta) where \theta are parameters.

If the Xs are independent with probability density function $P(X_i|\theta)$, the joint probability of the whole set is

$$P(X_1 \dots X_n|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$$

this may be maximised with respect to \theta to give the maximum likelihood estimates.

The idea is to

✓ assume a particular model with unknown parameters, θ

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(X_i \mid \theta)$

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(X_i \mid \theta)$
- ✓ We have observed a set of outcomes in the real world. $x_1, x_2, \cdots, x_n$

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(X_i \mid \theta)$
- ✓ We have observed a set of outcomes in the real world.
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(X_i / \theta)$
- ✓ We have observed a set of outcomes in the real world.
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(X_1 \dots X_n / \theta)$$

This is maximum likelihood. In most cases it is both consistent and efficient.

$$\log(L(\theta)) = \sum_{i=1}^{n} \log(P(X_i / \theta))$$

It is often convenient to work with the Log of the likelihood function.

The idea is to

- ✓ assume a particular model with unknown parameters, θ
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters. $P(X_i / \theta)$
- ✓ We have observed a set of outcomes in the real world.
- ✓ It is then possible to choose a set of parameters which are most likely to have produced the observed results.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(X_1 \dots X_n / \theta)$$

Likelihood

This is maximum likelihood. In most cases it is both consistent and efficient.

$$\log(L(\theta)) = \sum_{i=1}^{n} \log(P(X_i / \theta))$$

Log-Likelihood

It is often convenient to work with the Log of the likelihood function.

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of heads X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

e.g. Coin Flips cont.

$\{H, T\}$

- You flip a coin
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *a* coin for *k* times
  - How many heads would you expect
  - Number of heads X is a discrete random variable
  - Binomial distribution with parameters *k* and *p*

Review: Bernoulli Distribution e.g. Coin Flips

- You flip *n* coins
  - How many heads would you expect
  - Head with probability *p*
  - Number of heads X out of n trial
  - Each Trial following Bernoulli distribution with parameters *p*

N trails, e.g. $\{ \underset{X_1}{H}\ \underset{X_2}{H}\ \underset{X_3}{T}\ \underset{X_4}{H}\ H\ T\ H\ T \cdots \underset{X_n}{H} \}$

Calculating Likelihood

Given : $\{x_1, x_2, \cdots\cdots, x_n\}$

$\Downarrow$

$\{H, H, T, \cdots H\}$

$\Downarrow$ reformulate

$\{1, 1, 0, \cdots\cdots, 1\}$

$p(x_i|\theta) = p^{x_i}(1-p)^{1-x_i}$ (Here $x_i \in \{0, 1\}$)

n

Defining Likelihood for Bernoulli

- Likelihood = p(data | parameter)

➔ e.g., for n independent tosses of coins, with unknown p

Observed data ➔ x heads-up from n trials

function of x_i

PMF:

$$f(x_i \mid p) = p^{x_i}(1-p)^{1-x_i}$$

$$x = \sum_{i=1}^{n} x_i$$

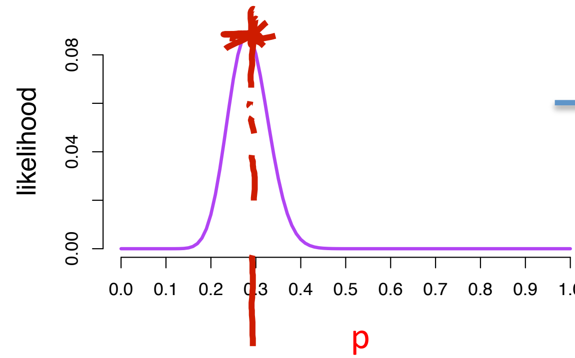
LIKELIHOOD:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x(1-p)^{n-x}$$

