Georgia Tech

Machine Learning CS 4641

Probability and Statistics

Mahdi Roozbahani

Slides adapted from Max Roozbahani, Le Song , Sam Roweis, and Chao Zhang.

Georgia Tech

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Probability

P(event) = 1/6

we dont know

- A sample space S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)
  - E.g., S may be the set of all possible outcomes of a dice roll: S
    $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$
  - E.g., S may be the set of all possible nucleotides of a DNA site: S
    $(A \quad C \quad G \quad T)$

  - E.g., S may be the set of all possible time-space positions of an aircraft on a radar screen.
- An Event A is any subset of S
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval

Three Key Ingredients in Probability Theory

A **sample space** is a collection of all possible **outcomes**

Random variables X represents **outcomes** in sample space

random variable

Probability of a random variable to happen $p(x) = p(X = x)$

specific outcome

$p(x) = p$

$p(x) \geq 0$

Continuous variable

Definition: Takes values from a continuous range (e.g., any real number within an interval).

Distribution: Governed by a **probability density function** (PDF).

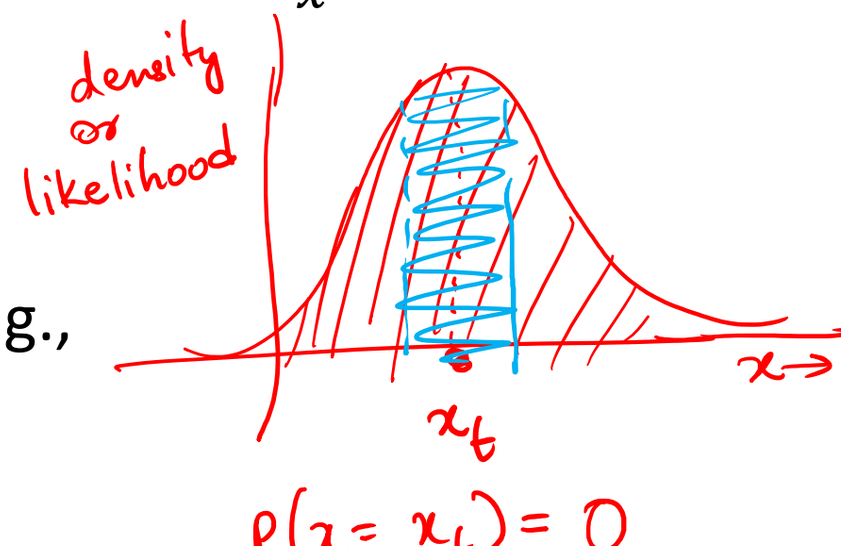
Key Concepts:

- The **density** represents likelihood, but not actual probability at a specific point.
- Example: **Temperature**, which can be any real value (e.g., 72.3°F).
- Common distribution: **Gaussian (Normal) Distribution**.

probability distribution function

Any dist. function is a probability function if area under curve = 1

$$\int_x p(x)dx = 1$$



Discrete variable

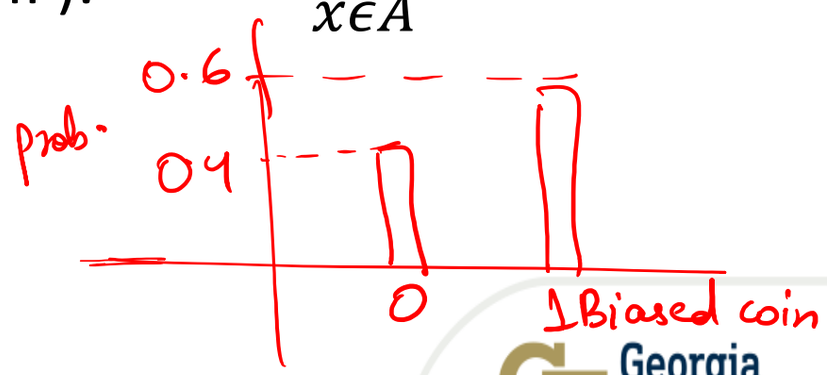
Definition: Takes values from a countable set (e.g., integers).

Distribution: Governed by a **probability mass function** (PMF).

Key Concepts:

- The function directly gives **probability values**.
- Example: **A coin flip** (e.g., 0 for tails, 1 for heads).
- Common distribution: **Bernoulli Distribution**.

