

# UVA CS 4774: Machine Learning

## Lecture 10: Maximum Likelihood Estimation (MLE)

Dr. Yanjun Qi

University of Virginia

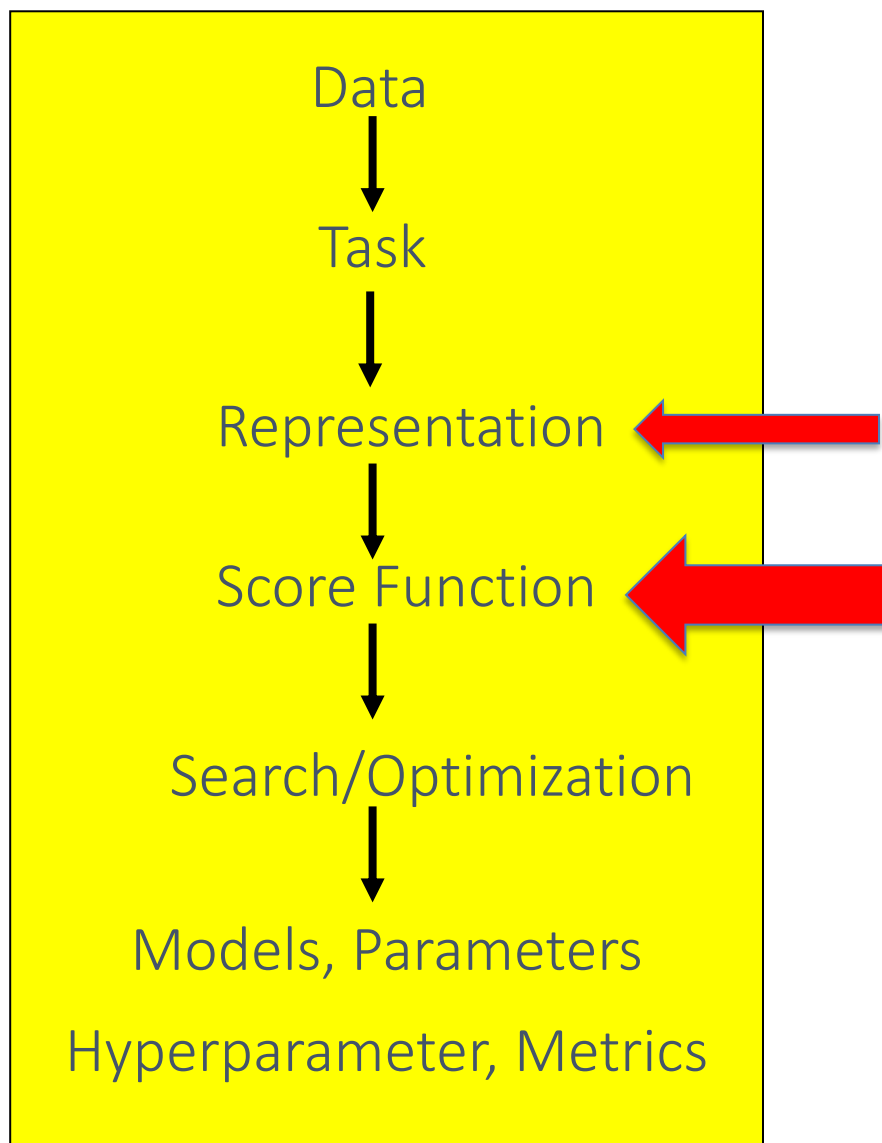
Department of Computer Science

# Machine Learning in a Nutshell

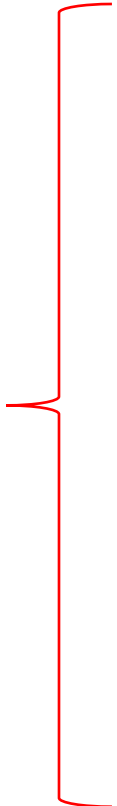
ML grew  
out of  
work in AI

Optimize a  
performance  
criterion  
using  
example data  
or past  
experience,

Aiming to  
generalize to  
unseen data



# 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

- $\Omega$  : **Sample Space**,
  - set of all outcomes
  - If you toss a coin **twice**  $\Omega = \{HH, HT, TH, TT\}$
- **Event**: a subset of  $\Omega$ 
  - First toss is head =  $\{HH, HT\}$
- $\mathcal{S}$ : **event space, a set of events**:
  - Contains the empty event and  $\Omega$

# 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
    - HardWorking\_Yes = ... who works hard
  - Very cumbersome
- Need “functions” that maps from  $O$  to an attribute space  $T$ .
- $P(H = \text{YES}) = P(\{\text{student} \in O : H(\text{student}) = \text{YES}\})$

