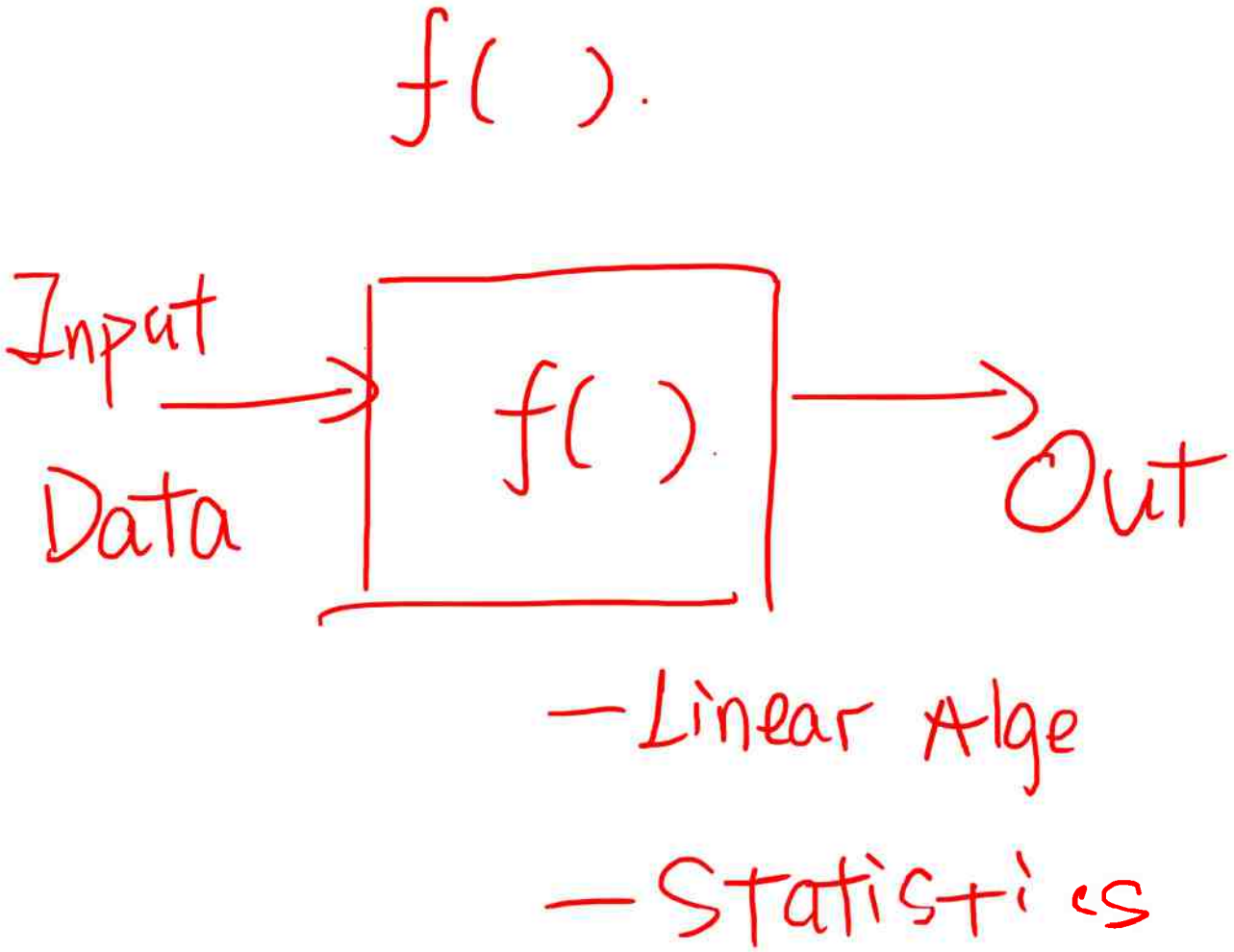

f( ).
Input
Data
f( )
Out
- Linear Alge
- Statistics

CSE 4820/5819
Warm-Up:
Linear Algebra
& Probability

Suining He
Department of Computer Science and Engineering
University of Connecticut
suining.he@uconn.edu

Outline

- Linear Algebra Warm-up
  - Matrix computation
- Statistics Warm-up
  - Probability theory

Scalars

regression

- A scalar is a single number
- Integers, real numbers, rational numbers, etc.
- We denote it with italic font:

$$a, n, x$$

Vectors

- A vector is a 1-D array of numbers:

