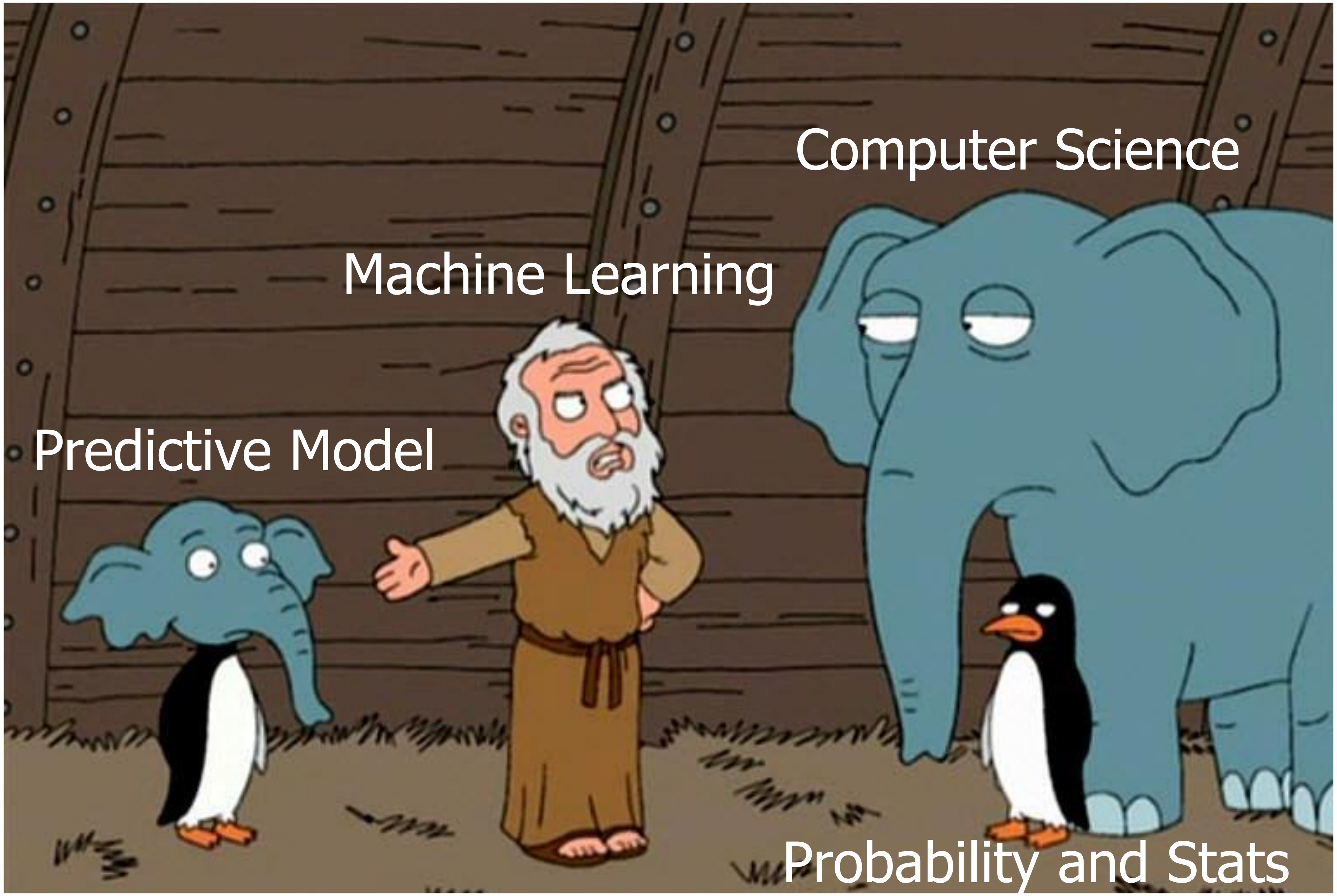
Probability and Statistics

Mahdi Roozbahani
Georgia Tech

These slides are inspired based on slides from Le Song , Sam Roweis, and Chao Zhang.