$$\sum_{x \epsilon A} p(x) = 1$$



Georgia Tech

$f(x)$ → objective function needs to be optimized

Parameters θ

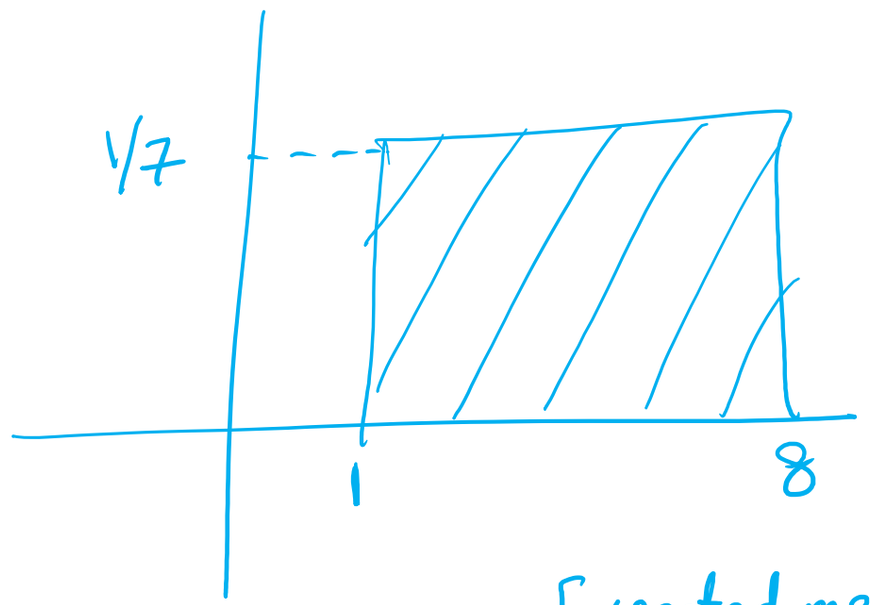
Continuous Probability Functions

- Examples:
  - Uniform Density Function:

$$f_x(x) = \begin{cases} \frac{1}{b-a} & for\ a \le x \le b \\ 0 & otherwise \end{cases}$$

$\theta \in a, b$ (min, max)

$1/7 \times (8-1) = 1$



  - Exponential Density Function:

$$f_x(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$

$\theta \in a$

$for\ x \ge 0$

$f(x) = \frac{1}{a} e^{-x/a}$

$a = \mu$

μ is average → Expected mean or weighted average

→ arithmetic mean

  - Gaussian(Normal) Density Function

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\theta \in \mu, \sigma$

Discrete Probability Functions

- Examples:
  - Bernoulli Distribution:
    - $\begin{cases} 1-p & for\ x=0 \\ p & for\ x=1 \end{cases}$

In Bernoulli, just a **single** trial is conducted

- Binomial Distribution:
  - $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$

k is number of successes

n-k is number of failures

$\binom{n}{k}$ The total number of ways of selection **k** distinct combinations of **n** trials, **irrespective of order**.

E.g. - Given a biased coin with $\mu = 0.3$ for tail $(x = 1)$, what is the probability of getting 4 heads given we flip the coin 10 times?

Georgia Tech

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Example

X = Throw a dice

Y = Flip a coin

X and **Y** are random variables

N = total number of trials

n_{ij} = Number of occurrence

X

Y	$x_{i=1}=1$	$x_{i=2}=2$	$x_{i=3}=3$	$x_{i=4}=4$	$x_{i=5}=5$	$x_{i=6}=6$	C_j
$y_{j=2}=tail$	$n_{ij}=3$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=5$	$n_{ij}=1$	$n_{ij}=5$	20
$y_{j=1}=head$	$n_{ij}=2$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=1$	15
C_i	5	6	6	7	5	6	N=35

X

Y	$x_{i=1}=1$	$x_{i=2}=2$	$x_{i=3}=3$	$x_{i=4}=4$	$x_{i=5}=5$	$x_{i=6}=6$	C_j
$y_{j=2}=tail$	$n_{ij}=3$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=5$	$n_{ij}=1$	$n_{ij}=5$	20
$y_{j=1}=head$	$n_{ij}=2$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=1$	15
C_i	5	6	6	7	5	6	N=35

JOINT PROBABILITY: $p(x=1, y=t) = \frac{3}{35} = \frac{n_{ij}}{N}$

MARGINAL PROBABILITY: $p(x=4) = \frac{7}{35} = \frac{C_i}{N}$

$p(y=t) = \frac{20}{35} = \frac{C_j}{N}$

SUM RULE: $p(x) = \sum_y p(x,y)$

$p(y) = \sum_x p(x,y)$

CONDITIONAL PROBABILITY: $p(x=3 \mid y=t) = \frac{2}{20} = \frac{n_{ij}}{C_j}$ Sample shrinked from 35 outcomes to 20

$p(y=t \mid x=3) = \frac{2}{6} = \frac{n_{ij}}{C_i}$ x is a more informative feature than y.

PRODUCT RULE: $p(x,y) = p(x \mid y) \cdot p(y)$

$= p(y \mid x) \cdot p(x)$

Georgia Tech

Probability:

$$p(X = x_i) = \frac{c_i}{N}$$

Joint probability:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional probability:

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Sum rule

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \Rightarrow p(X) = \sum_{Y} P(X, Y)$$

Product rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p(Y = y_j | X = x_i)p(X = x_i)$$

$$p(X, Y) = p(Y|X)p(X)$$

Georgia Tech

$P(X|Y) = P(X) \rightarrow$ X is C.I. of Y.

Conditional Independence

- Examples:

Virus is independent conditionally of drinking beer

$P(Virus|\ DrinkBeer) = P(Virus)$

iff Virus is independent of Drink Beer

$P(Flu \mid Virus, DrinkBeer) = P(Flu \mid Virus)$

iff Flu is independent of Drink Beer, given Virus

$P(Headache \mid Flu, Virus, DrinkBeer)$
$= P(Headache|Flu, DrinkBeer)$

iff Headache is independent of Virus, given Flu and Drink Beer

$P(h, f, v, d) =$
$P(h \mid f, v, d) \cdot P(f, v, d)$
$P(h \mid f, d) \cdot \frac{P(f \mid v, d)}{P(v, d)}$
$P(h \mid f, d) \cdot P(f \mid v) \cdot P(v, d)$
$P(h \mid f, d)\ P(f \mid v)\ P(v \mid d)\ P(d)$
$P(h \mid f, d)\ P(f \mid v)\ P(v)\ P(d)$