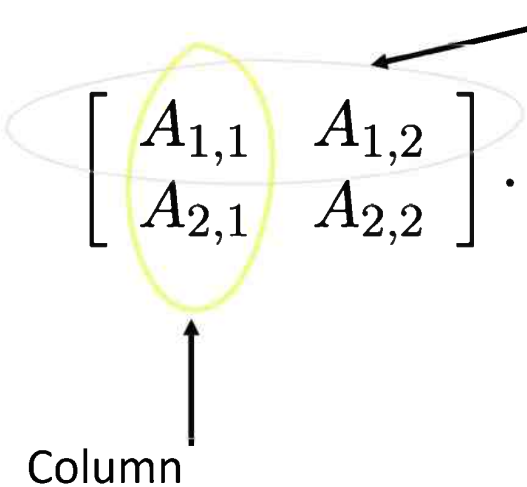
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

- Can be real, binary, integer, etc.
- Example notation for type and size:

$$\mathbb{R}^n$$

Matrices

- A matrix is a 2-D array of numbers:



- Example notation for type and shape:

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}$$

Tensors

- A tensor is an array of numbers, that may have
  - Zero dimensions, and be a scalar
  - One dimension, and be a vector
  - Two dimensions, and be a matrix
  - or more dimensions.

Vector & Matrix

- Vector in $\mathbb{R}^n$ is an ordered set of n real numbers.
  - e.g., $v = (1,6,3,4)$ is in R^4
  - A column vector: $\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$
  - A row vector: $\begin{pmatrix} 1 & 6 & 3 & 4 \end{pmatrix}$
- m-by-n matrix is an object in $R^{m \times n}$ with m rows and n columns, each entry filled with a (typically) real number:

$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$

Special Matrices

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$ diagonal

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$$ upper-triangular

$$\begin{pmatrix} a & b & 0 & 0 \\ c & d & e & 0 \\ 0 & f & g & h \\ 0 & 0 & i & j \end{pmatrix}$$ tri-diagonal

$$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}$$ lower-triangular

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ $\boldsymbol{I}$ (identity matrix)

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}.$$

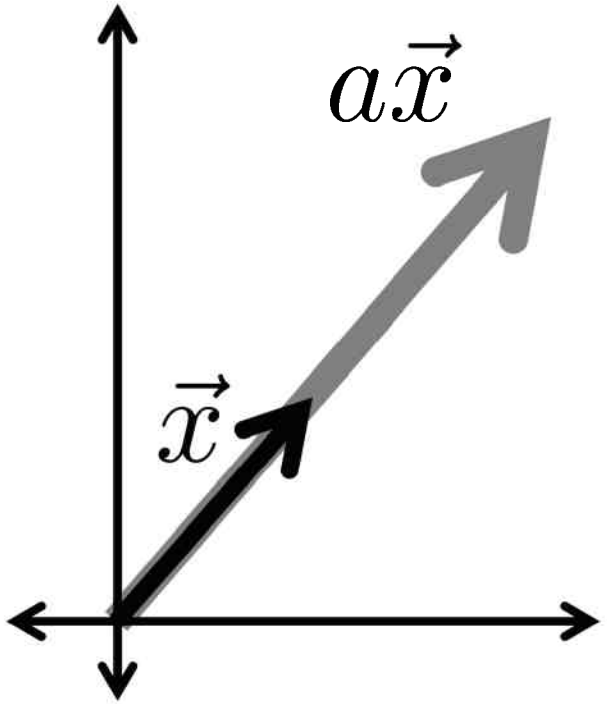
Vector & Matrix

- Matrix Addition

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

- Vector
  - Scalar times vector

$$a\vec{x} = a \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_N \end{pmatrix}$$



Matrix Transpose

$$(\boldsymbol{A}^\top)_{i,j} = A_{j,i}.$$

Transpose of a Matrix- examples

A	A^T
$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$
$[5]$	$[5]$
$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix}$	$\begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$
$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$	$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top.$$

Vector

- Element-by-element product (Hadamard product)

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} .* \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \end{pmatrix}$$