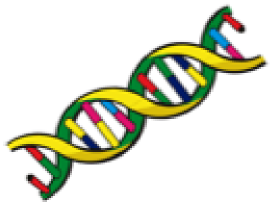

Computer Science
Machine Learning
Predictive Model
Probability and Stats

Outline

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

Probability

- 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

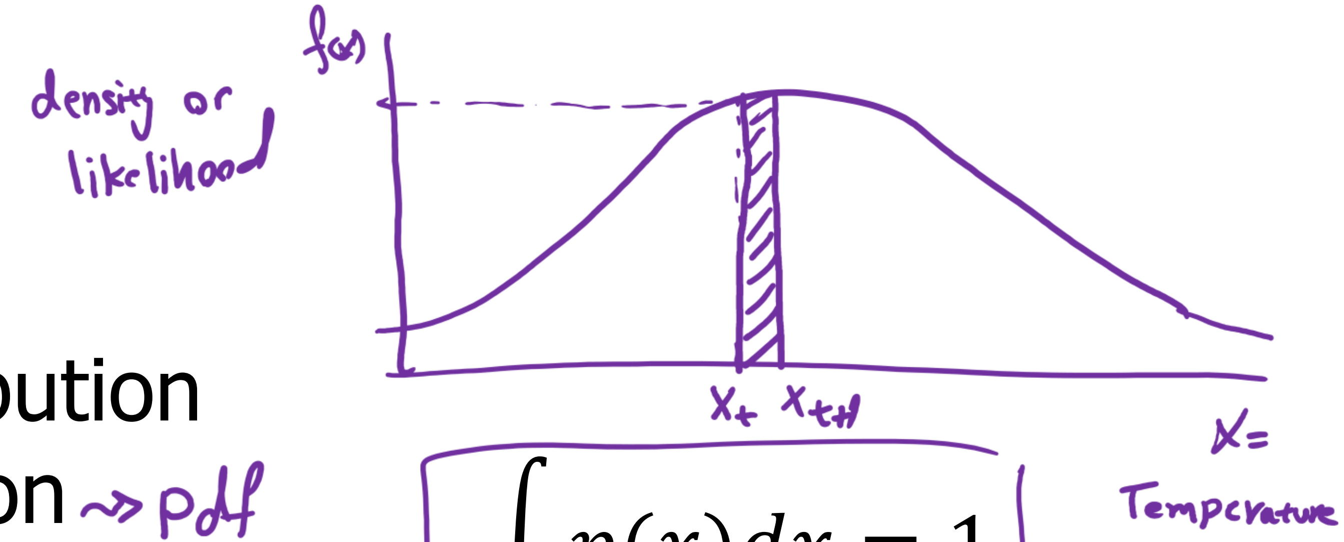
Three Key Ingredients in Probability Theory

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

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

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

$$p(x) \geq 0$$



Continuous variable

Continuous probability distribution

Probability density function ⇝ pdf

Density or likelihood value

Temperature (real number)

Gaussian Distribution

$$\boxed{\int_x p(x)dx = 1}$$

$$P(X = X_t) = 0$$

Probability distribution function (pdf)

Probability density function / pmf

Discrete variable

Discrete probability distribution

Probability mass function ⇝ pmf

Probability value

Coin flip (integer)

Bernoulli distribution

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

$$\frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{6} = 1$$

Continuous Probability Functions

Parameters ⇝ Θ

- Examples:

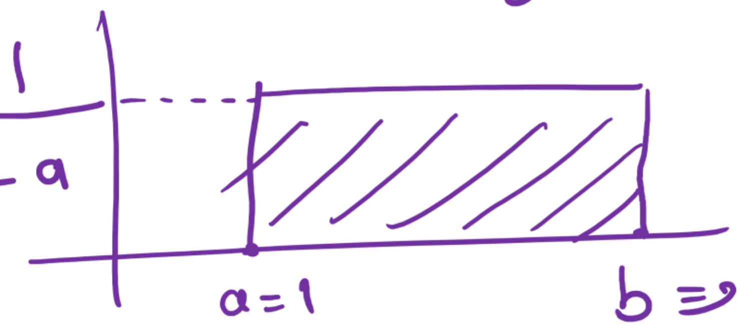
$\{a, b\} \in \Theta$

$\frac{1}{8} * (\overbrace{b - a}^{8}) = 1$

  - Uniform Density Function:

$a=1 \quad b=9 \quad \frac{1}{8} = \frac{1}{b-a}$

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



  - Exponential Density Function:

$\{\mu\} \in \Theta$

$$f_x(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad for\ x \ge 0$$

$$F_x(x) = 1 - e^{\frac{-x}{\mu}} \qquad for\ x \ge 0$$

μ → Arithmetic average

μ → Expected value or weighted avg

  - Gaussian(Normal) Density Function

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

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

$\mu = \frac{1}{N} \sum x_i \qquad \sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$