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

- 1. Joint probability

- Use Chain Rule

$$P(X, Y)$$

- 2. Marginal probability

- Use the total law of probability

$$P(X)$$

- 3. Conditional probability

- Use the Bayes Rule

$$P(Y|X)$$

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

- 1. Joint probability

- Use Chain Rule

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

- 2. Marginal probability

- Use the total law of probability

$$P(B) = P(B, A) + P(B, \sim A)$$
$$\parallel$$
$$P(B, A \cup \sim A) //$$

- 3. Conditional probability

- Use the Bayes Rule

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

# Simplify Notation: To Calculate Conditional Probability

- Bayes Rule

$$P(x | y) = \frac{P(x)P(y | x)}{P(y)}$$

- You can condition on **more variables**

$$P(x | y, z) = \frac{P(x | z)P(y | x, z)}{P(y | z)}$$



# Examples

*Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set  $\{r, r, r, b\}$ . What is the probability of drawing 2 red balls in the first 2 tries?*

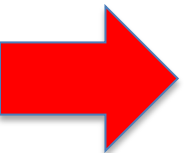
$$P(\underline{B_1 = r}, \underline{B_2 = r}) =$$

*What is the probability that the 2<sup>nd</sup> ball drawn from the set  $\{r, r, r, b\}$  will be red?*

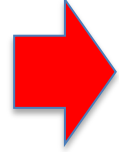
$$P(B_2 = r)$$

# Today : MLE

- 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



# Roadmap



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

$\{H H T T \dots H\}$

$$Z: \{H, T\} \rightarrow \{Z_1, Z_2, \dots, Z_n\}$$
$$P(Z): \{p, 1-p\}$$

$\downarrow p$

A general Statement

Consider a sample set  $T=(Z_1 \dots Z_n)$  which is drawn from a probability distribution  $P(Z|\theta)$  where  $\theta$  are parameters.

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

$$\theta^* = \arg \max_{\theta} \left\{ P(Z_1 \dots Z_n | \theta) \right\} = \prod_{i=1}^n P(Z_i | \theta)$$

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,  $\theta$

The idea is to

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

The idea is to

- ✓ assume a particular **model with unknown parameters**,  $\theta$
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters.  $P(Z_i|\theta)$
- ✓ We have observed **a set of outcomes** in the real world.

The idea is to

- ✓ assume a particular **model with unknown parameters**,  $\theta$
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters.  $P(Z_i|\theta)$
- ✓ We have observed **a set of outcomes** in the real world.  $z_1, z_2, \dots, z_n$
- ✓ 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**,  $\theta$
- ✓ we can then define the probability of observing a given event conditional on a particular set of parameters.  $P(Z_i|\theta)$
- ✓ 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.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(Z_1 \dots Z_n | \theta) = \prod_{i=1}^n P(Z_i | \theta)$$

This is maximum likelihood.


In most cases it is **both consistent and efficient**.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log(L(\theta)) = \sum_{i=1}^n \log(P(Z_i | \theta))$$


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


The idea is to

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(Z_1 \dots Z_n | \theta)$$


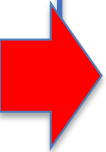
This is maximum likelihood.

In most cases this scorer is **both consistent and efficient**.

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


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

# Roadmap

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

# Review: Bernoulli Distribution

## e.g. Coin Flips

*Bernoulli( $p$ )*

- You flip a coin

- $Z$ : {Who is Up: Head or Tail} is a discrete Random Variable

$\{H, T\}$

- Head with probability  $p$
    - **Binary** random variable
    - **Bernoulli trial** with success probability  $p$

# Review: Bernoulli Distribution

## e.g. Coin Flips

- You flip  $n$  coins
  - Head with probability  $p$  (*UNKNOWN, Need to estimate from data*)
  - Number of heads  $X$  out of  $n$  trial
  - Each Trial following Bernoulli distribution with parameters  $p$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(Z_1 \dots Z_n | \theta)$$

## Review: Defining Likelihood for basic Bernoulli

Given:  $\{z_1, z_2, \dots, z_n\}$

$\Downarrow$   
 $\{H, H, T, \dots, H\}_n$

$\Downarrow$  reformulate

$\{1, 1, 0, \dots, 1\}_n$

$\theta = \{p\}$   
 $= \{p(\text{Head})\}$

$$p(z_i | \underline{\theta}) = p^{z_i} (1-p)^{1-z_i} \quad (\text{Here } z_i \in \{0, 1\})$$

$$p(z_i) = \begin{cases} p, & \text{if } z_i = H/1 \\ 1-p, & \text{if } z_i = T/0 \end{cases} \Rightarrow \arg \max_p \prod_{i=1}^n p^{z_i} (1-p)^{1-z_i}$$

