CPEN 355 – Lecture 2: Review of Mathematical Foundations

Mirza Sarwar, Ph.D.

Department of Electrical and Computer Engineering

University of British Columbia

Outline

- Linear Algebra basics
- Probability basics
- Supervised
  - Linear regression
  - Classification (logistic regression)
  - Model evaluation
  - Naïve Bayes classifier
  - SVM
  - Trees, boosting and bagging
  - Random forest
  - KNN
- Unsupervised
  - Clustering
  - VAE (if time permits)
- Dimension reduction
  - PCA
- Neural Network
- State of the Art

Linear Algebra basics

Vector Operations

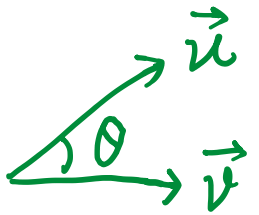
$\vec{u} = [u_1, u_2]^T \quad \vec{v} = [v_1, v_2]^T$

- Addition: $\vec{u} + \vec{v} = [u_1 + v_1, \; u_2 + v_2]^T$
- Scalar Product: $a\vec{u} = [au_1, au_2]^T$
- Dot (Inner) Product:

$$\vec{u} \cdot \vec{v} = \vec{u}^T \vec{v} = [u_1, u_2] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = u_1 v_1 + u_2 v_2$$

$$\|\vec{u}\|^2 = \vec{u}^T \vec{u} = u_1^2 + u_2^2$$

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \, \|\vec{v}\| \cos\theta$$



Matrix Operations

- Addition

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

- Subtraction

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a-e & b-f \\ c-g & d-h \end{bmatrix}$$

- Multiplication

Matrix-matrix product

Hadamard product
Element-wise product

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Matrix – Vector product

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

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

- We use the notation a_{ij} (or A_{ij}, $A_{i,j}$, etc) to denote the entry of A in the ith row and jth column:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

- We denote the jth column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

- We denote the ith row of A by a_i^T or $A_{i,:}$:

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}.$$

https://cs229.stanford.edu/section/cs229-linalg.pdf

Matrix – Vector product

I

If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

II

Alternatively, let's write A in column form. In this case we see that,

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a_1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a_2 \\ \\ \end{bmatrix} x_2 + \ldots + \begin{bmatrix} \\ a_n \\ \\ \end{bmatrix} x_n \ .$$

In other words, y is a ***linear combination*** of the *columns* of A, where the coefficients of the linear combination are given by the entries of x.

Important to understand attention used in transformers

https://cs229.stanford.edu/section/cs229-linalg.pdf

Matrix Operating on Vectors

- Matrix is like a function that transforms the vectors on a plane
- In system of linear equation, matrix holds the coefficients

$$\begin{cases} x' = ax + by \\ y' = cx + dy \end{cases} \Rightarrow \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