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

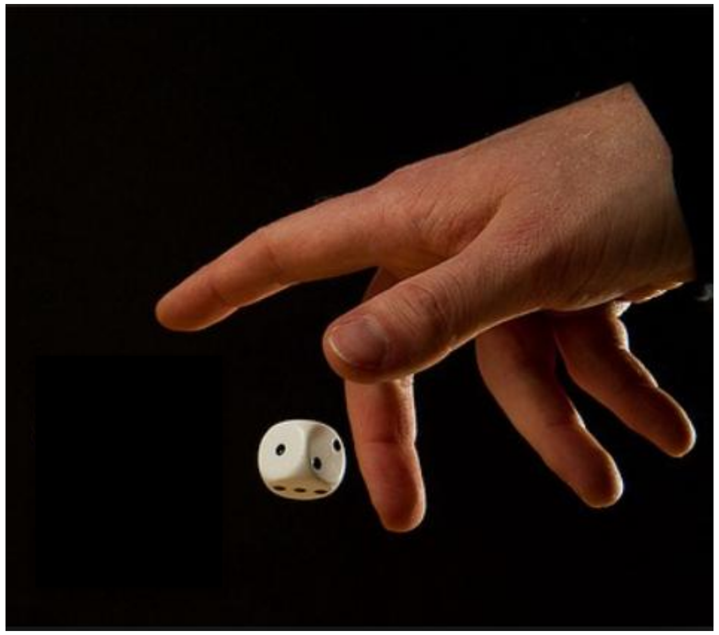
n-k is number of failures

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

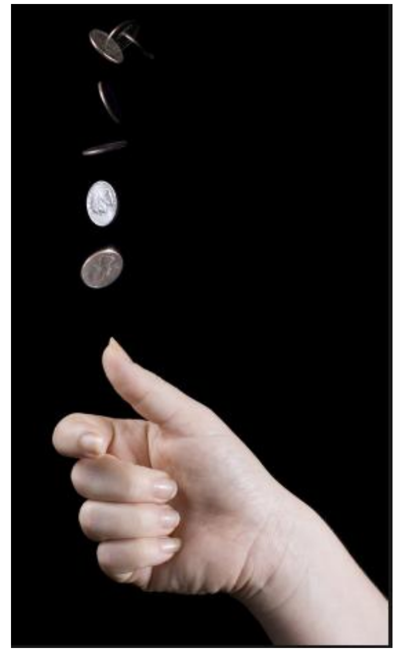
Outline

- 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

$\boldsymbol{n_{ij}}$ = Number of occurrence

Joint Probability table

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

$$P(y=tail, x=4) = \frac{5}{35} = \frac{n_{ij}}{N}$$

$$P(y=tail) = \frac{20}{35} = \frac{C_j}{N} \qquad P(x=4) = \frac{7}{35} = \frac{C_i}{N}$$

$$P(y=tail \mid x=4) = \frac{5}{7} = \frac{n_{ij}}{C_i} \qquad P(x=4 \mid y=tail) = \frac{5}{20} = \frac{n_{ij}}{C_j}$$

$$P(y,x) = \frac{n_{ij}}{N} = \frac{n_{ij}}{C_j}\,\frac{C_j}{N} = P(x \mid \underline{\underline{y}})\, P(\underline{\underline{y}})$$

$\rightsquigarrow$ product rule

$$= \frac{n_{ij}}{C_i}\,\frac{C_i}{N} = P(y \mid \underline{\underline{x}})\, P(\underline{\underline{x}})$$

$$P(b \mid c) P(c)$$

$$P(a,b,c) = P(a \mid \underbrace{b,c}) \underbrace{P(b,c)}$$

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

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

Product rule

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

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

Conditional Independence

$P(V,D) = P(V|D)P(D)$

$P(H,F,V,D) = P(H|F,\cancel{V},D) \cdot P(F,V,D)$

$P(H|F,D) \cdot P(F|V,\cancel{D}) \cdot P(V,D)$

$P(H|F,D) \cdot P(F|V) \cdot P(V|D) \cdot P(D)$

$P(H|F,D) \cdot P(F|V) \cdot P(V) \cdot P(D)$

- Examples:

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

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

Assume the above independence, we obtain:

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

Outline

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

Bayes' Rule

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

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

$P(y) = \sum_{x_i} P(y, x = x_i) = \sum P(y|x_i)\,P(x_i)$

- 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

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

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

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

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

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

Other cases:

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

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

Outline

① product
② sum rule

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

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

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

$$\underline{p(y|x)} = \frac{p(x,y)}{p(x)} = \frac{p(x|y)\, p(y)}{p(x)}$$

$E[\cdot]$

Mean and Variance

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

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

- 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): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

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

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

Mean and average

$\underline{g(x)} = x$

$X = [1, 2, 3]$

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

$$E[g(x)] = \sum_{x_i} g(x=i)\,P(x=i) =$$

$$g(x=1)P(x=1) + g(x=2)P(x=2) + g(x=3)P(x=3)$$

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

$$= \frac{14}{6}$$

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

$$\text{avg}(x) \neq E[g(x)]$$

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

$$\text{avg}(x) = E[x]$$

Variance and average:

$\mu = E[\cdot]$

$\frac{1}{N}\sum_{i=1}^{N}(\cdot) = avg = E[\cdot]$

height $= h = x_1$

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}_{n \times d} \quad \begin{matrix} x^{\{1\}} \\ x^{\{2\}} \\ x^{\{3\}} \end{matrix}$$