Vector

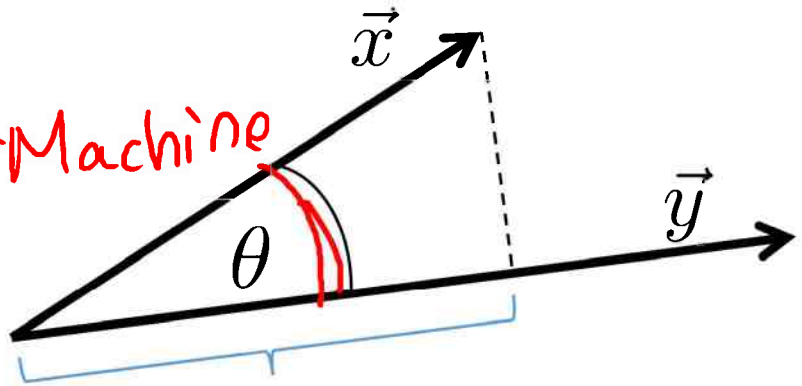
- Dot product (inner product)

Support Vector Machine

SVM

$$\vec{x} \cdot \vec{y} = (x_1 \quad x_2 \quad \cdots \quad x_N) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$$

$$= \sum_{i=1}^{N} x_i y_i$$



similarity. $\vec{x} \cdot \vec{y} = |\vec{x}||\vec{y}| \cos(\theta)$

$\cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|}$

$[-1, 1]$

Vector Norm

$\vec{x} = [x_1, x_2 \cdots x_n]$

- Functions that measure how "*large*" a vector is
- Similar to a distance between zero and the point represented by the vector
- Vector norms: A norm of a vector ||x|| is informally a measure of the "length" of the vector.

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

loss function → objective for ML.

- l_1 norm, l_2 norm, max norm ($p \to \infty$)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad ||\boldsymbol{x}||_\infty = \max_i |x_i|.$$

Examples

$$x = [1 \quad 2 \quad 3 \quad 4]$$

$$\|x\|_1 = \sum |x_i| = 1+2+3+4$$

$$\|x\|_2 = \sqrt{\sum (x_i)^2} = \sqrt{1^2+2^2+3^2+4^2}$$

$$\|x\|_\infty = \max\{x_i\} = 4$$

$$\vec{x} - \vec{y}$$

Matrix Times a Vector

$$\vec{y} = \overleftrightarrow{W}\vec{x}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1N} \\ W_{21} & W_{22} & \cdots & W_{2N} \\ \vdots & \vdots & & \vdots \\ W_{M1} & W_{M2} & \cdots & W_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

$M \times 1$ $\quad$ $M \times N$ $\quad$ $N \times 1$

Examples

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad W \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$y_1 = w_{11} \cdot x_1 + w_{12} \cdot x_2 .$$

Matrix Product

- We will use upper case letters for matrices. The elements are referred by A_{ij}.
  - Matrix product:

$$A \in \mathbb{R}^{m \times n} \qquad B \in \mathbb{R}^{n \times p}$$

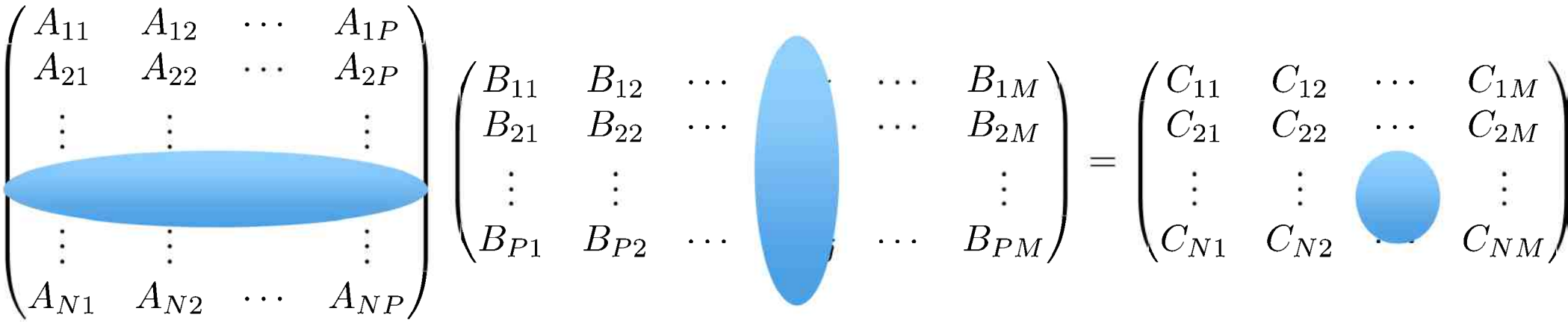
$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Matrix Times Matrix