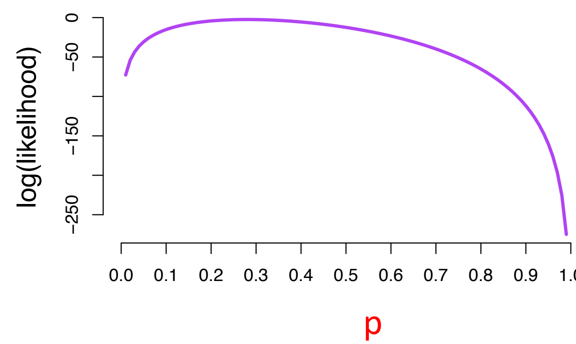
function of p

Deriving the Maximum Likelihood Estimate for Bernoulli

$$\overbrace{L(p)}^{\text{maximize}} = p^x (1-p)^{n-x}$$

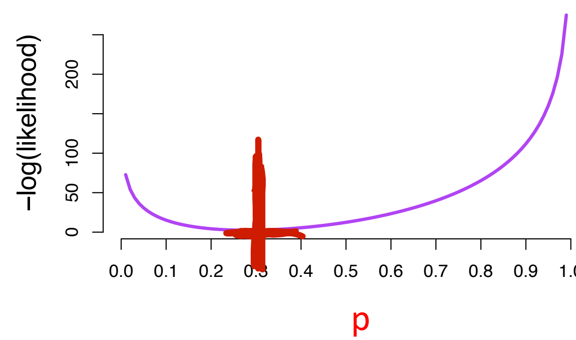


$$\overbrace{\log(L(p)}^{\text{maximize}} = \log\left[p^x (1-p)^{n-x} \right]$$



Minimize the negative log-likelihood

$$-l(p) = -\log\left[p^x (1-p)^{n-x} \right]$$



Deriving the Maximum Likelihood Estimate for Bernoulli

Minimize the negative log-likelihood

$$\underset{p}{\operatorname{argmin}}\left\{-l(p)\right\} = -\log(L(p)) = -\log\left[p^x(1-p)^{n-x}\right]$$

$$= -\log(p^x) - \log((1-p)^{n-x})$$

$$= -x\log(p) - (n-x)\log(1-p)$$

Deriving the Maximum Likelihood Estimate for Bernoulli

$$\arg_p\{-l(p)\} = \arg\min_p\{-x\log(p)-(n-x)\log(1-p)\}$$

$$\frac{dl(p)}{dp} = -\frac{x}{p} - \frac{-(n-x)}{1-p} = 0$$

$$0 = -\frac{x}{\hat{p}} + \frac{n-x}{1-\hat{p}}$$

$$0 = \frac{-x(1-\hat{p}) + \hat{p}(n-x)}{\hat{p}(1-\hat{p})}$$

$$0 = -x + \hat{p}x + \hat{p}n - \hat{p}x$$

$$0 = -x + \hat{p}n$$

Minimize the negative log-likelihood

➔ MLE parameter estimation

$$\hat{p} = \frac{x}{n}$$

i.e. Relative frequency of a binary event

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Review: Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
  - For discrete RV: Probability mass function (pmf): P(X = x_i)

- A pdf (prob. Density func.) is any function $f(x)$ that describes the probability density in terms of the input variable x.

Review: Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$ $\longrightarrow$ $\sum_{i=1}^{k_i} P(X = x_i) = 1$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 5 and 6 is

$$P(5 \leq X \leq 6) = \int_5^6 f(x)dx$$

Review: Mean and Variance of RV

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$

    $$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$

    $$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Adapt From Carols' prob tutorial

Review: Mean and Variance of RV

- Variance: $Var(X) = E((X-\mu)^2)$ $\sigma_X = \sqrt{V(X)}$
  - Discrete RVs: $V(X) = \sum_{v_i} (v_i - \mu)^2 P(X = v_i)$
  - Continuous RVs: $V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$