$n=3$ & $d=1$

3×1

$$\mu_h = avg_h = \frac{1+2+3}{3} = 2$$

$$\sigma_h^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{\{i\}} - \mu_h\right)^2$$

For loop

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

Matrix operation

$X \to \bar{X}$

$$\begin{array}{c} X \\ h \\ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \end{array} \rightarrow \begin{array}{c} \bar{X} \\ h - \mu_h = \bar{h} \\ \begin{bmatrix} 1-\mu_h = -1 \\ 2-\mu_h = 0 \\ 3-\mu_h = 1 \end{bmatrix} \end{array}$$

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

$$= \frac{1}{N}\sum_{i=1}^{N}\left(x^{\{i\}} - \mu_h\right)^2$$

Covariance:

$$X = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{cc} h & \text{weight} = w \\ \left[\begin{matrix} 1 \\ 2 \\ 3 \end{matrix}\right. & \left.\begin{matrix} 4 \\ 5 \\ 6 \end{matrix}\right] \end{array} \quad 3\times 2 = n\times d$$

$$X \to \bar{X} \Rightarrow \bar{X} = \begin{array}{cc} \bar{h} & \bar{w} \\ h-\mu_h & w-\mu_w \\ \left[\begin{matrix} 1-\mu_h \\ 2-\mu_h \\ 3-\mu_h \end{matrix}\right. & \left.\begin{matrix} 4-\mu_w \\ 5-\mu_w \\ 6-\mu_w \end{matrix}\right] \end{array}$$

$2\times 2 = d\times d$

$$\text{COV} = \frac{1}{N}\, \overline{\underbrace{\bar{X}^T}_{2\times 3}\underbrace{\bar{X}}_{3\times 2}} = \frac{1}{N}\begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \\ 4-\mu_w & 5-\mu_w & 6-\mu_w \end{bmatrix}\begin{bmatrix} 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix} = \begin{array}{c|cc} & h & w \\ \hline h & 1 & 2 \\ w & 3 & 4 \end{array}$$

(1) $\frac{1}{N}\left[(1-\mu_h)^2 + (2-\mu_h)^2 + (3-\mu_h)^2\right] = \sigma_h^2$

(2) $\frac{1}{N}\left[(1-\mu_h)(4-\mu_w) + (2-\mu_h)(5-\mu_w) + (3-\mu_h)(6-\mu_w)\right] = \sigma_{hw}$

(3) $= (2) = \sigma_{wh}$

(4) σ_w^2

$$\text{COV} = \begin{array}{c|cc} & h & w \\ \hline h & \sigma_h^2 & \sigma_{hw} \\ w & \sigma_{wh} & \sigma_w^2 \end{array}$$

Symmetrical

Correlation: EDA

$$X \longrightarrow \bar{X} \longrightarrow \bar{X}^*$$

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

$$\begin{array}{cc} \bar{h}^* & \bar{w}^* \end{array}$$
$$\bar{X}^* = \begin{bmatrix} \frac{1-\mu_h}{\sigma_h} & \frac{4-\mu_w}{\sigma_w} \\ \frac{2-\mu_h}{\sigma_h} & \frac{5-\mu_w}{\sigma_w} \\ \frac{3-\mu_h}{\sigma_h} & \frac{6-\mu_w}{\sigma_w} \end{bmatrix}$$

$$\bar{X}^{*T}$$

Standardization

$$\text{cor} = \frac{1}{N} \bar{X}^{*T} \bar{X}^* = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{d \times d}$$

$$1 \quad \frac{1}{N}\left(\frac{1-\mu_h}{\sigma_h}\right)^2 + \left(\frac{2-\mu_h}{\sigma_h}\right)^2 + \left(\frac{3-\mu_h}{\sigma_h}\right)^2 = \frac{(1-\mu_h)^2 + (2-\mu_h)^2 + (3-\mu_h)^2}{N \, \sigma_h^2} = \frac{\sigma_h^2}{\sigma_h^2} = 1$$

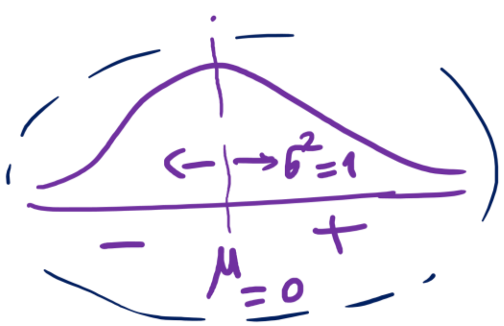
$$\text{cor} = \begin{array}{c|cc} & h & w \\ \hline h & 1 & -1 \le \sigma_{hw} \le 1 \\ w & -1 \le \sigma_{wh} \le 1 & 1 \end{array}$$