Assume the above independence, we obtain:

$P(Headache, Flue, Virus, DrinkBeer)$
$= P(Headache|Flu, Virus, DrinkBeer)\ P(Flu|Virus, DrinkBeer)$
$P(Virus|DrinkBeer)P(DrinkBeer)$
$= P(Headache|Flu, DrinkBeer)P(Flu|Virus)P(Virus)P(DrinkBeer)$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

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

$$P(x,y) = P(y|x)\,P(x) = P(x|y)\,P(y)$$

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)\,P(y)}{P(x)} = \sum_{y} P(x, y=y) = \sum_{y=cat,dog} P(x|y=y)\,P(y=y)$$

Bayes' Rule

- P(X|Y)= Fraction of the worlds in which X is true given that Y is also true.

- For example:
  - H="Having a headache"
  - F="Coming down with flu"
  - $P(Headche|Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?

- Definition:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{\mathrm{P(Y|X)P(X)}}{\mathrm{P(Y)}}$$

Corollary:

$$P(X,Y) = P(Y|X)P(X)$$

This is called Bayes Rule

Bayes' Rule

- $P(Headache|Flu) = \frac{P(Headache,Flu)}{P(Flu)}$

$$= \frac{P(Flu|Headache)P(Headache)}{P(Flu)}$$

Other cases:

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y)+P(X|\neg Y)P(\neg Y)}$
- $P(Y = y_i|X) = \frac{P(X|Y)P(Y)}{\Sigma_{i\in S} P(X|Y = y_i)P(Y=y_i)}$
- $P(Y|X,Z) = \frac{P(X|Y,Z)P(Y,Z)}{P(X,Z)} =$

$$\frac{P(X|Y,Z)P(Y,Z)}{P(X|Y,Z)P(Y,Z)+P(X|\neg Y,Z)P(\neg Y,Z)}$$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

E[·] ⇝ expected value / weighted avg E(·) ⇝ function g(·) h(·)

Mean and Variance

- Expectation: The mean value, center of mass, first moment:

$$E_X[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)dx = \mu$$

∑ g(x) P(x)

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x-\mu)^n$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} xp_X(x)dx$
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$
- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

E[b]+E[c]

E[a+b+c] = E[a]+E[b+c]

E[(X - E[X])²]

$= \frac{1}{N}\sum_{i=1}^{N}(x^{i,j} - M)^2$

Var(x) = E[x²] - E[x]²

Expected value and average

$f(x) = x \qquad X = [1, 2, 3]$

$P(x) = [\frac{1}{6}, \frac{2}{6}, \frac{3}{6}]$

$avg(x) = \mu = \frac{1+2+3}{3} = 2$

$E[f(x)] = \sum_{i=1}^{3} P(x) f(x) = P(x=1) f(x=1) + \dots + P(x=3) f(x=3)$

$= \frac{1}{6} * 1 + \frac{2}{6} * 2 + \frac{3}{6} * 3 =$

$= \frac{14}{6}$

$\frac{1}{N} \sum (\cdot)$

$\Rightarrow \mu = \frac{1}{N} \sum_{i=1}^{N} (x^{[i]}) =$

$E[f(x)] \neq \mu$

$X = [1, 2, 2, 3, 3, 3] \qquad avg(x) = \mu = \frac{1+2+2+3+3+3}{6} = \frac{14}{6}$

Expectation determined by probability distribution. Arithmetic average determined by observed outcomes of trials.

Both are the same in our class because we assume we have sufficient data that models the distribution. So, the arithmetic average is a good estimate of the true expectation (Law of Large Numbers).

Georgia Tech

Variance and average:

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}_{n \times d} \quad h = \text{height}$$

$$\Rightarrow \mu_{x_h} = 2$$

$$\sigma_h^2 = \frac{1}{N} \sum_{i=1}^{N} \left(x_1^{\{i\}} - \mu_{x_h} \right)^2 \quad \text{(variance)}$$

$$\sigma_h^2 = \left(\frac{1}{N} \sum_{i=1}^{N} (\cdot) \right) = E[\cdot]$$

$$x^{\{1\}} = 1 \qquad x_1 = [1, 2, 3]$$

$$= E\left[\left(x^{\{i\}} - \mu_{x_h} \right)^2 \right] = E\left[\left(x^{\{i\}} - E[x] \right)^2 \right]$$

$$\bar{X}_{3 \times 1} = \begin{bmatrix} \bar{h} = h - \mu_h \\ 1 - \mu_h \\ 2 - \mu_h \\ 3 - \mu_h \end{bmatrix} = \begin{bmatrix} h \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$\mu_{\bar{h}} = 0$$

$$\sigma_h^2 = \frac{1}{N} \underset{1 \times 3}{\bar{X}^T} \underset{3 \times 1}{\bar{X}} = \frac{1}{N} \begin{bmatrix} 1 - \mu_h & 2 - \mu_h & 3 - \mu_h \end{bmatrix} \begin{bmatrix} 1 - \mu_h \\ 2 - \mu_h \\ 3 - \mu_h \end{bmatrix} = \frac{1}{N} \left[(1 - \mu_h)^2 + \cdots + (3 - \mu_h)^2 \right]$$