# Deriving the Maximum Likelihood Estimate for Bernoulli

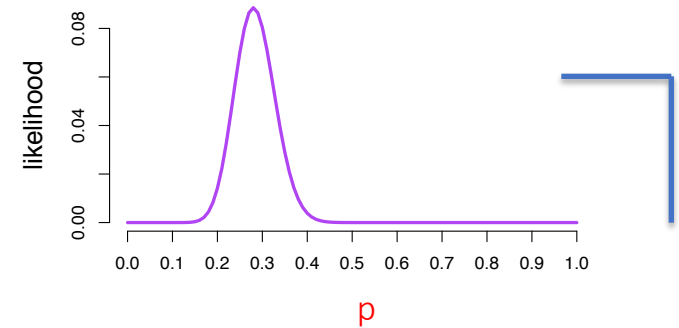
$$\begin{aligned}\log(L(p)) &= \log \left[ \prod_{i=1}^n p^{z_i} (1-p)^{1-z_i} \right] \\ &= \sum_{i=1}^n (z_i \log p + (1-z_i) \log(1-p)) \\ &= \log p \sum_{i=1}^n z_i + \log(1-p) \sum_{i=1}^n (1-z_i) \\ &= x \log p + (n-x) \log(1-p)\end{aligned}$$

Observed data  $\rightarrow x$   
heads-up from  $n$  trials

# Deriving the Maximum Likelihood Estimate for Bernoulli

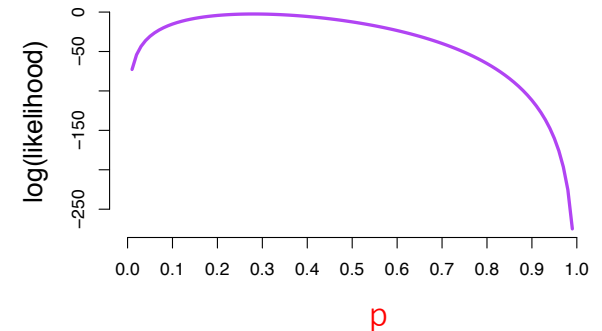
maximize

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



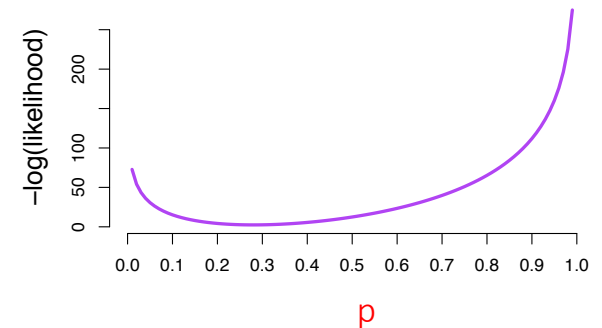
maximize

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



Minimize the negative log-likelihood

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





# Deriving the Maximum Likelihood Estimate for Bernoulli

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

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

$$0 = -x + pn$$

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

Minimize the negative log-likelihood

→ MLE parameter estimation

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

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

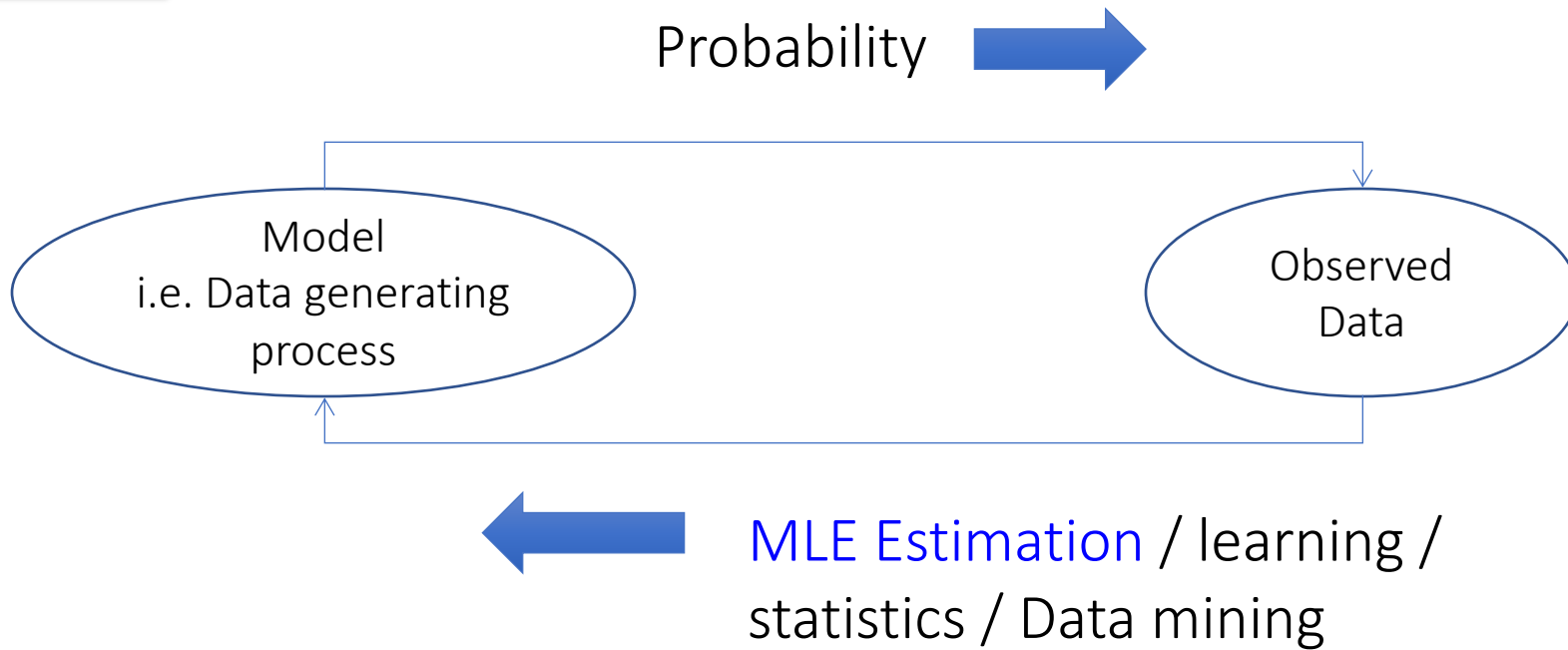
i.e. Relative frequency of a binary event