For Joint Distributions

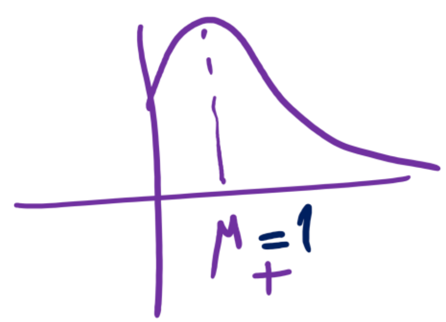
$\mathrm{cov}(x,y) = E[xy] - E[x]E[y] = \underbrace{E[z^3]}_{0} - \underbrace{E[z]}_{0} E[z^2] = 0$

- Expectation and Covariance:
  - $E[X+Y] = \boxed{E[X] + E[Y]}$ $\quad E[a+b+c] = E[a] + E[b+c]$
  - $cov(X,Y) = E[(X - E_X[X])(Y - E_Y(Y)] = E[XY] - E[X]E[Y]$
  - $Var(X+Y) = Var(X) + 2cov(X,Y) + Var(Y)$

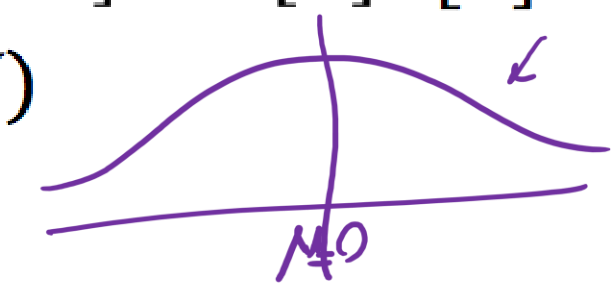
$X = Z$ ⤳ Standard Gaussian distribution



$Y = Z^2$ chi-squared



$Y = Z^3$



$Var(X) = 1 = E[X^2] - E[X]^2$

$1 = E[Z^2] - 0^2 \Rightarrow E[Z^2] = 1$

[Uncorrelated vs Independent RV]