Georgia Tech

Covariance:

$$X = \begin{bmatrix} h & w = \text{weight} \\ 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$X^{[i]} = [1, 4] \quad X_1 = [1, 2, 3]$

$$\bar{X} = \begin{bmatrix} \bar{h} & \bar{w} \\ 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix}$$

$$Cov = \frac{1}{N} \bar{X}^T \bar{X} = \frac{1}{N} \begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \\ 4-\mu_w & 5-\mu_w & 6-\mu_w \end{bmatrix} \begin{bmatrix} 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix}$$

$$Cov = \begin{matrix} h \\ w \end{matrix} \begin{bmatrix} h & w \\ \sigma_h^2 & \sigma_{hw} \\ \sigma_{wh} & \sigma_w^2 \end{bmatrix}_{dxd}$$

$\sigma_{hw} = \sigma_{wh}$

$\sigma_{hw} = 0$

Georgia Tech

Correlation:

$$Cor = \frac{1}{N} \bar{X}^{*T} \bar{X}^{*}$$

$$\bar{X} = \begin{bmatrix} h & \omega \\ 1-\mu_h & 4-\mu_\omega \\ \vdots & \vdots \end{bmatrix}$$

$\bar{X}^{*}$ → Standardization

$$\bar{h}^{*} = \frac{\bar{h}}{\sigma_h} \qquad \bar{\omega} = \frac{\bar{\omega}}{\sigma_\omega}$$

$$\begin{bmatrix} \frac{1-\mu_h}{\sigma_h} & \frac{4-\mu_\omega}{\sigma_\omega} \\ \vdots & \vdots \end{bmatrix}$$

$$Cor = \frac{1}{N} \begin{bmatrix} \frac{1-\mu_h}{\sigma_h} & \cdots \\ \frac{4-\mu_\omega}{\sigma_\omega} & \end{bmatrix} \begin{bmatrix} \frac{1-\mu_h}{\sigma_h} & \frac{4-\mu_\omega}{\sigma_\omega} \\ \vdots & \\ | & \end{bmatrix}$$

$$Cor = \begin{matrix} h \\ \omega \end{matrix} \begin{bmatrix} h & \omega \\ \frac{\sigma^2 h}{\sigma^2 h} = 1 & -1 \le \sigma_{h\omega} \le 1 \\ -1 \le \sigma_{h\omega} \le 1 & \frac{\sigma^2_\omega}{\sigma^2_\omega} = 1 \end{bmatrix}$$

EDA

Useful in EDA. If features are correlated, then data may have redundancy.

Georgia Tech

Uncorrelated vs Independent RV

Uncorrelated (Definition: Cov(X,Y)=0.)
•Means no **linear relationship** between X and Y.
•Does **not** rule out non-linear dependence.
•Example: X∼U(−1,1), Y=x^2 → Uncorrelated but dependent.

Independent (Definition: P(X∈A,Y∈B)=P(X∈A)P(Y∈B))
•Stronger condition: knowing one variable gives **no information** about the other.
•Independence ⇒ Uncorrelated (if variances exist).

Key Differences
•Independence is a **stronger** property.
•Uncorrelated only removes **linear relationships**.
•Uncorrelated ⇏ Independent.

ML Relevance
- Most ML models care about uncorrelatedness because they only model linear relationships.
- True independence is rarer and much harder to check, but it’s what we assume in stronger probabilistic models

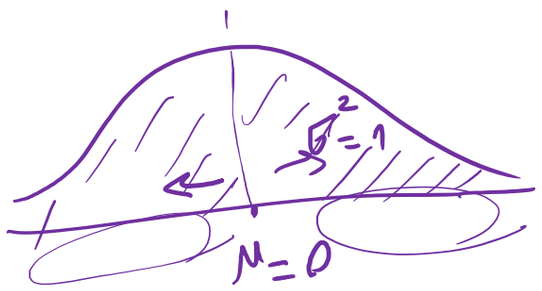
For Joint Distributions

$$Cov(x,y) = \cancel{E[z^3]} - \cancel{E[z]}\,E[z^2] = 0$$

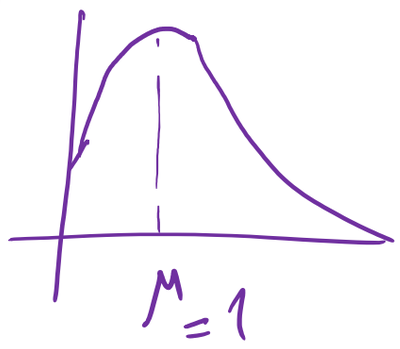
- Expectation and Covariance:
  - $E[X+Y] = E[X] + E[Y]$
  - $cov(X,Y) = E[(X - E_X[X])(Y - E_Y(Y)] = E[XY] - E[X]E[Y]$
  - $Var(X+Y) = Var(X) + 2cov(X,Y) + Var(Y)$