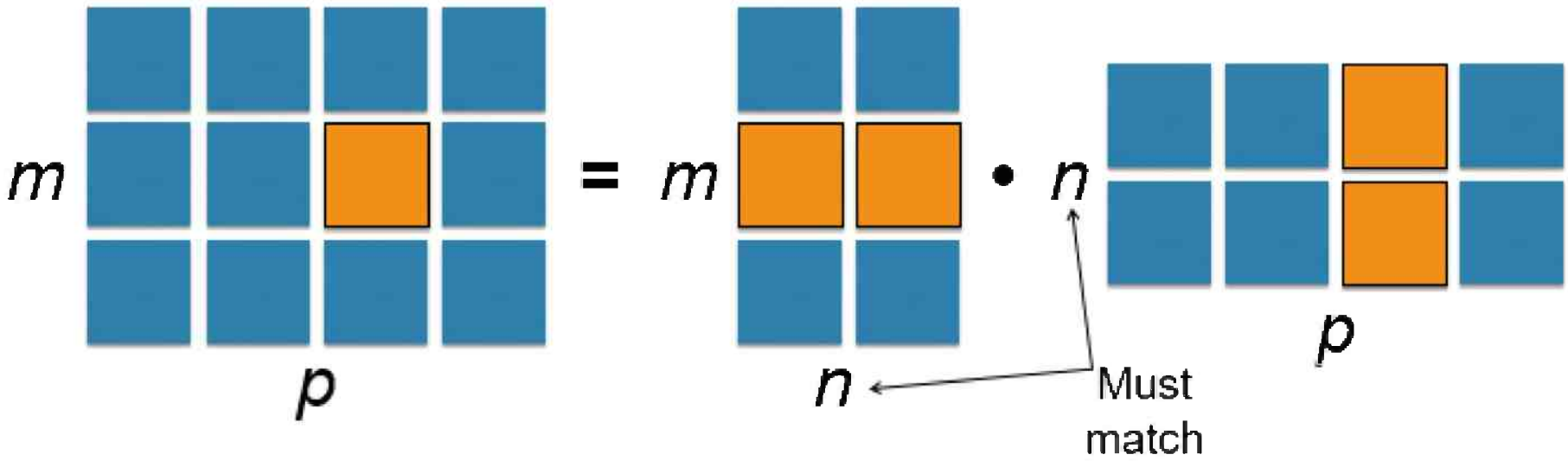


- C_{ij} is the inner product of the i-th row with the j-th column

Matrix (Dot) Product

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}.$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}.$$



Systems of Equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

expands to

$$\boldsymbol{A}_{1,:}\boldsymbol{x} = b_1$$
$$\boldsymbol{A}_{2,:}\boldsymbol{x} = b_2$$
$$\cdots$$
$$\boldsymbol{A}_{m,:}\boldsymbol{x} = b_m$$

A linear system of equations can have:

- No solution
- Many solutions
- Exactly one solution: this means multiplication by the matrix is an invertible function

Inverse of a Matrix

- Inverse of a square matrix A, denoted by A^{-1} is the *unique* matrix s.t.
  - $AA^{-1} = A^{-1}A = \boldsymbol{I}$ (identity matrix)
- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$,
  - $(A^T)^{-1} = (A^{-1})^T$
- For diagonal matrices $\mathbf{D}^{-1} = \text{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

ad−bc: determinant

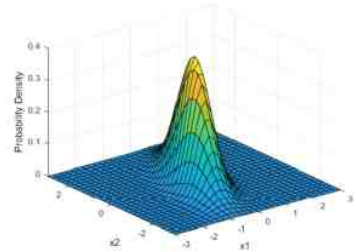
$$\begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}^{-1} = \frac{1}{4\text{x}6-7\text{x}2} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix}$$

$$= \frac{1}{10} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

Invertibility

- Matrix can't be inverted if…
  - More rows than columns
  - More columns than rows
  - Redundant rows/columns ("linearly dependent", "low rank")
- The matrix is invertible if and only if its determinant is not zero.
  - Any square matrix which contains a complete row or a complete column filled with zeros, cannot be inverted