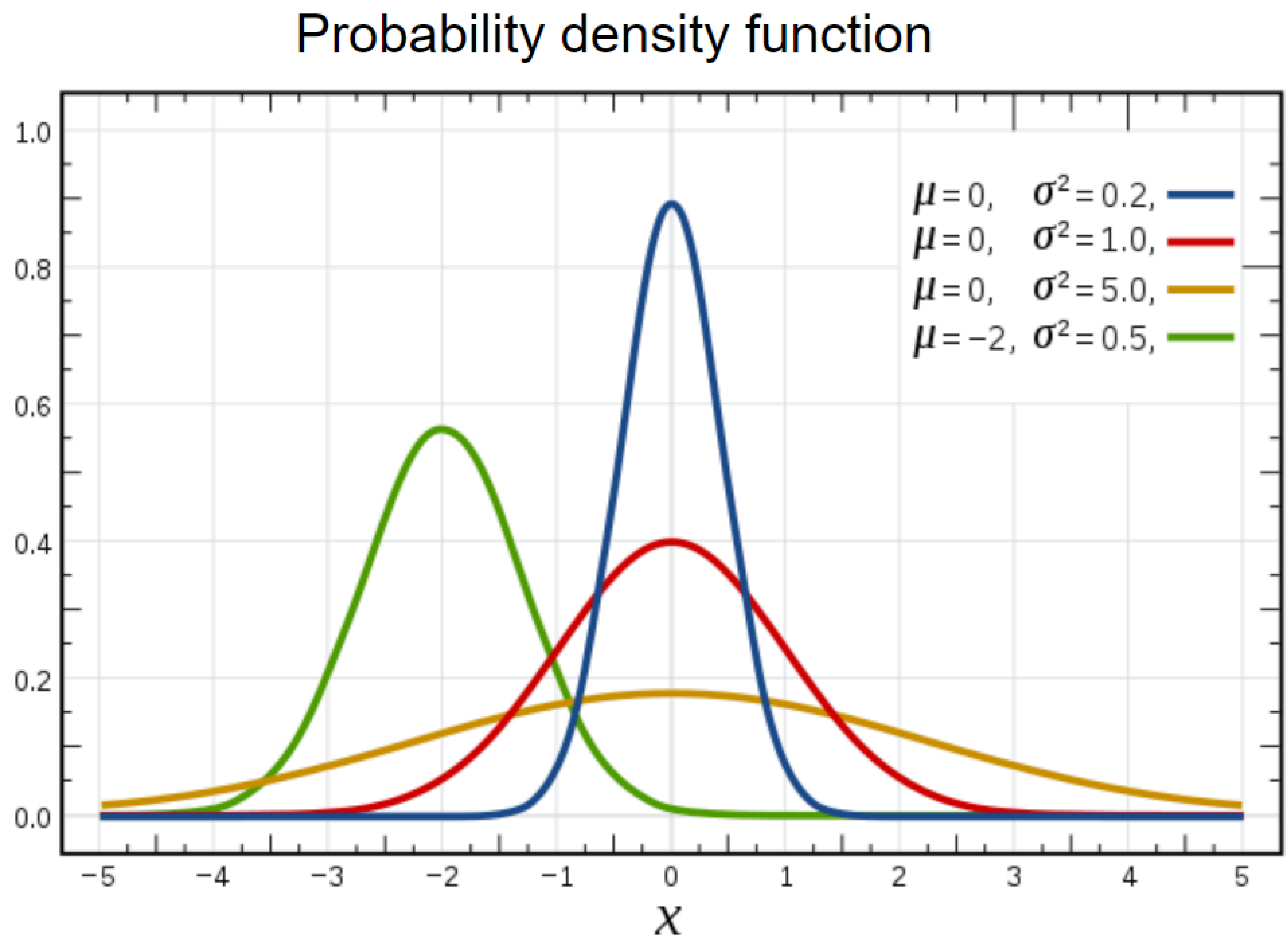
Outline

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

Gaussian Distribution

$\{\mu, \sigma^2\} \in \theta$

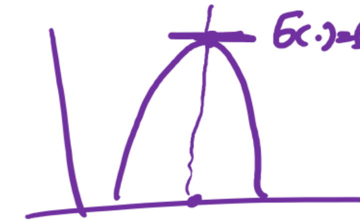
- Gaussian Distribution: $$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



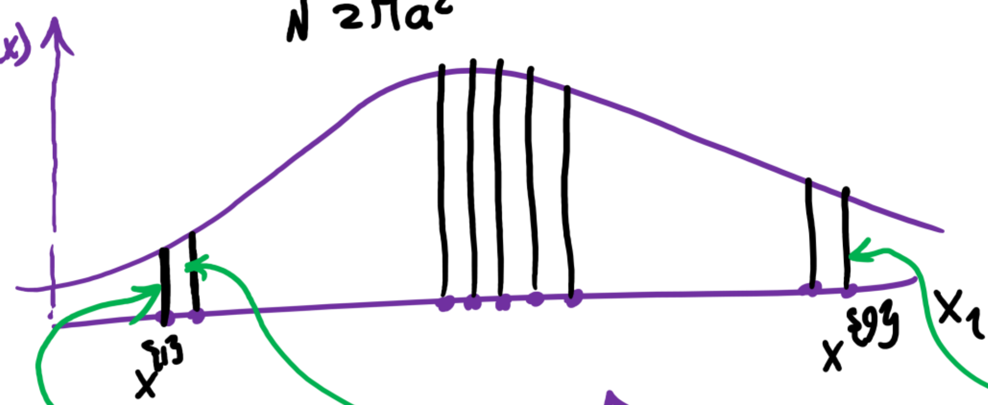
[Probability versus likelihood]