$$0 = -x + px + pn - px$$



# EXTRA

# The Big Picture



# Discrete Random Variables

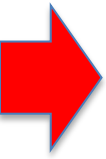
- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g.  $Z$  as the total number of heads you get if you flip 100 coins
- $Z$  is a RV with arity  $k$  if it can take on exactly one value out of *a set size  $k$* 
  - E.g. the possible values that  $Z$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

## e.g. Coin Flips cont.

- You flip a coin
  - $Z$ : {Who is Up: Head or Tail} is a discrete RV
  - 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  $Z$  is also a discrete random variable
  - **Binomial distribution** with parameters  $k$  and  $p$

$\{H, T\}$

# Roadmap - All the rest are EXTRA

- 
- ☐ 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):  $\mu = E(X)$

- Discrete RVs:

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

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

- Continuous RVs:

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

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

# 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:

$$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

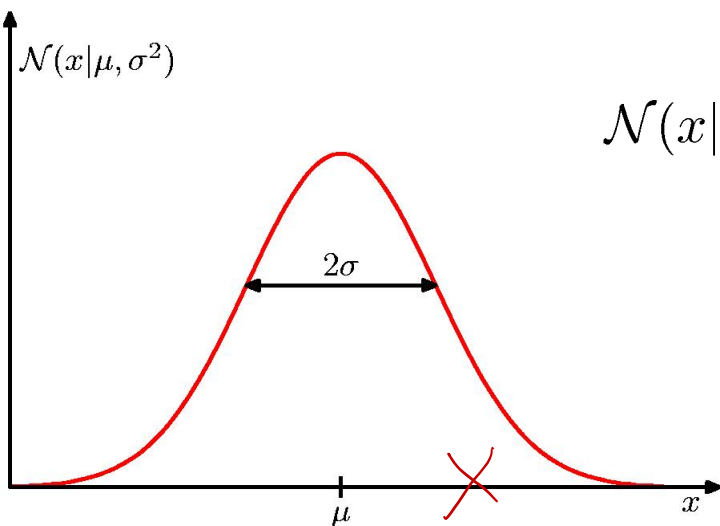
Correlation

$$\rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

- Covariance:

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

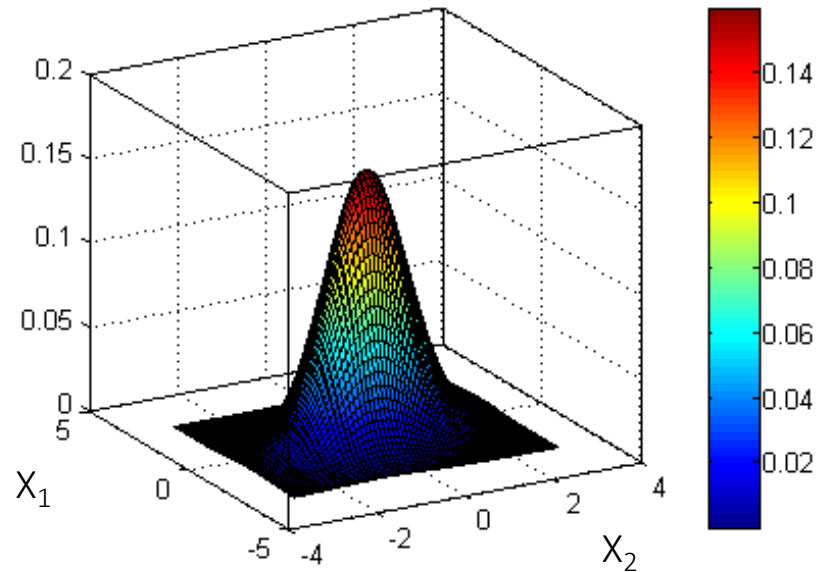
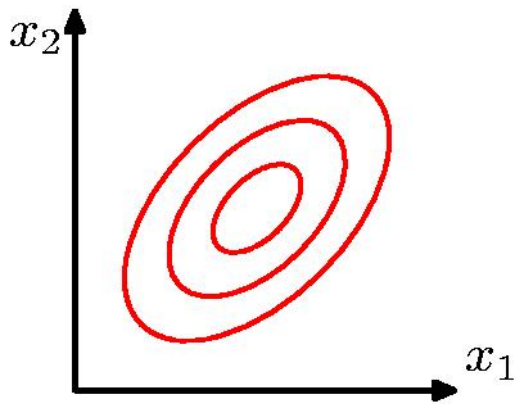
# Single-Variate Gaussian Distribution



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

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

# Bi-Variate Gaussian Distribution



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

# Multivariate Normal (Gaussian) PDFs

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

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

Where  $|\ast|$  represents **determinant**

Mean

Covariance Matrix

- 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

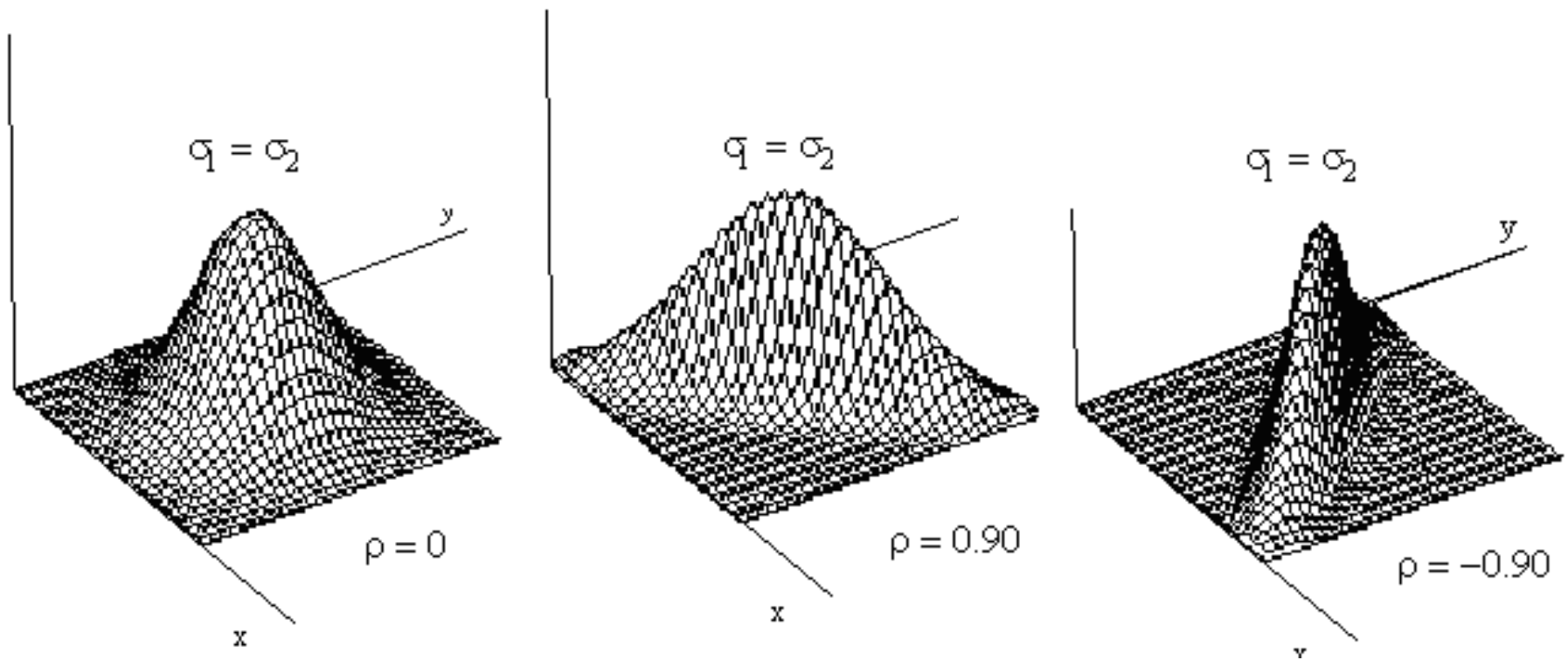
$$f(x_1, x_2) = \frac{1}{(2\pi)|\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})}$$