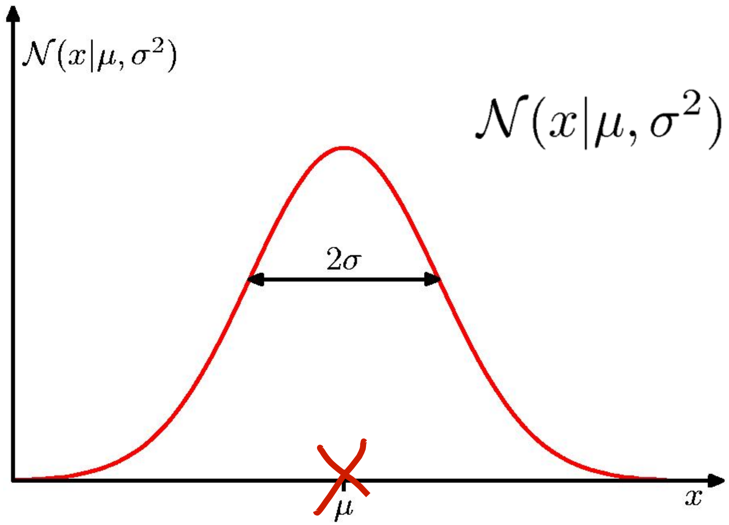
Correlation $\rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

- Covariance:

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x \mu_y$$

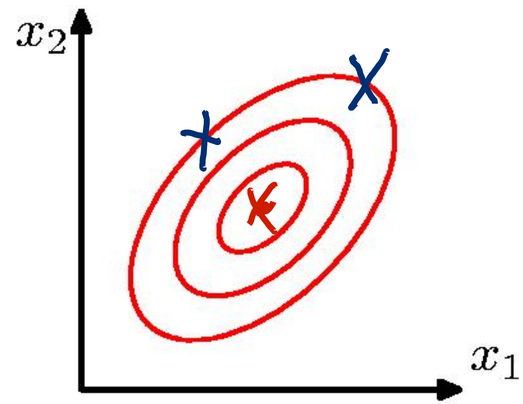
Adapt From Carols' prob tutorial

Single-variate Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$X \sim N(\mu, \sigma^2)$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Mean

Covariance Matrix

Courtesy: http://research.microsoft.com/~cmbishop/PRML/index.htm

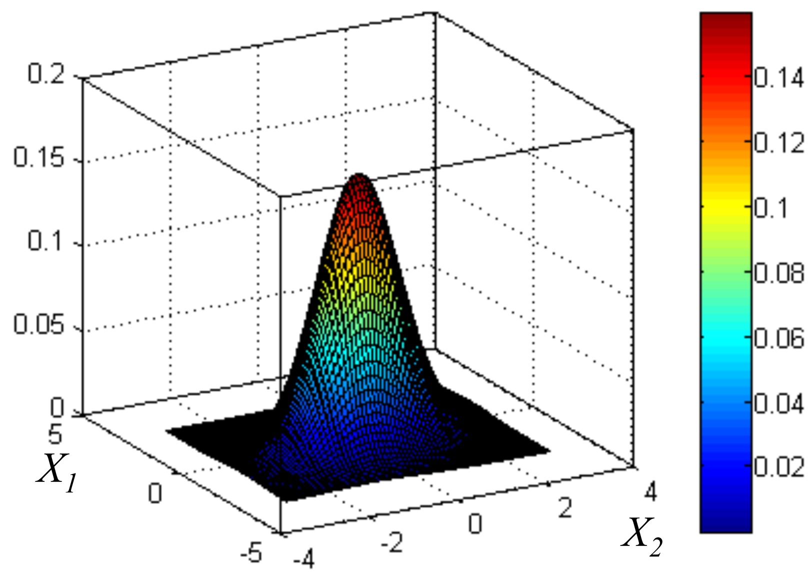
Multivariate Normal (Gaussian) PDFs

The only widely used continuous joint PDF is the multivariate normal (or Gaussian):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Where |*| represents determinant

Bivariate normal PDF:



- Mean of normal PDF is at peak value. Contours of equal PDF form ellipses.