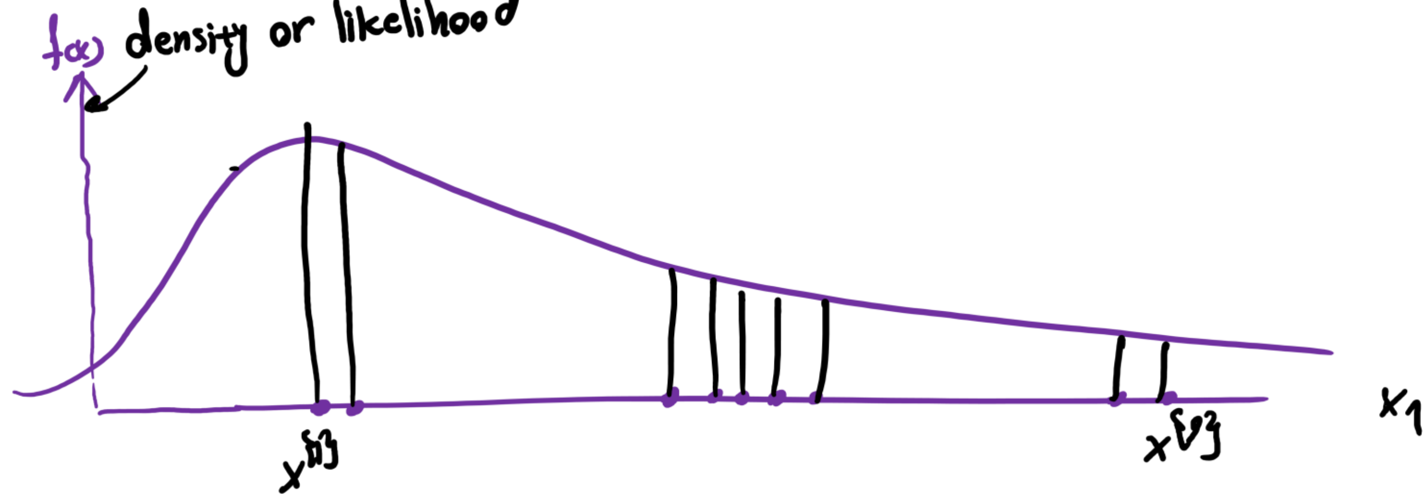
Prob vs Likelihood ①

coin ⇝ $P(x=T)=\frac{1}{2}$ ⇒ ② $[T, T, T, H] \Rightarrow P(x=T)=\frac{3}{4}$



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





objective function

$\text{Max } L(a,b\,|\,X) \rightarrow a=1,\ b=1$

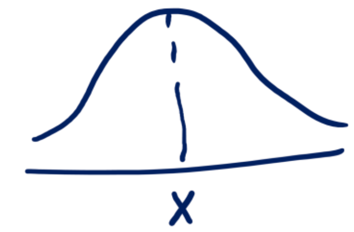
$\rightarrow a=1,\ b=-1$

$$f(x\,|\,a=1,b=1)=f(x^{\{1\}},x^{\{2\}},\ldots,x^{\{9\}}\,|\,a,b)$$

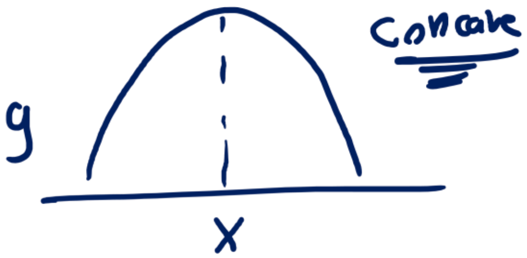
$$f(x^{\{1\}},x^{\{2\}},\ldots,x^{\{9\}}\,|\,a,b)=f(x^{\{1\}}\,|\,a,b)\;f(x^{\{2\}}\,|\,a,b)\;\cdots\;f(x^{\{9\}}\,|\,a,b)$$

Prob vs Likelihood

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



log



$$\text{Max } L(a,b \mid x) \rightsquigarrow \text{Max } \operatorname{Log} L(a,b \mid x)$$

$$f(x^{\{i\}} \mid a,b) \cdots f(x^{\{N\}} \mid a,b) \longrightarrow \log(f(x^{\{i\}} \mid a,b) + \cdots + \log f(x^{\{N\}} \mid a,b)$$

$$\prod_{i=1}^{N} f(x^{\{i\}} \mid a,b) \qquad \sum_{i=1}^{N} \log f(x^{\{i\}} \mid a,b)$$

$$\text{Max } \ell(a,b \mid x) = \sum_{i=1}^{N} \log f(x^{\{i\}} \mid a,b)$$

$$\frac{\partial \ell(a,b \mid x)}{\partial a} = 0 \rightsquigarrow a = \sigma^2 \qquad \frac{\partial \ell(a,b \mid x)}{\partial b} = 0 \Rightarrow b = \mu$$

Multivariate Gaussian Distribution

σ^2

cov

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

- 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

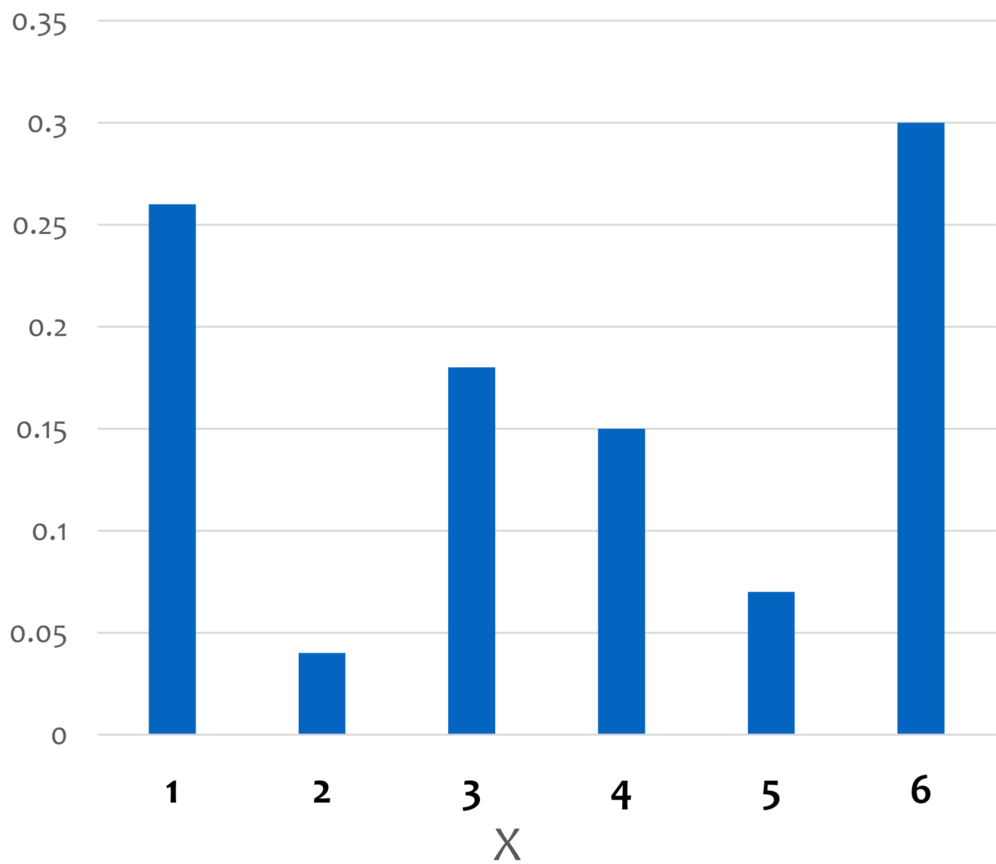
$$\mathrm{Y} = \mathrm{X}_1 + \mathrm{X}_2,\; \mathrm{X}_1 \perp X_2 \;\rightarrow \mu_y = \mu_1 + \mu_2, \Sigma_y = \Sigma_1 + \Sigma_2$$