with  $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$  and

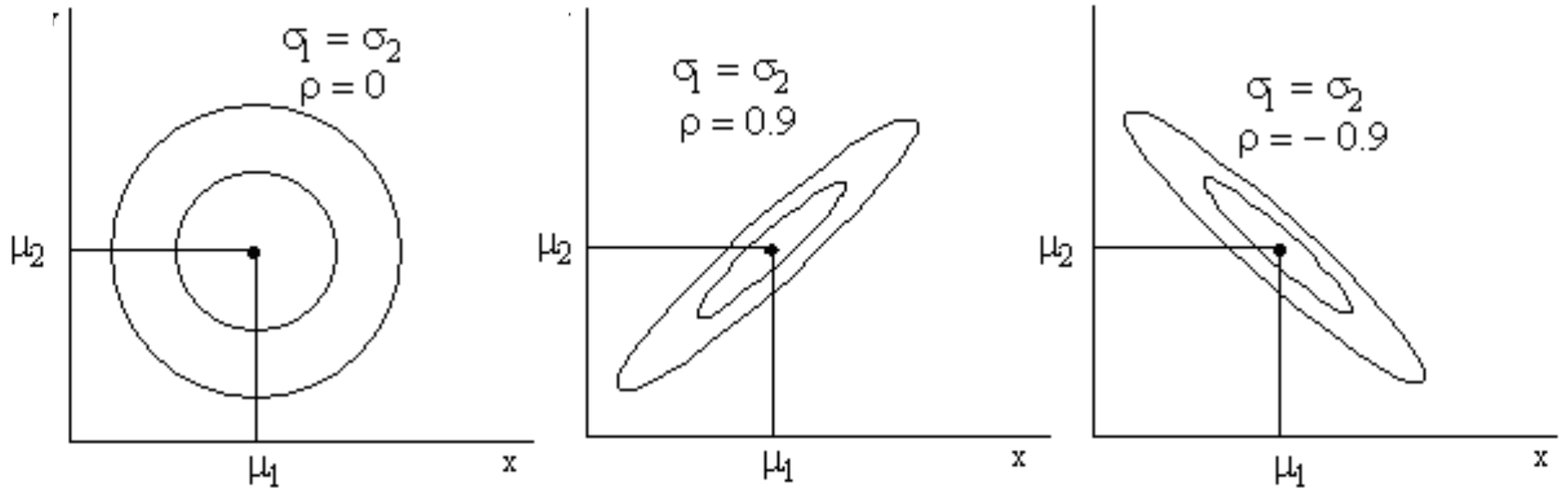
$$\Sigma_{2 \times 2} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \overset{\text{Var}(X_1)}{\sigma_1^2} & \overset{\text{Cov}(X_1, X_2)}{\rho \sigma_1 \sigma_2} \\ \rho \sigma_1 \sigma_2 & \overset{\text{Var}(X_2)}{\sigma_2^2} \end{bmatrix}_{2 \times 2}$$