Probability Theory for ML

- *Convenience*: declaring all conditions, exceptions, assumptions would be too complicated.
  - Example: "I will be in lecture if I go to bed early enough the day before and I do not become ill and my car does not have a breakdown and ..."
  - Or simply: I will be in lecture with probability of 0.87
- *Lack of information*: relevant information is missing for a precise statement.
  - Example: weather forecasting
- *Intrinsic randomness*: non-deterministic processes.
  - Example: appearance of photons in a physical process
- Intuitively, probabilities give the expected relative frequency of an event

Sample Space

- The sample space Ω is the set of possible outcomes of an experiment.
  - Points ω in Ω are called sample outcomes, realizations, or elements.
- Subsets of Ω are called Events.
  - Example: If we toss a coin twice then Ω = {HH,HT, TH, TT}. The event that the first toss is head is A = {HH,HT}
- We say that events A_i and A_j are disjoint (mutually exclusive) if $A_i \cap A_j$ = {}
  - Example: first flip being heads and first flip being tails

Probability

- We will assign a real number $P(A)$ to every event A, called the probability of A.
- To qualify as a probability, P must satisfy three axioms:
  - Axiom 1: $P(A) \geq 0$ for every A
  - Axiom 2: $P(\Omega) = 1$
  - Axiom 3: If $A_1, A_2, \ldots$ are disjoint then

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

Conditional Probabilities

- ***Conditionals:***
  - Example: if someone is taking a shower, he gets wet (by causality) -- P(“wet” | “taking a shower”) = 1
  - while: P(“taking a shower” | “wet”) = 0.4 because a person also gets wet if it is raining
- ***Causality and conditionals:***
  - Causality typically causes conditional probabilities close to 1: P(“wet” | “taking a shower”) = 1,
  - e.g. P(“score a goal” | “shoot strong”) = 0.92 (’vague causality’: if you shoot strong, you very likely score a goal’).
  - Offers the possibility to express vagueness in reasoning.

Conditional Probabilities

- You cannot conclude ***causality*** from large ***conditional*** probabilities:
  - P(“being rich” | “owning an airplane”) ≈ 1 but: owning an airplane is not the reason for being rich

Joint Events

- For pairs of events A, B, the joint probability expresses the probability of both events occurring at same time: $P(A, B)$
  - Example: P(“*Lakers* is losing”, “*Magic* is winning”) = 0.3
- For two events the conditional probability of $A \mid B$ is defined as the probability of event A if we consider only cases in which event B occurs:

$$P(A|B) = \frac{P(A,B)}{P(B)}, P(B) \neq 0$$

- With the above, we also have

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

example: $P(\text{“caries”}|\text{“toothaches”}) = 0.8$

$P(\text{“toothaches”}|\text{“caries”}) = 0.3$

Product Rule & Chain Rule

- From definition of conditional probability

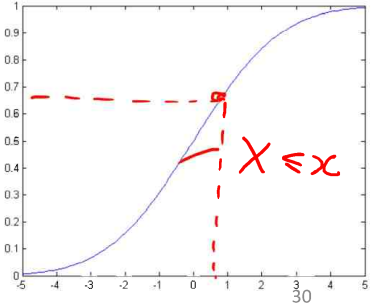
$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Repeated application: *chain rule*

$$\begin{aligned} & P(A_1, \ldots, A_n) = P(A_n, \ldots, A_1) \\ = \; & P(A_n|A_{n-1}, \ldots, A_1)\, P(A_{n-1}, \ldots, A_1) \\ = \; & P(A_n|A_{n-1}, \ldots, A_1)\, P(A_{n-1}|A_{n-2}, \ldots, A_1)\, P(A_{n-2}, \ldots, A_1) \\ = \; & \ldots \\ = \; & \Pi_{i=1}^{n} P(A_i|A_1, \ldots, A_{i-1}) \end{aligned}$$

Random Variables & Distributions

- Random variable:
  - A corresponding probability distribution → maps the values of random variables to their probabilities of occurrence (often denoted as X)
- Cumulative distribution function (CDF)
  - A function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as $F_X(x) = P(X \leq x)$
  - Basic properties
    - $0 \leq F_X(x) \leq 1.$
    - $\lim_{x \to -\infty} F_X(x) = 0.$
    - $\lim_{x \to \infty} F_X(x) = 1.$
    - $x \leq y \implies F_X(x) \leq F_X(y).$