- The covariance matrix captures linear dependencies among the variables

Example: the Bivariate Normal distribution

$$f\left(x_1, x_2\right) = \frac{1}{\left(2\pi\right)\left|\Sigma\right|^{1/2}} \mathrm{e}^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

with $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}_{2\times 1}$ and

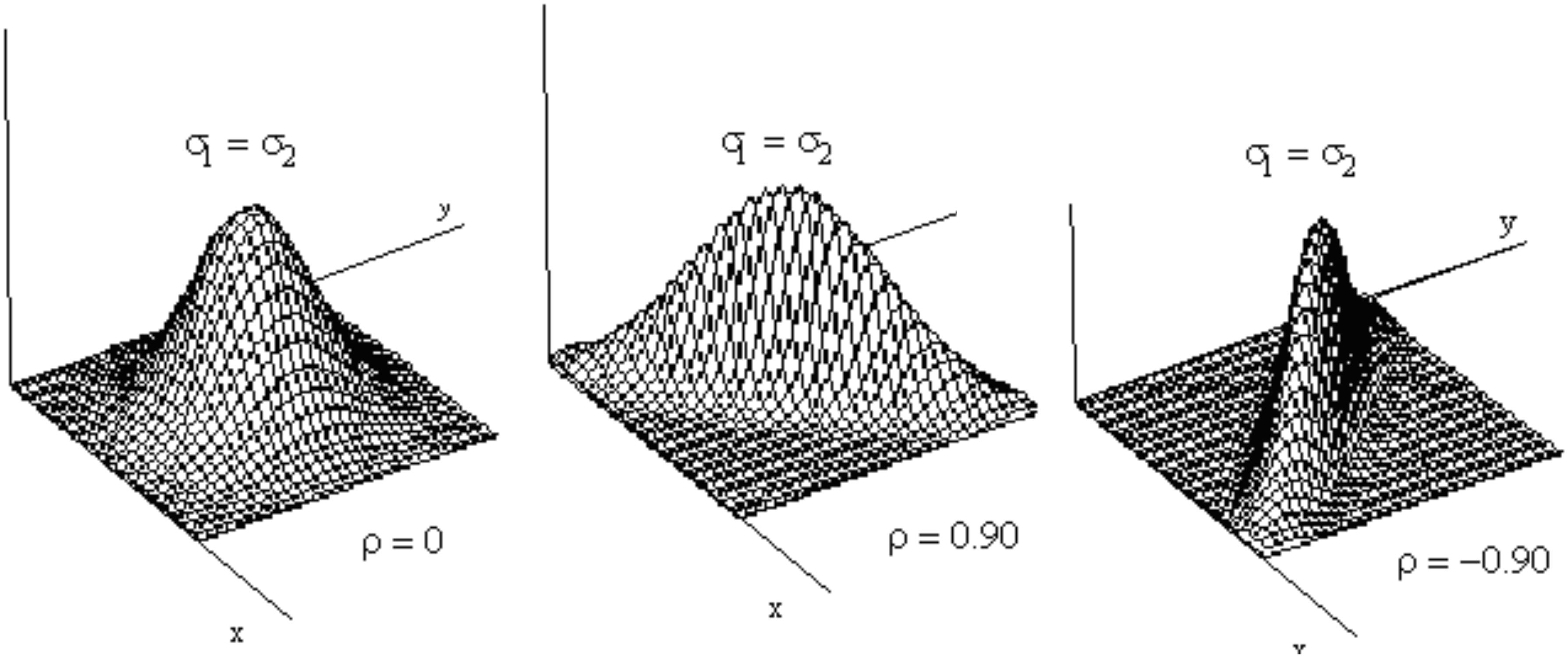
V(X1) Cov(X1, X2)

$$\underset{2\times 2}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \underline{\rho\sigma_1\sigma_2} \\ \underline{\rho\sigma_1\sigma_2} & \sigma_2^2 \end{bmatrix}_{2\times 2}$$