$X = z \rightsquigarrow$ Standard gaussian

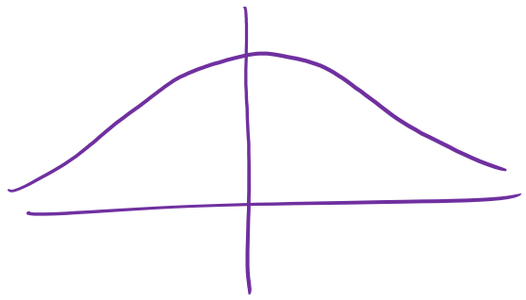
$\mu = 0 \quad \sigma^2 = 1$



$Y = z^2 \rightsquigarrow$ chi-squared



$h = z^3$



$Cov(X,Y) =$

$Var(z) = E[z^2] - E[z]^2$

$1 = E[z^2] - 0$

$E[z^2] = 1$

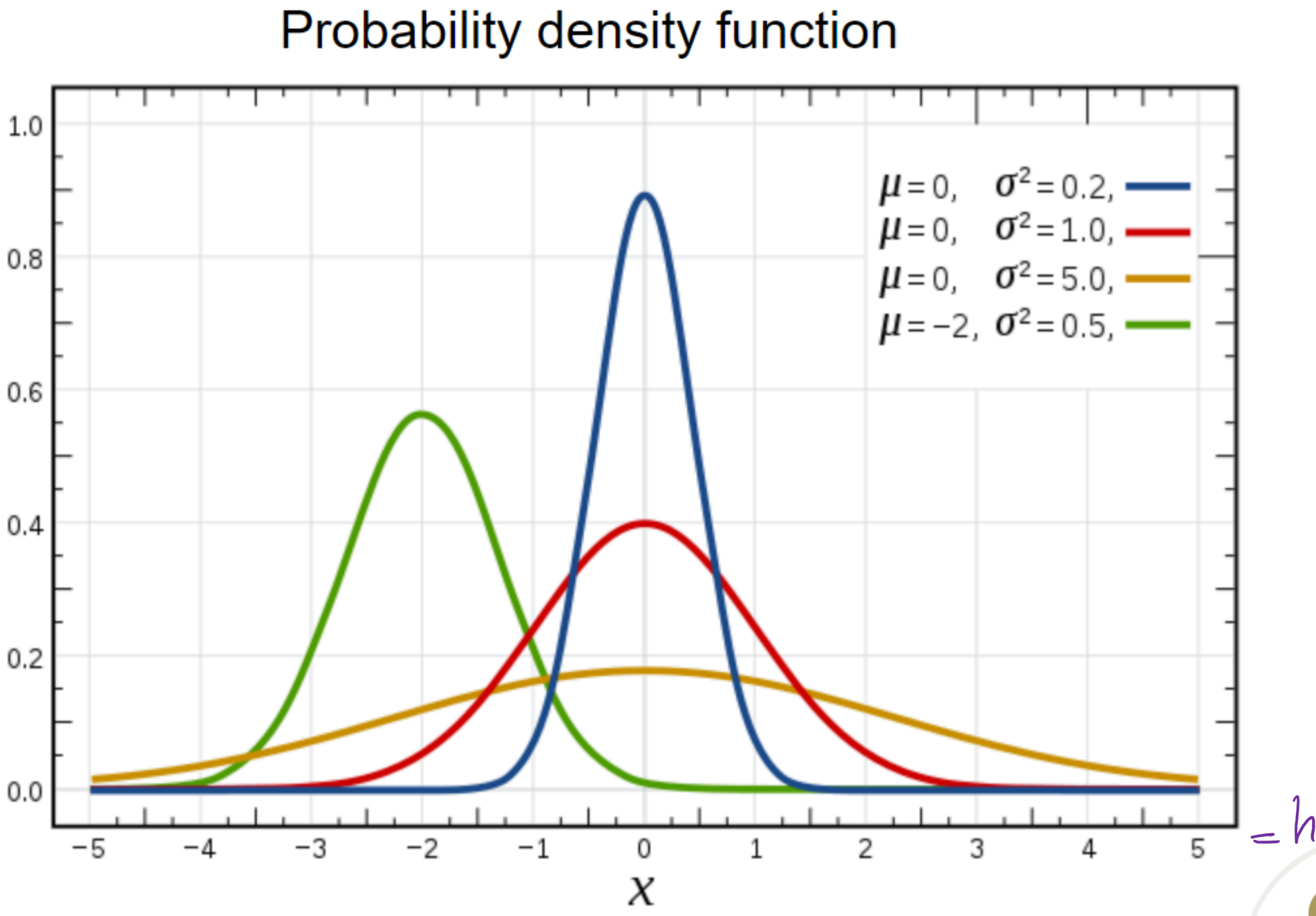
Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Gaussian Distribution

$\{\mu, \sigma\} \in \theta$

- Gaussian Distribution: $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



$= h$

[Probability versus likelihood]

Multivariate Gaussian Distribution

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\}$$

Cov (above Σ^{-1})

Cov (below $|\Sigma|$)

- Moment Parameterization $\mu = E(X)$

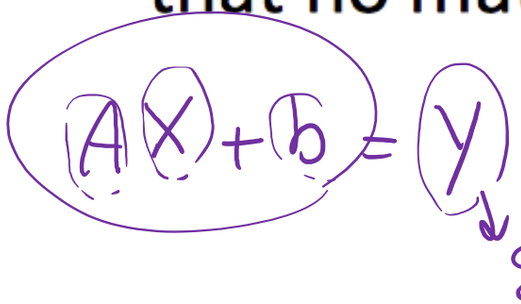
$$\Sigma = Cov(X) = E[(X-\mu)(X-\mu)^\top]$$

- Mahalanobis Distance $\Delta^2 = (x-\mu)^\top \Sigma^{-1}(x-\mu)$

- Tons of applications (MoG, FA, PPCA, Kalman filter,…)