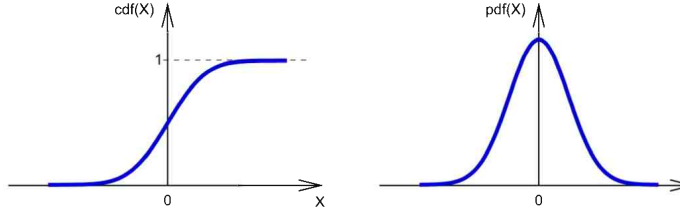


Random Variables & Distributions

- Probability density function (PDF)
  - Derivative of the function F_X
  - Gaussian/Normal distribution is an important probability distribution with the pdf of

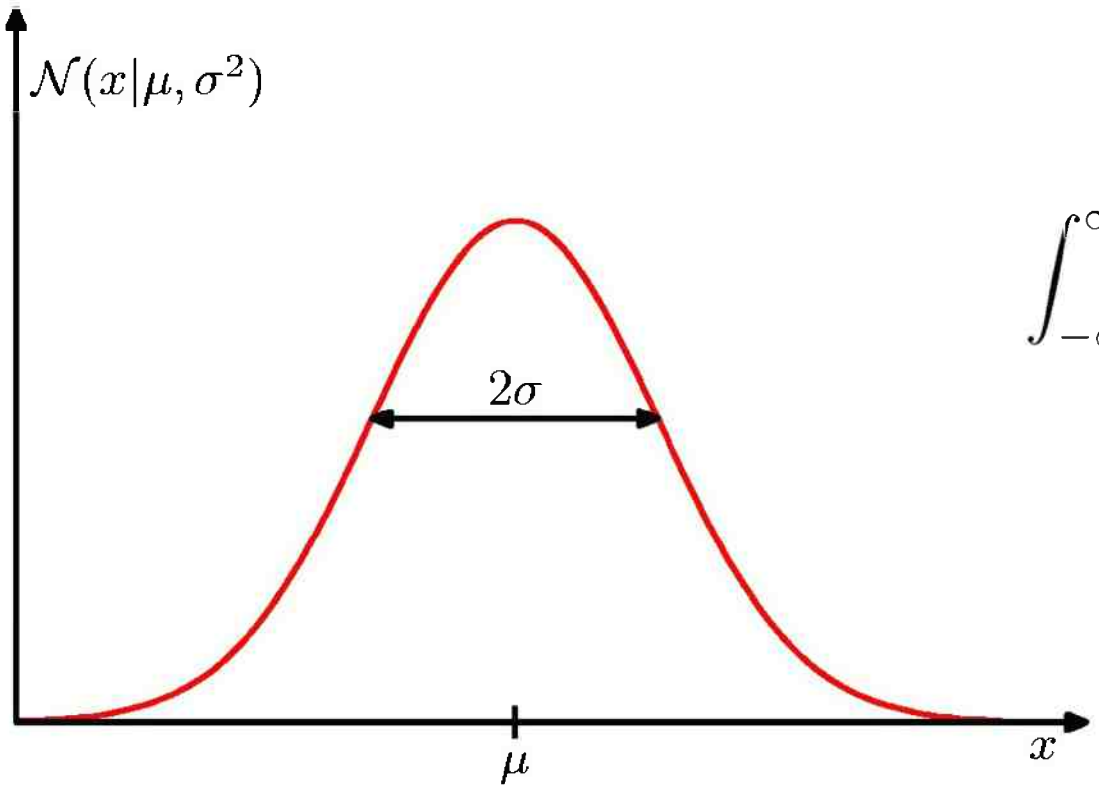
$$pdf(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

  - $\mu \in R$ and $\sigma^2 > 0$ are the parameters of the



Gaussian Distribution

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Mean and Variance

- The mean of a random variable X is the average value X takes.

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

- The variance of X is a measure of how dispersed the values that X takes are.

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- The standard deviation is simply the square root of the variance.
- Linear property $E[X+Y] = E[X] + E[Y] \qquad E[aX] = aE[X]$