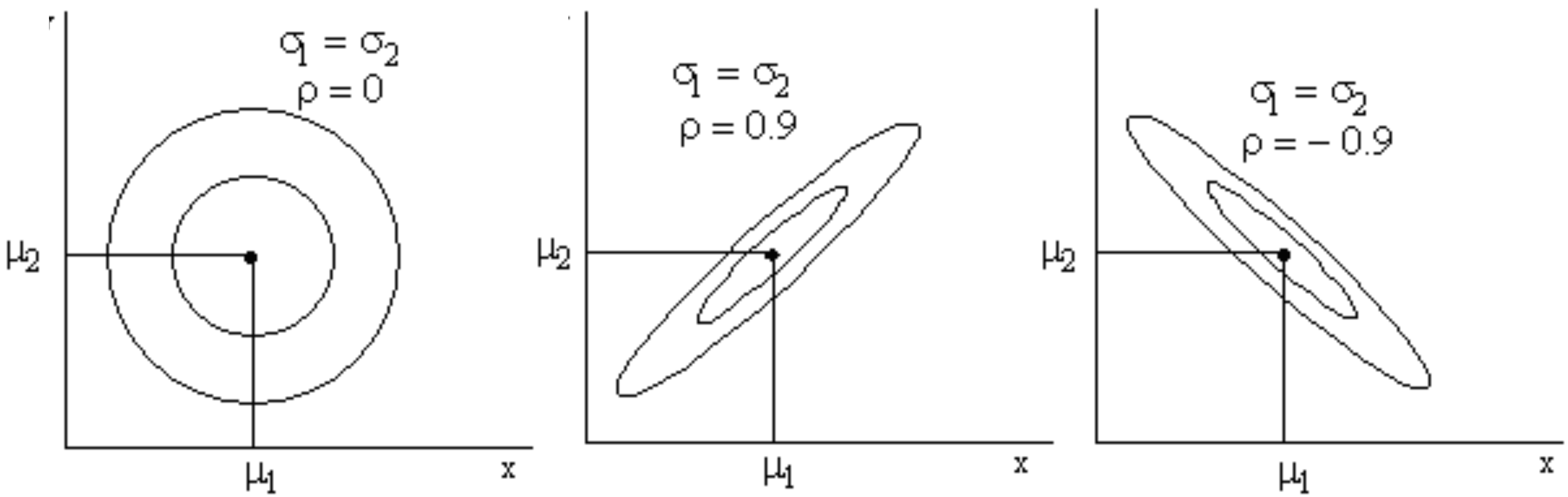
V(X2)

$$\left|\Sigma\right| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_1^2\sigma_2^2\left(1-\rho^2\right)$$

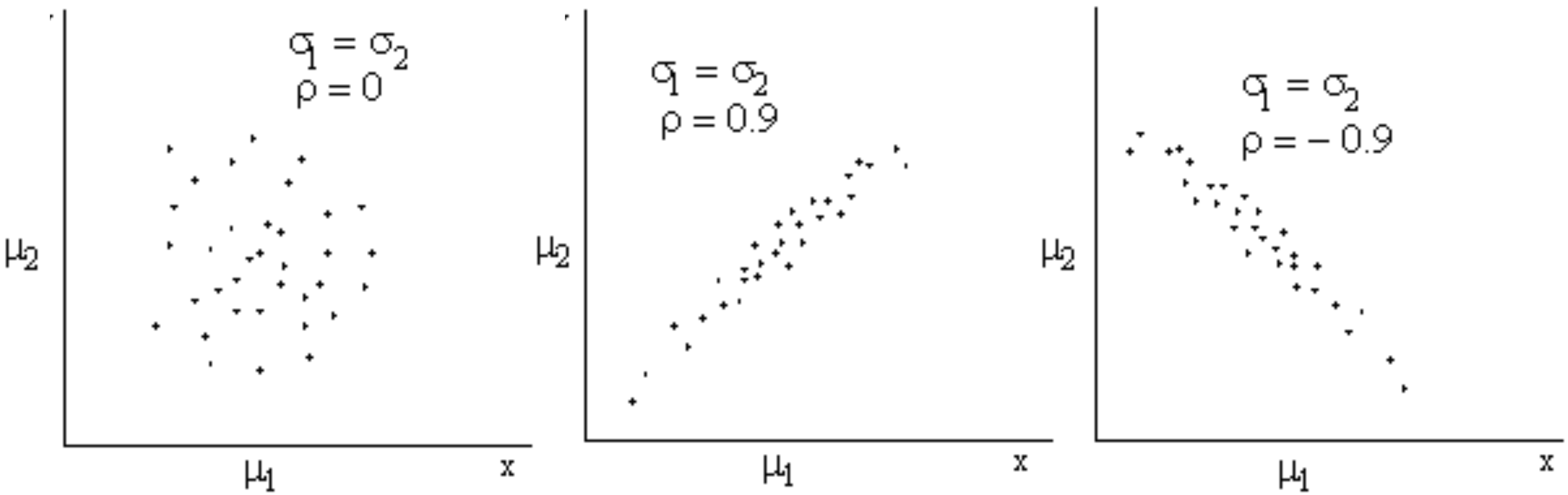
Surface Plots of the bivariate Normal distribution