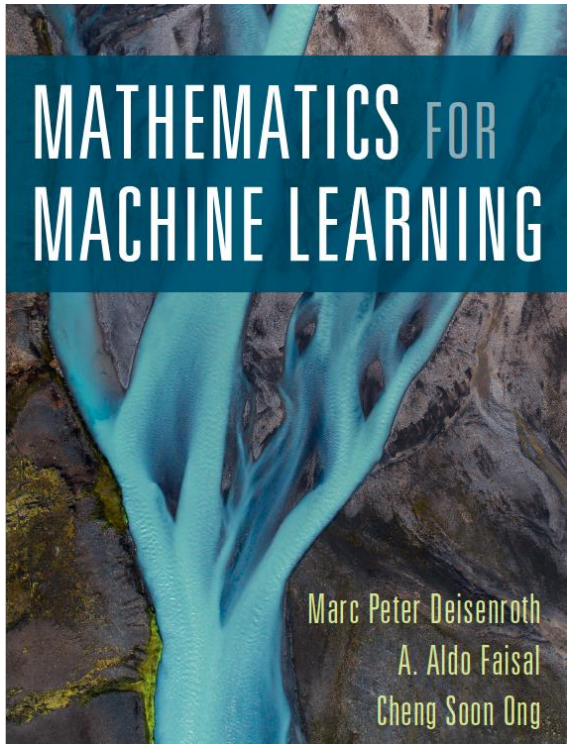
Systems of Linear Equations

An Example

A company produces products $N_1, \ldots, N_n$ for which resources $R_1, \ldots, R_m$ are required. To produce a unit of product N_j, a_{ij} units of resource R_i are needed, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The objective is to find an optimal production plan, i.e., a plan of how many units x_j of product N_j should be produced if a total of b_i units of resource R_i are available and (ideally) no resources are left over.

If we produce $x_1, \ldots, x_n$ units of the corresponding products, we need



An Example

A company produces products $N_1, \ldots, N_n$ for which resources $R_1, \ldots, R_m$ are required. To produce a unit of product N_j, a_{ij} units of resource R_i are needed, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The objective is to find an optimal production plan, i.e., a plan of how many units x_j of product N_j should be produced if a total of b_i units of resource R_i are available and (ideally) no resources are left over.

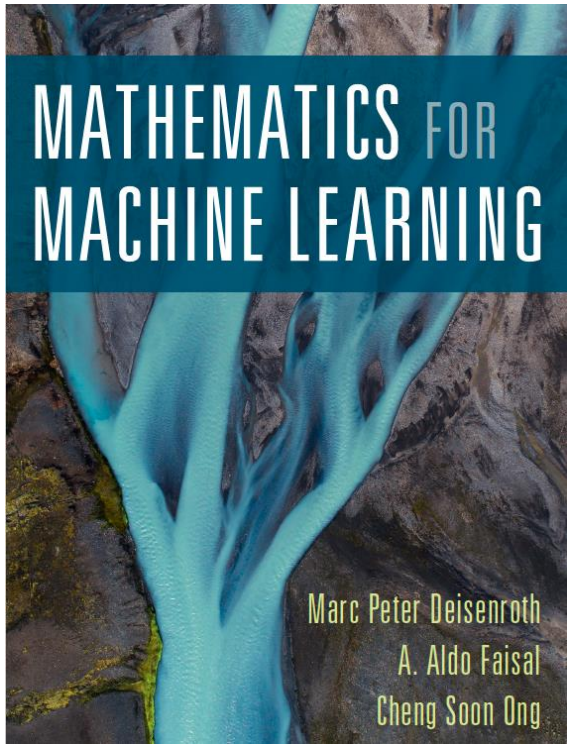
If we produce $x_1, \ldots, x_n$ units of the corresponding products, we need a total of

$$a_{i1}x_1 + \cdots + a_{in}x_n \tag{2.2}$$

many units of resource R_i. An optimal production plan $(x_1, \ldots, x_n) \in \mathbb{R}^n$, therefore, has to satisfy the following system of equations:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}, \tag{2.3}$$

where $a_{ij} \in \mathbb{R}$ and $b_i \in \mathbb{R}$.



https://mml-book.github.io/book/mml-book.pdf

An Example

A company produces products $N_1, \ldots, N_n$ for which resources $R_1, \ldots, R_m$ are required. To produce a unit of product N_j, a_{ij} units of resource R_i are needed, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The objective is to find an optimal production plan, i.e., a plan of how many units x_j of product N_j should be produced if a total of b_i units of resource R_i are available and (ideally) no resources are left over.