Properties of Gaussian Distribution

- The linear transform of a Gaussian r.v. is a Gaussian. Remember that no matter how x is distributed



$$E(AX + b) = AE(X) + b$$

$$Cov(AX + b) = ACov(X)A^{\top}$$

this means that for Gaussian distributed quantities:

$$X \sim N(\mu, \Sigma) \rightarrow AX + b \sim N(A\mu + b, A\Sigma A^{\top})$$

- The sum of two independent Gaussian r.v. is a Gaussian

$$\mathrm{Y} = \mathrm{X}_1 + \mathrm{X}_2,\ \mathrm{X}_1 \perp X_2 \rightarrow \mu_y = \mu_1 + \mu_2, \Sigma_y = \Sigma_1 + \Sigma_2$$

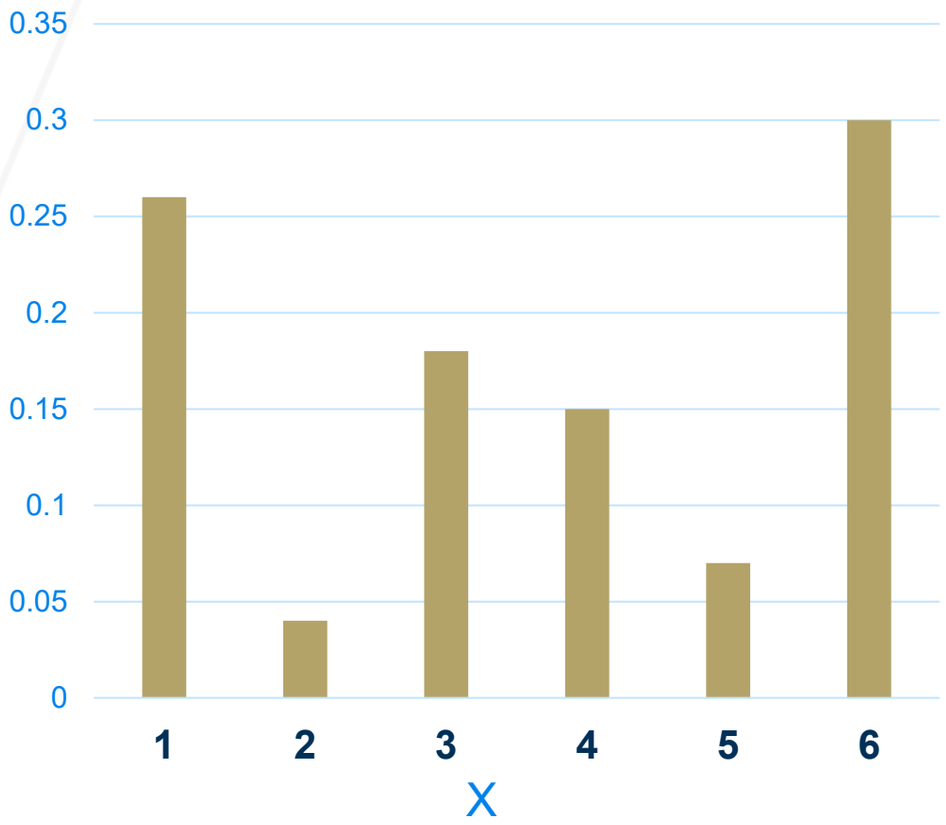
- The multiplication of two Gaussian functions is another Gaussian function (although no longer normalized)

$$N(a, A)N(b, B) \propto N(c, C),$$

$$where\ C = (A^{-1} + B^{-1})^{-1}, c = \mathrm{C}\mathrm{A}^{-1}a + CB^{-1}b$$

Central Limit Theorem

Probability mass function of a **biased** dice



Let's say, I am going to get a sample from this pmf having a size of $\boldsymbol{n = 4}$

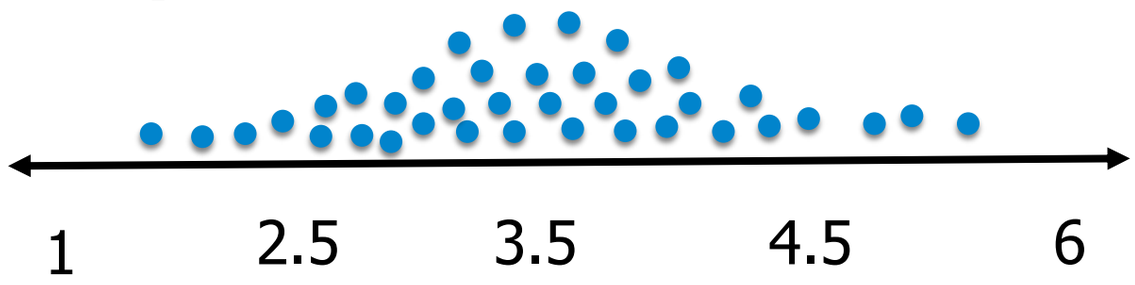
$$S_1 = \{1,1,1,6\} \Rightarrow E(S_1) = 2.25$$

$$S_2 = \{1,1,3,6\} \Rightarrow E(S_2) = 2.75$$

$$\vdots$$

$$S_m = \{1,4,6,6\} \Rightarrow E(S_m) = 4.25$$

- According to CLT, if you sample enough from any distribution with finite variance you will get an approximate Gaussian distribution.
- No matter what the population looks like, the average of many samples looks Normal.
- Explains **why the Normal distribution is everywhere** in statistics and ML.