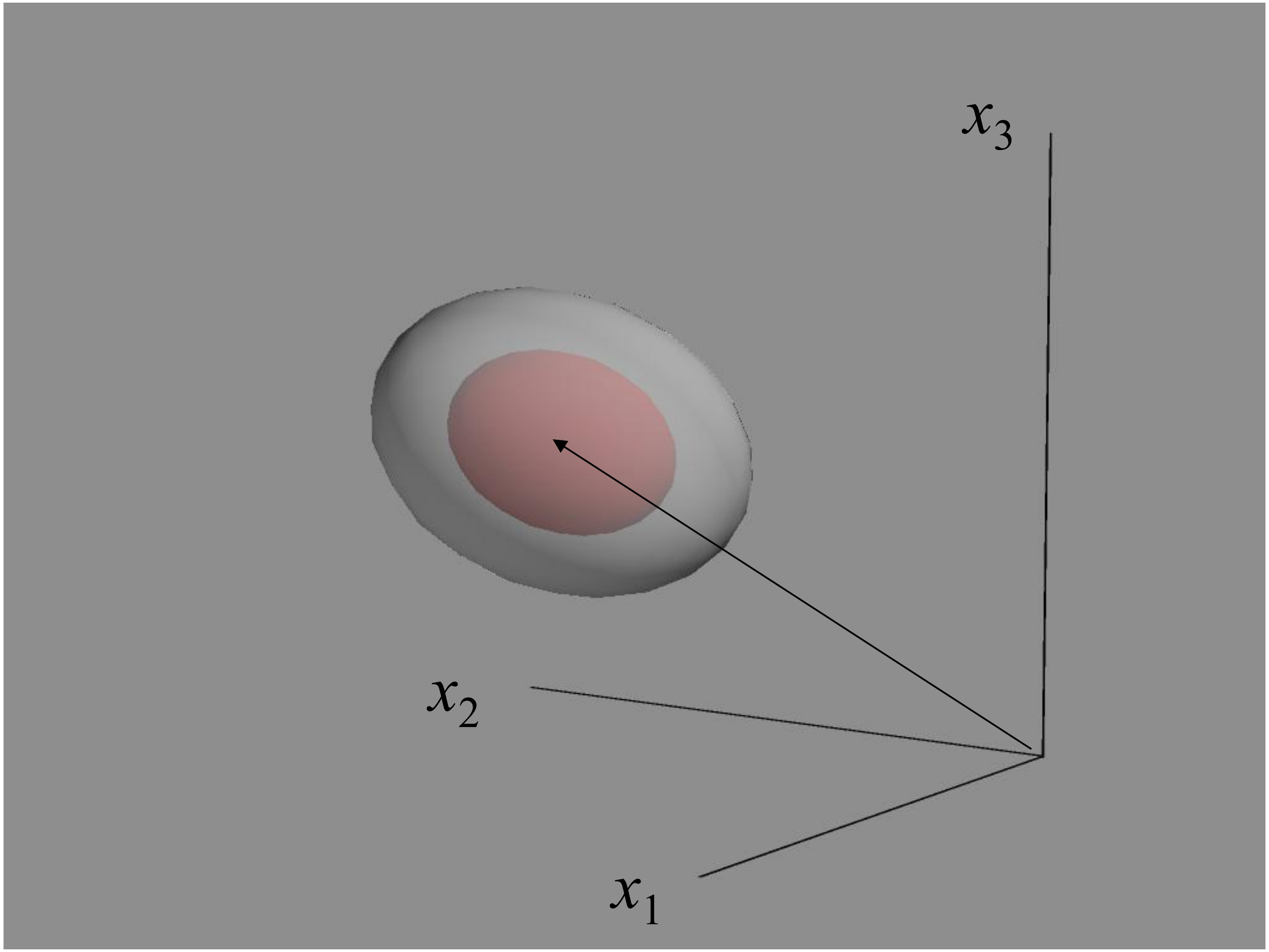
Contour Plots of the bivariate Normal distribution