Simple Example

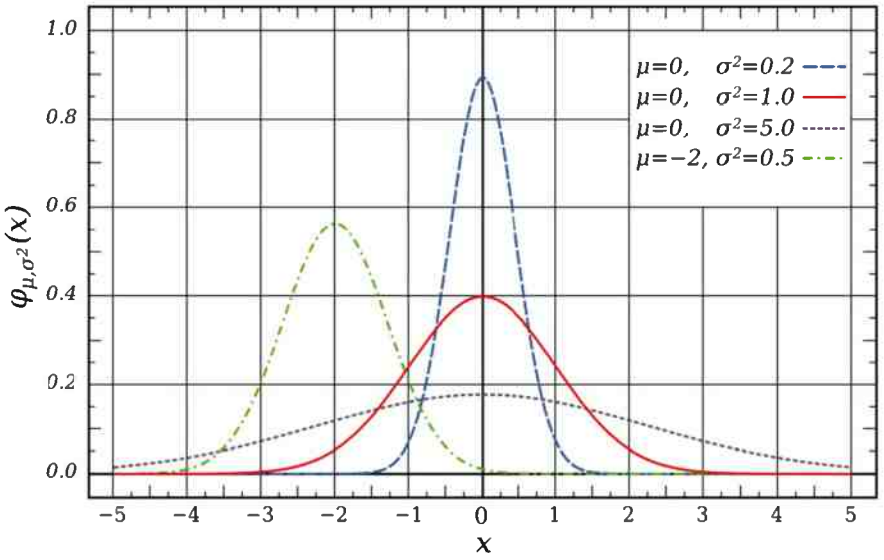
- X= {1, 2} with P(X=1) = 0.8 and P(X=2) = 0.2
- Mean
  - $0.8 \times 1 + 0.2 \times 2 = 1.2$
- Variance
  - $0.8 \times (1 - 1.2) \times (1 - 1.2) + 0.2 \times (2 - 1.2) \times (2 - 1.2)$

Gaussian Distribution

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$



Parametric Approach

- Assumption: data distribution follows a parametric model, e.g., Gaussian, Bernoulli
  - The model is fully specified by a small number of parameters θ
- Parameter estimation:
  - Assuming some parametric form for $p(x \mid \theta)$, θ is estimated using X
  - Statistical inference

Statistical Inference

- Determining the probability distribution of a random variable (*estimation*)
- Collecting outcome of repeated random experiments (*data sample*)
- Adapt a generic probability distribution to the data.
  - Bernoulli-distribution (possible outcomes: 1 or 0) with success parameter p (=probability of outcome "1")
  - Gaussian distribution with parameters μ and σ
  - Uniform distribution with parameters a and b
- Maximum-likelihood estimation (MLE) approach:

$$\underset{\text{parameters}}{\arg\max}\, P(\text{data samples}|\text{distribution})$$

Likelihood

- What does likelihood mean and how is "likelihood" different than "probability"?
  - *Discrete distributions*: likelihood is a synonym for the joint probability of your data.
  - *Continuous distribution*: likelihood refers to the joint probability density of your data.
- Since we assumed that each data point is independent, the likelihood of all of our data is the product of the likelihood of each data point.
  - Mathematically, the likelihood of our data give parameters θ is:

$$L(\theta|\boldsymbol{X}) = p(\boldsymbol{X}|\theta) = \prod_i p(x_i|\theta)$$

Maximum Likelihood Estimation

- MLE seeks to find θ that makes sampling x_i from $p(x \mid \theta)$ as likely as possible by maximizing the likelihood of θ given the sample $\boldsymbol{X} = \{x_i\}$
- Likelihood of θ given $\boldsymbol{X}$:
  - $L(\theta|\boldsymbol{X}) = p(\boldsymbol{X}|\theta) = \prod_i p(x_i|\theta)$
- Log likelihood (computational simplification)

$$\mathcal{L}(\theta|\mathbf{X}) = \log L(\theta|\mathbf{X}) = \sum_i \log p(x_i|\theta)$$

- Therefore, MLE finds:
  - $\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{X})$

Some Useful Derivatives

$f(x)$	$f'(x)$
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\tan x$	$\sec^2 x$
$\cot x$	$-\csc^2 x$
$\sec x$	$\sec x \tan x$
$\csc x$	$-\csc x \cot x$