Georgia Tech

Prob vs Likelihood

Probability predicts data from parameters.
Likelihood evaluates parameters from data.

- **Probability**
  - Forward direction: given parameters, what's the chance of data?
  - $P(data \mid \theta)$
  - Example: *If the coin has bias* $\theta = 0.7$*, what's the probability of 8 heads in 10 tosses?*
  - Varies with **different outcomes** of data.
- **Likelihood**
  - Reverse view: given data, how plausible are parameters?
  - $L(\theta \mid data) \propto P(data \mid \theta)$
  - Example: *I observed 8 heads in 10 tosses. How likely is it that the coin's bias is* $\theta = 0.7$*?*
  - Varies with **different parameter values**.

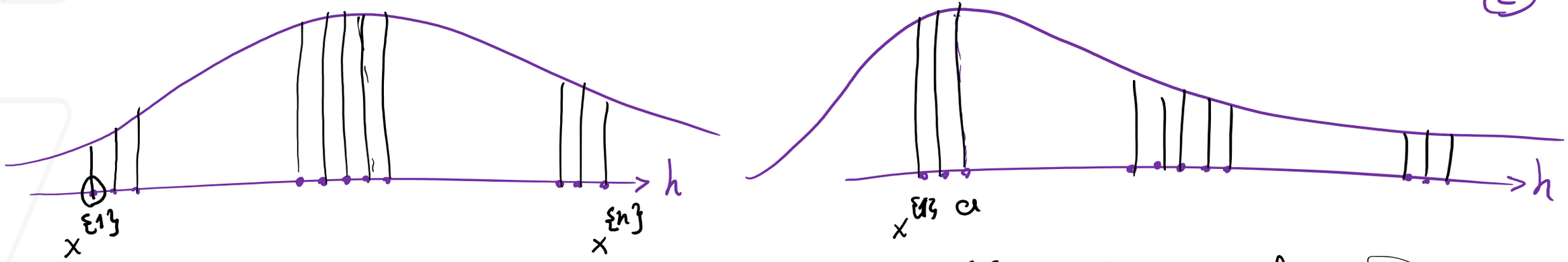
Georgia Tech

Prob vs Likelihood

$$P(X=T) = \frac{1}{2}, \qquad T, T, T, H \Rightarrow P(X=T) = \frac{3}{4}$$

$$f(x^{\{i\}}) = f(x^{\{i\}}|\theta) = f(x^{\{i\}}|\mu, \sigma) = f(x^{\{i\}}|a,b) = \frac{1}{\sqrt{2\pi b^2}} \exp\left(-\frac{(x^{\{i\}}-a)^2}{2b^2}\right)$$

① ②



$$L(\theta | x^{\{1\}}, \ldots, x^{\{n\}}) = L(a,b|x) \rightsquigarrow f(x^{\{1\}}, \ldots, x^{\{n\}}|\theta) = f(x|\theta) = f(x|a,b)$$

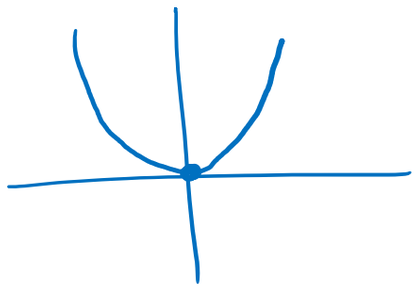
$$\left[f(a,b|c) = f(a|c) f(b|c) \right]$$

$$\text{Max } L(a,b|x) = \max f(x^{\{1\}}|\theta) f(x^{\{2\}}|\theta) \cdots f(x^{\{n\}}|\theta) = \text{Max} \prod_{i=1}^{N} f(x^{\{i\}}|\theta) \quad \theta \in \{a,b\}$$