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

# Surface Plots of the bivariate Normal distribution

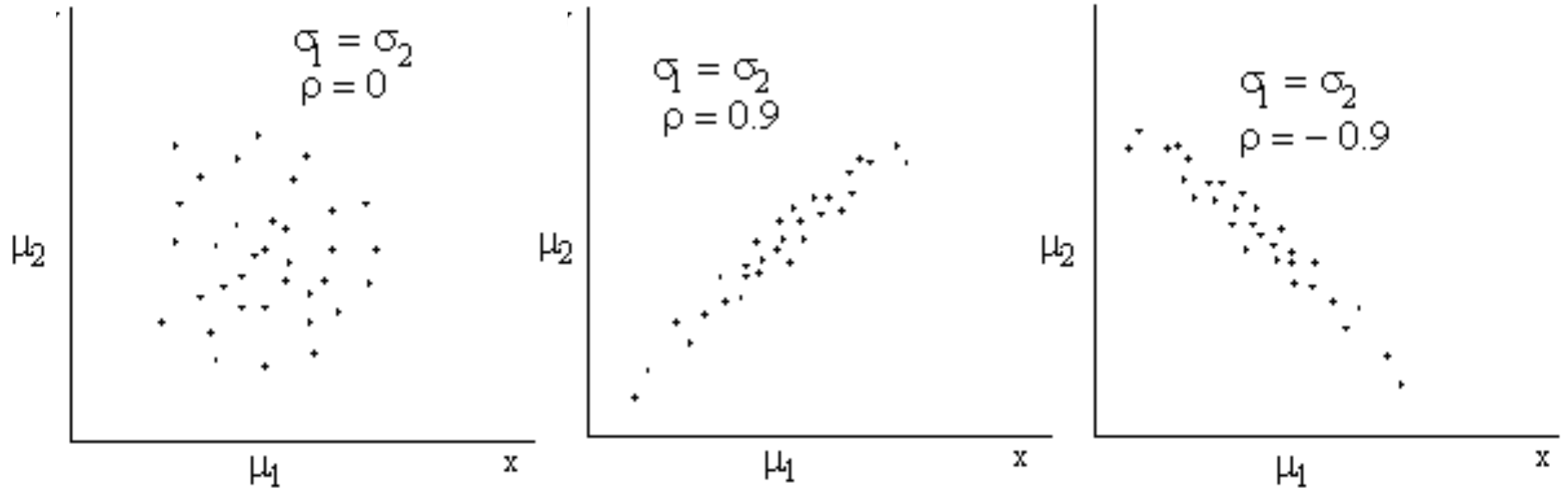


# Contour Plots of the bivariate Normal distribution

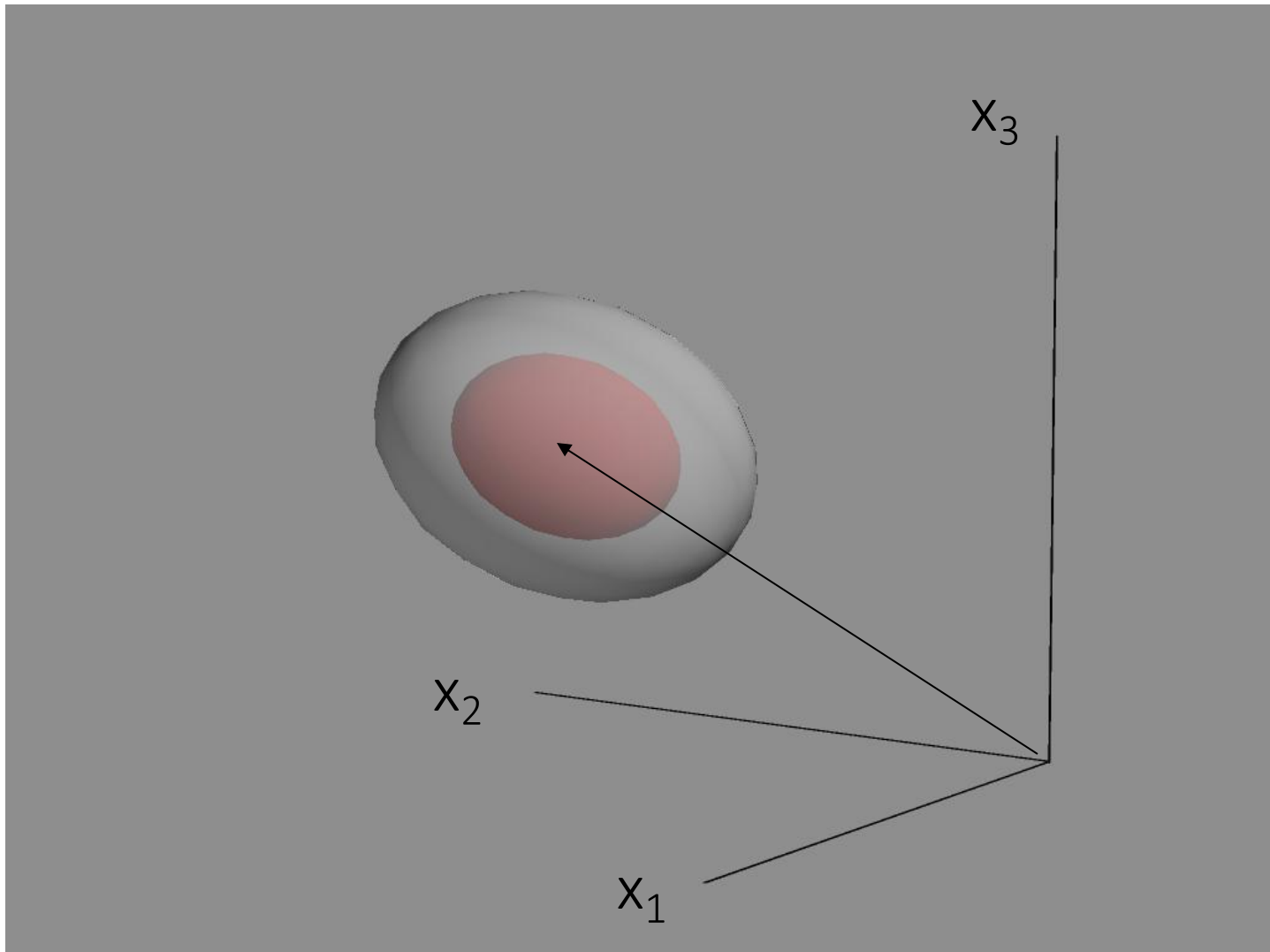




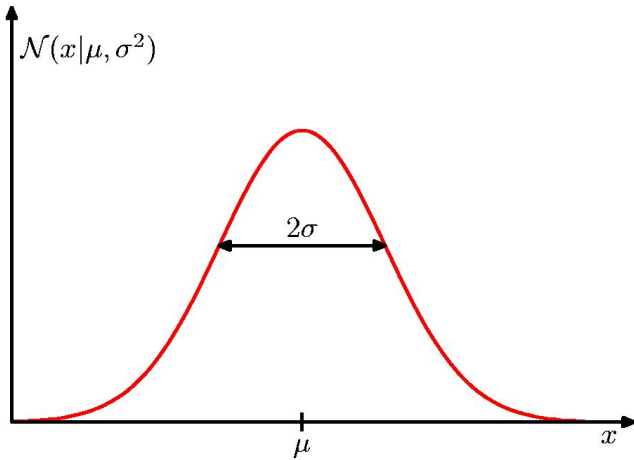
# Scatter Plots of data from the bivariate Normal distribution



# Trivariate Normal distribution



# How to Estimate 1D Gaussian: MLE



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

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

# How to Estimate p-D Gaussian: MLE

$$\langle X_1, X_2, \dots, X_p \rangle \sim N(\vec{\mu}, \Sigma)$$

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

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

*(Note: The handwritten text "i-th feature" is circled and has an arrow pointing to the superscript (i) in the equation above.)*

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \vdots & \ddots & \ddots & \vdots \\ \text{Cov}(X_i, X_1) & \text{Cov}(X_i, X_2) & \dots & \text{Cov}(X_i, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \dots & \text{Var}(X_p) \end{bmatrix} \quad \begin{matrix} \vdots \\ \vdots \\ \vdots \\ \vdots \end{matrix}$$

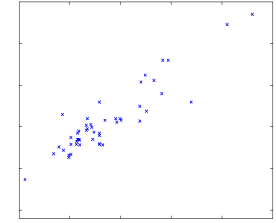
*(Note: The handwritten text "i" is written above the i-th row of the covariance matrix.)*

# 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

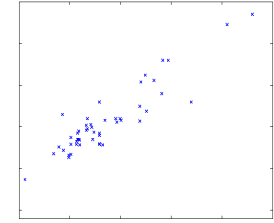


- Let us assume that the target variable and the inputs are related by the equation:

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

where  $\varepsilon$  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:

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

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

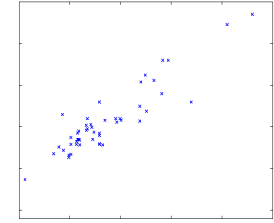
where  $\varepsilon$  is an error term of unmodeled effects or random noise<sub>2</sub>

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

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

$$\text{RV } y | x; \theta \sim N(\theta^T x, \sigma)$$

# DETOUR: Probabilistic Interpretation of Linear Regression



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