$f(x)$	$f'(x)$
e^x	e^x
a^x	$a^x \ln a$
$\ln x$	$\frac{1}{x}$
$\log_a x$	$\frac{1}{x \ln a}$

$$(f' + g') = f' + g'$$

$$(c)' = 0 \qquad c \text{ a constant}$$

$$(c\,f)' = c\,f' \qquad c \text{ a constant}$$

$$(x^n)' = n\,x^{n-1} \qquad n \text{ a positive integer}$$

$$(f\,g)' = f'\,g + f\,g'$$

$$\left(\frac{f}{g}\right)' = \frac{(f'\,g - f\,g')}{(g)^2}$$

$$(f\,(g(x)))' = g'(x) f'(g) \qquad \textbf{the Chain Rule.}$$

Example: Bernoulli Distribution

- **MLE with Bernoulli-distribution:**
- ***Assumptions***: coin toss with a twisted coin. How likely is it to observe head?
- Repeat several experiments, to get a sample of observations, e.g.:
  - "head", "head", "number", "head", "number", "head", "head", "head", "number", "number", ...
  - You observe k times "head" and n times "number"
- Probabilistic model: "head" occurs with (unknown) probability p, "number" with probability $1 - p$
  - We basically find the parameters p to maximize the likelihood,

$$\arg\max_{\mathrm{p}} \quad p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot \ldots$$

$$\arg\max_{\mathrm{p}} \quad p^k (1-p)^n$$

Example: Bernoulli Distribution

- Probabilistic model: "head" occurs with (unknown) probability p, "number" with probability $1 - p$
  - We basically find the parameters p to maximize the likelihood,

$$\arg\max_{p} \quad p^k(1-p)^n$$

$L = p^k(1-p)^n$

$\ell = \log L = \log p^k + \log (1-p)^n$

$= k \log p + n \cdot \log(1-p)$

$\frac{\partial \ell}{\partial p} = \frac{k}{p} + (-1) \cdot \frac{n}{1-p} = \frac{k(1-p) + (-1) np}{p(1-p)}$

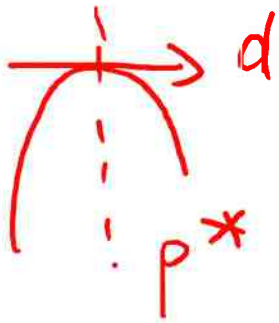
$= 0 \Rightarrow k(1-p) - np = 0$

$k - kp - np = 0$

$\Rightarrow k - p(k+n) = 0$

$\Rightarrow p = \frac{k}{k+n}$

convex, $\frac{\partial \ell}{\partial p} = 0 \Rightarrow$ derivative $= 0$



Example: Bernoulli Distribution

Example: Bernoulli Distribution

Example: Gaussian Distribution

- Suppose the weights of randomly selected students are normally distributed with unknown mean and standard deviation . A random sample of 10 students yielded the following weights (in pounds):
- 115 122 130 127 149 160 152 138 149 180
- Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ the mean weight of all students. Using the given sample, find a maximum likelihood estimate of μ as well.
- How about σ?

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

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

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

Example: Gaussian Distribution

$$\mathcal{L} = \log L = \log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{N} \log\left(\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0. \quad = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{N} (-1) \cdot \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 + (-1)(-1) \cdot \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} \cdot \frac{1}{2} = 0.$$

$$\sum_{i=1}^{N} (x_i - \mu) = 0. \Rightarrow \sum_{i=1}^{N} x_i - N\mu = 0.$$

$\hat{\mu}$ sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Example: Gaussian Distribution

$$\frac{\partial L}{\partial \theta} = 0.$$

Example: Gaussian Distribution

Maximum Likelihood Estimation

- MLE seeks to find θ that makes sampling x_i from $p(x \mid \theta)$ as likely as possible by maximizing the likelihood of θ given the sample $\boldsymbol{X} = \{x_i\}$
- Likelihood of θ given $\boldsymbol{X}$:
  - $L(\theta|\boldsymbol{X}) = p(\boldsymbol{X}|\theta) = \prod_i p(x_i|\theta)$
- Log likelihood (computational simplification)

$$\mathcal{L}(\theta|\mathbf{X}) = \log L(\theta|\mathbf{X}) = \sum_i \log p(x_i|\theta)$$

- Therefore, MLE finds:
  - $\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{X})$

Comparison between Probability and Machine Learning (ML)

	Machine Learning	Statistics
Unsupervised Learning	Create a model of the observed patterns	Estimating the probability distribution P(patterns)
Classification	Guessing the class from an input pattern	Estimating P(class|input pattern)
Regression	Predicting the output from the input pattern	Estimating P(output|input pattern)

- Probabilities allow to precisely describe the relationships in a certain domain, e.g. distribution of the input data, distribution of outputs conditioned on inputs, ...
- ML principles like minimizing squared error can be interpreted in a stochastic sense

Summary

- Linear Algebra Warm-up
- Statistics Warm-up