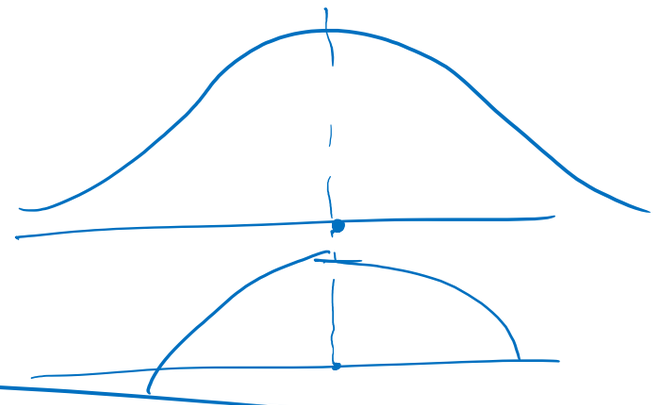
Georgia Tech

Prob vs Likelihood

$f(x) = x^2$



$$\frac{\partial f(x)}{\partial x} = 2x = 0 \Rightarrow x = 0$$



max

Maximizing Likelihood

$$L(\theta|x) = \text{Max} \prod_{i=1}^{N} f(x^{\{i\}} | a,b) \rightsquigarrow$$

Maximizing log likelihood

Objective function

always scalar

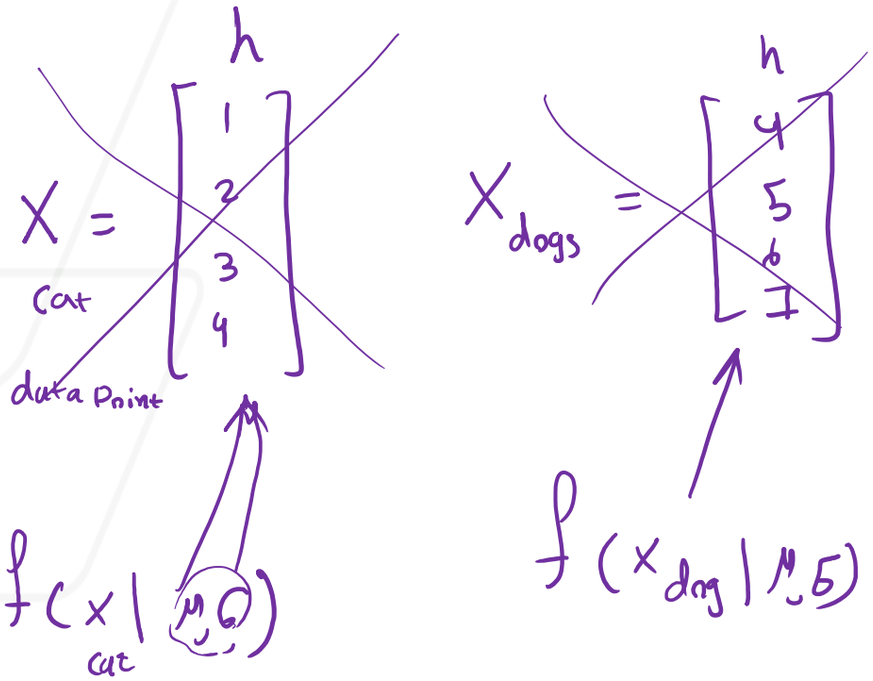
$$\log(a.b) = \log a + \log b$$

$$\text{Max}\ l(\theta|x) = \text{Max}\ \text{Log}\ L(\theta|x) = \text{Max} \sum_{i=1}^{N} \log f(x^{\{i\}} | a,b) \Rightarrow \text{MLE}$$

$$\frac{\partial l(a,b|x)}{\partial a} = 0 \Rightarrow a = ? \qquad a = \frac{1}{N} \sum_{i=1}^{N} x^{\{i\}} = \mu$$

$$\frac{\partial l(a,b|x)}{\partial b} = 0 \Rightarrow b = ? \qquad b = \frac{1}{N} \sum_{i=1}^{N} (x^{\{i\}} - \mu)^2 = \sigma^2$$

Georgia Tech



ML → ① what is the problem you are trying to solve

② we need to come up with an Objective function

③ training ⇝ Optimization

④ Testing

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Maximum Likelihood Estimation

- Probability: inferring probabilistic quantities for data given fixed models (e.g. prob. of events, marginals, conditionals, etc).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).

Main assumption:

Independent and identically distributed random variables

i.i.d

Maximum Likelihood Estimation

For Bernoulli (i.e. flip a coin):

Objective function: $P\left(x^{\{i\}}|\theta\right) = \theta^{x^{\{i\}}}(1-\theta)^{1-x^{\{i\}}}$ $x^{\{i\}} \in \{0,1\}\ or\ \{head, tail\}$

$$L(\theta|X) = L(\theta|X = x^{\{1\}}, X = x^{\{2\}}, X = x^{\{3\}}, \dots, X = x^{\{n\}})$$

i.i.d assumption

$$L(\theta|X) = \prod_{i=1}^{n} P\left(x^{\{i\}}|\theta\right)$$

$$L(\theta|X) = \prod_{i=1}^{n} P\left(x^{\{i\}}|\theta\right) = \prod_{i=1}^{n} \theta^{x^{\{i\}}}(1-\theta)^{1-x^{\{i\}}}$$

$$L(\theta|X) = \theta^{x^{\{1\}}}(1-\theta)^{1-x^{\{1\}}} \times \theta^{x^{\{2\}}}(1-\theta)^{1-x^{\{2\}}} \dots \times \theta^{x^{\{n\}}}(1-\theta)^{1-x^{\{n\}}} =$$
$$= \theta^{\Sigma\, x^{\{i\}}}(1-\theta)^{\Sigma(1-x^{\{i\}})}$$

Georgia Tech

We don't like multiplication, let's convert it into summation

What's the trick? Take the log

$\log a^b = b \log a$

$\frac{\partial \log(a)}{\partial \theta} = \frac{1}{\theta}$

$$L(\theta|X) = \theta^{\sum x^{\{i\}}}(1-\theta)^{\sum(1-x^{\{i\}})}$$

$$logL(\theta|X) = l(\theta|X) = \log(\theta)\sum_{i=1}^{n} x^{\{i\}} + \log(1-\theta)\sum_{i=1}^{n}(1-x^{\{i\}})$$

a b

How to optimize θ?

$$\frac{\partial l(\theta|X)}{\partial \theta} = 0 \qquad \frac{\sum_{i=1}^{n} x^{\{i\}}}{\theta} - \frac{\sum_{i=1}^{n} 1 - x^{\{i\}})}{1-\theta} = 0$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x^{\{i\}}$$

1,1,1,1,1,1,1 0,0,0

$\theta = \frac{1+1+1 \cdots + 0+0}{10} = 0.7$

Georgia Tech