- 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



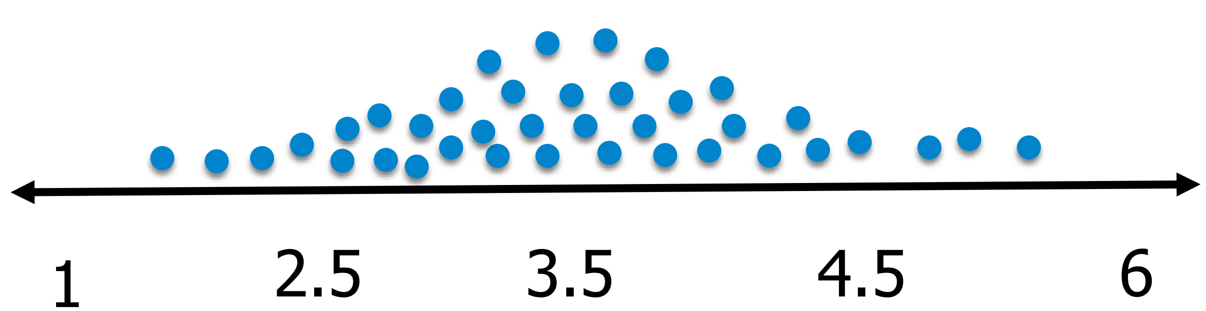
Let's say, I am going to get a sample from this pmf having a size of $\boldsymbol{n = 4}$

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

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

$$\vdots$$

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



According to CLT, it will follow a bell curve distribution (normal distribution)

Outline

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

Maximum Likelihood Estimation

M L E

- 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

$\max_{\theta} L(\theta|X) = \prod_{i=1}^{N} f(x^{\{i\}}|\theta)$

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

$\max_{\theta} \ell(\theta|X) = \sum_{i=1}^{N} \log f(x^{\{i\}}|\theta)$

Objective function: $\boxed{P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}}$ $\quad x_i \in \{0,1\}\ or\ \{head, tail\}$

$$L(\theta|X) = L(\theta|X = x_1, X = x_2, X = x_3, \ldots, X = x_n)$$

i.i.d assumption

$$\boxed{L(\theta|X) = \prod_{i=1}^{n} P(x_i|\theta)}$$

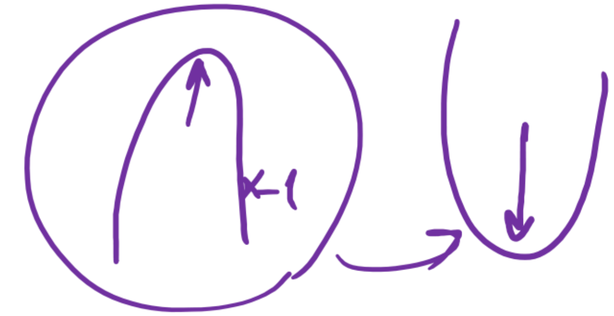
Objective function

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

$$L(\theta|X) = \theta^{x_1}(1-\theta)^{1-x_1} \times \theta^{x_2}(1-\theta)^{1-x_2} \ldots \times \theta^{x_n}(1-\theta)^{1-x_n} =$$
$$= \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}$$

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

What's the trick? Take the log



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

Max

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

How to optimize θ?

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

$X=1$

$X=0$

$70 \rightarrow H$

$30 \rightarrow T$

$$\Rightarrow \theta = \frac{1+1+1+\ldots+0+0}{100} = 0.7$$

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