If we produce $x_1, \ldots, x_n$ units of the corresponding products, we need a total of

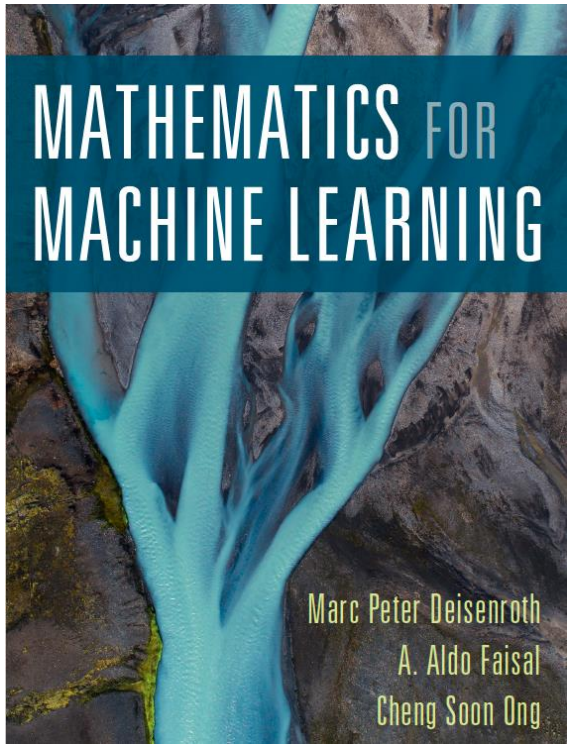
$$a_{i1}x_1 + \cdots + a_{in}x_n \tag{2.2}$$

many units of resource R_i. An optimal production plan $(x_1, \ldots, x_n) \in \mathbb{R}^n$, therefore, has to satisfy the following system of equations:

$$\begin{matrix} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{matrix}, \tag{2.3}$$

where $a_{ij} \in \mathbb{R}$ and $b_i \in \mathbb{R}$.

general form of a system of linear equations



Solving Systems of Linear Equations

For a systematic approach to solving systems of linear equations, we will introduce a useful compact notation. We collect the coefficients a_{ij} into vectors and collect the vectors into matrices. In other words, we write the system from (2.3) in the following form:

$$\begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} x_2 + \cdots + \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} x_n = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Row vector

Column vector

$$\Longleftarrow \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

Matrix-vector product
= a transformation of a vector

$$\boldsymbol{Ax} = \boldsymbol{b} \iff \boldsymbol{A}^\top \boldsymbol{Ax} = \boldsymbol{A}^\top \boldsymbol{b} \iff \boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{b}$$

Vector Spaces

Vector space (informal): a structured space in which vectors live

Definition 2.9 (Vector Space)**.** A real-valued *vector space* $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$$
$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$$

Example:

$\mathcal{V} = \mathbb{R}^n, n \in \mathbb{N}$ is a vector space with operations defined as follows:

- Addition: $\boldsymbol{x} + \boldsymbol{y} = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$
- Multiplication by scalars: $\lambda \boldsymbol{x} = \lambda(x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n)$ for all $\lambda \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^n$

Linear Independence and Rank

- A set of vectors $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^m$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors.

- If a vector can be represented as a linear combination of the remaining vectors, namely

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for scaler values $\alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$.

Linear Independence and Rank

- The **column rank** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a **linear independent** set.

- In the same way, the **row rank** is the largest number of rows of A that constitute a linearly independent set.

- Rank(A) = column rank = row rank

Basic properties of the rank

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be ***full rank***.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

Span, Range and Nullspace of a Matrix

- The **span** of a set of vectors $\{x_1, x_2, \dots, x_n\}$ is the set of all vectors that can expressed as a linear combination of $\{x_1, x_2, \dots, x_n\}$

$$\text{span}(\{x_1, \dots x_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}$$

- If $\{x_1, x_2, \dots, x_n\}$ are linearly independent, where each $x_i \in \mathbb{R}^n$, then span ($\{x_1, x_2, \dots, x_n\}$) = $\mathbb{R}^n$. Namely, any vector v $\in \mathbb{R}^n$ can be written as a linear combination of x_1 through x_n.

https://www.youtube.com/watch?v=k7RM-ot2NWY

Span, Range and Nullspace of a Matrix

- The **range** of matrix $A \in \mathbb{R}^{m \times n}$, denoted $R(A)$, is the span of the columns of A. In other words:

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}.$$

- The nullspace of a matrix $A \in \mathbb{R}^{m \times n}$ (n<m), denote $N(A)$ is the set of all vectors that equal 0 when multiple by A, i.e.,