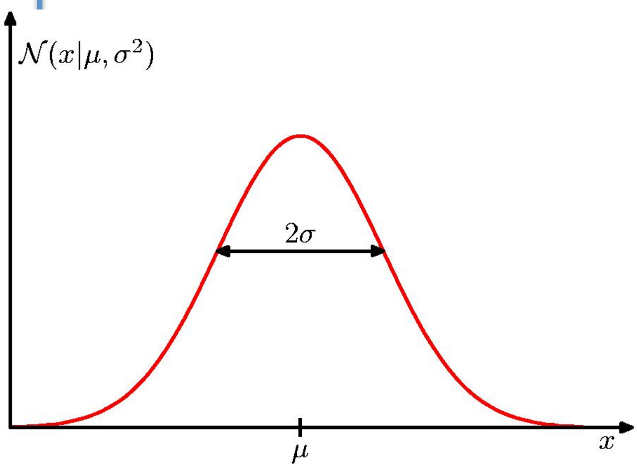
Scatter Plots of data from the bivariate Normal distribution



Trivariate Normal distribution

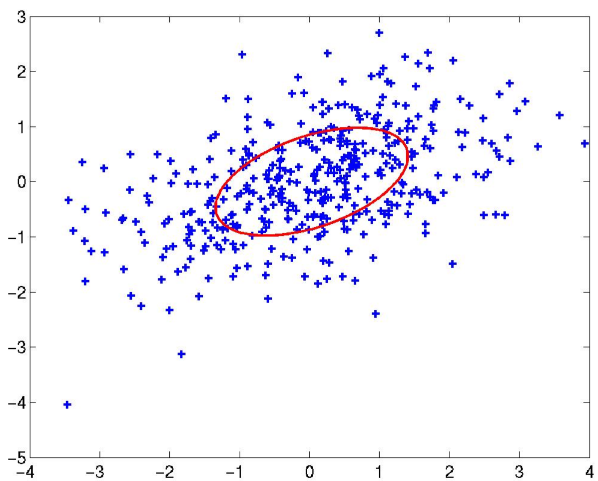


How to Estimate Gaussian: MLE



- In the 1D Gaussian case, we simply set the mean and the variance to the sample mean and the sample variance:

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mu})^2$$



The p-multivariate Normal distribution

$$< X_1, X_2 \cdots, X_p > \sim N\left(\vec{\mu}, \Sigma\right)$$

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}_{p \times 1}$$

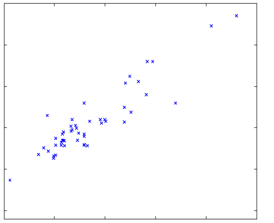
$$\mu_i = \frac{1}{n} \sum_{j=1}^{N} \underbrace{X_j}_{\substack{j\text{th sample} \\ \in \{1,2,\cdots,N\}}}^{\overbrace{(i)}^{\substack{i\text{-th feature} \\ \in \{1,2,\cdots,p\}}}}$$

$$\underset{p \times p}{\Sigma} = \begin{array}{c} \begin{array}{cc} & i \end{array} \\ \begin{bmatrix} Var(X_1) & & \\ & \ddots & Cov(X_i, X_j) \\ & & \ddots \\ & & Var(X_p) \end{bmatrix} \; j \end{array}$$

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

DETOUR: Probabilistic Interpretation of Linear Regression

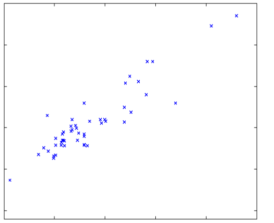


- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

where ε is an error term of unmodeled effects or random noise

DETOUR: Probabilistic Interpretation of Linear Regression



- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

RV $\varepsilon \sim N(0, \sigma^2)$

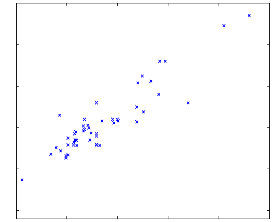
where ε is an error term of unmodeled effects or random noise