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

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



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

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

$$\begin{aligned}
 & p \left\{ (\vec{x}_1, y_1) \wedge (\vec{x}_2, y_2) \wedge \dots \wedge (\vec{x}_N, y_N) \right\} \\
 & \stackrel{\text{IID}}{=} \prod_{i=1}^N p(y_i, \vec{x}_i) = \prod_{i=1}^N p(y_i | \vec{x}_i; \theta) p(\vec{x}_i) \\
 & \theta^* = \arg\max_{\theta} \prod_{i=1}^N p(y_i | \vec{x}_i; \theta)
 \end{aligned}$$

Thus under independence Gaussian residual assumption,  
residual square error is equivalent to **MLE** of  $\theta$  !

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



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

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



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

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

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

(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.  $\sum_i x_i \exp(wx_i) = \sum_i x_i y_i \exp(wx_i)$

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

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

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

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

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

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

$$L(\theta)$$

$$\downarrow$$

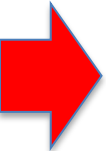
$$L(\theta)$$

$$\downarrow$$

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

# Today

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



# Mean and Variance

- Correlation:

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

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

# 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:

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

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

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

## Some more properties

- The law of Total Variance:

$$\text{Var}(Y) = \text{Var}[E(Y \mid X)] + E[\text{Var}(Y \mid X)]$$



# References

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