- Now assume that ε follows a Gaussian $N(0,\sigma^2)$, then we have:

$$p(y_i \mid x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right)$$

RV $y \mid x; \theta \sim N(\theta^T x, \sigma)$

DETOUR: Probabilistic Interpretation of Linear Regression



- By IID (independent and identically distributed) assumption, we have data likelihood

$$L(\theta) = \prod_{i=1}^{n} p(y_i \mid x_i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right)$$

$$l(\theta) = \log(L(\theta)) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta^T \mathbf{x}_i)^2$$

$$L(\theta)=\prod_{i=1}^{n}p(y_i|x_i;\theta)=\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\theta^T\mathbf{x}_i)^2}{2\sigma^2}\right)$$

We can learn \theta by maximizing the probability / likelihood of generating the observed samples:

$$p\{(\vec{x}_1,y_1)\wedge(\vec{x}_2,y_2)\wedge\cdots(\vec{x}_N,y_N)\}$$

IID

$$=\prod_{i=1}^{N}p(y_i,\vec{x}_i)=\prod_{i=1}^{N}p(y_i|\vec{x}_i;\theta)p(\vec{X}_i)$$

$$\theta^*=\arg\max_{\theta}\prod_{i=1}^{N}p(y_i|\vec{X}_i;\theta)$$

Thus under independence Gaussian residual assumption, residual square error is equivalent to MLE of ϑ !

$y|x;\theta \sim N(\theta^T x, \sigma)$

$\Downarrow$

Two unknown parameters : $\{\theta, \sigma\}$

$$l(\theta) = \log(L(\theta)) = n\log\frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T \mathbf{x}_i)^2$$

$\Downarrow$

$\operatorname{argmax} l(\theta) \Rightarrow \operatorname{argmin} J(\theta)$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i{}^T\theta - y_i)^2$$

$\boxed{y_i \sim N(exp(wx_i), 1)}$

(b) (6 points) (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\Sigma_i x_i exp(wx_i) = \Sigma_i x_i y_i exp(wx_i)$

B. $\Sigma_i x_i exp(2wx_i) = \Sigma_i x_i y_i exp(wx_i)$

C. $\Sigma_i x_i^2 exp(wx_i) = \Sigma_i x_i y_i exp(wx_i)$

D. $\Sigma_i x_i^2 exp(wx_i) = \Sigma_i x_i y_i exp(wx_i/2)$

E. $\Sigma_i exp(wx_i) = \Sigma_i y_i exp(wx_i)$

$y_i \sim N(exp(wx_i), 1)$

Answer: B (this is an extra credit question.)

$L(\theta)$

$\downarrow$

$LL(\theta)$

$\downarrow$

$\frac{\partial LL(\theta)}{\partial \theta} = 0 \Rightarrow (B)$

Today

- ❑ Basic MLE
- ❑ MLE for Discrete RV
- ❑ MLE for Continuous RV (Gaussian)
- ❑ MLE connects to Normal Equation of LR
- ❑ More about Mean and Variance

Mean and Variance

- Correlation:

$$\rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y$$

$$-1 \le \rho(X,Y) \le 1$$

Properties

- Mean
  - $E(X+Y)=E(X)+E(Y)$
  - $E(aX)=aE(X)$
  - If X and Y are independent, $E(XY)=E(X)\cdot E(Y)$
- Variance
  - $V(aX+b)=a^2V(X)$
  - If X and Y are independent, $V(X+Y)=V(X)+V(Y)$

Some more properties

- The conditional expectation of Y given X when the value of X = x is:

$$E(Y \mid X = x) = \int y * p(y \mid x) dy$$

- The Law of Total Expectation or Law of Iterated Expectation:

$$E(Y) = E[E(Y \mid X)] = \int E(Y \mid X = x) p_X(x) dx$$

Some more properties

- The law of Total Variance:

$$Var(Y) = Var\left[E(Y \mid X)\right] + E\left[Var(Y \mid X)\right]$$

References

- Prof. Andrew Moore's review tutorial
- Prof. Nando de Freitas's review slides
- Prof. Carlos Guestrin recitation slides