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Null Space

The **null space** of an $m \times n$ matrix A, written as Nul A, is the set of all solutions to the homogeneous equation $A\mathbf{x} = \mathbf{0}$.

$$\text{Nul } A = \{\mathbf{x} : \mathbf{x} \text{ is in } \mathbf{R}^n \text{ and } A\mathbf{x} = \mathbf{0}\} \qquad \text{(set notation)}$$

Orthogonal and orthonormal

Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

$$U^T U = I = U U^T.$$

Also, $U^{-1}U = I$

So, $U^{-1} = U^T$ For orthogonal matrix

https://www.cs.unc.edu/~ronisen/teaching/spring_2023/web_materials/lecture2_maths.pdf

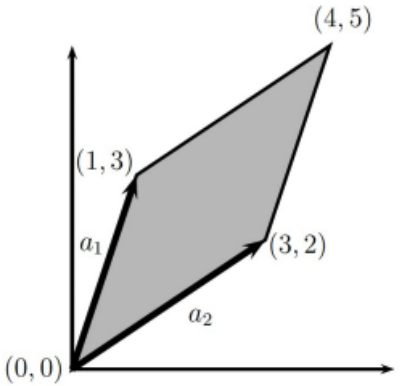
Determinant

How do we find determinant of a nxn matrix?
[See Leibniz formula for determinants]

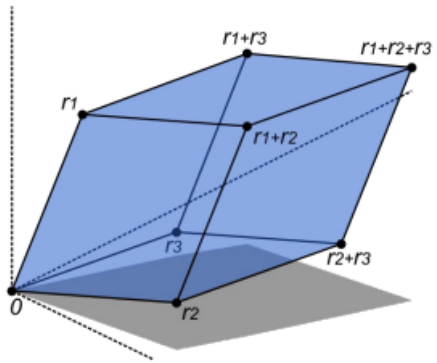
$$|[a_{11}]| = a_{11}$$

$$\left\| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \right| = a_{11}a_{22} - a_{12}a_{21}$$

$$\left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right| = \begin{aligned} & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ & -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned}$$



- Determinant of 2x2 matrix is the area of the parallelogram formed by the column vectors of the matrix.
- Determinant of 3x3 matrix is the volume of a parallelopiped formed by the 3 column vectors of the matrix
- Sign indicates whether the transformation preserves or reverse orientation.



https://www.cs.unc.edu/~ronisen/teaching/spring_2023/web_materials/lecture2_maths.pdf

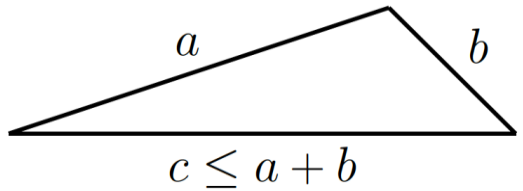
Norms

Definition 3.1 (Norm). A *norm* on a vector space V is a function

$$\begin{aligned} \|\cdot\| : V &\to \mathbb{R}\,, \\ \boldsymbol{x} &\mapsto \|\boldsymbol{x}\|\,, \end{aligned}$$

which assigns each vector $\boldsymbol{x}$ its *length* $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- *Absolutely homogeneous:* $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
- *Triangle inequality:* $\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- *Positive definite:* $\|\boldsymbol{x}\| \geqslant 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$



L1 and L2 Norms

The *Manhattan norm* on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|,$$

The *Euclidean norm* of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

Other Norms

$$x = [x_1, x_2, \ldots, x_D] \in \mathbb{R}^D$$

- l_∞ norm: $\|x\|_\infty = \max_i |x_i|$
- l_p norm: $\|x\|_p = \left(\sum_{i=1}^{D} |x_i|^p\right)^{1/p}$
- Frobenius norm: $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{tr(A^T A)}$

$$tr(A) = \sum_{i=1}^{n} A_{ii}$$

Eigen values and Eigen vectors

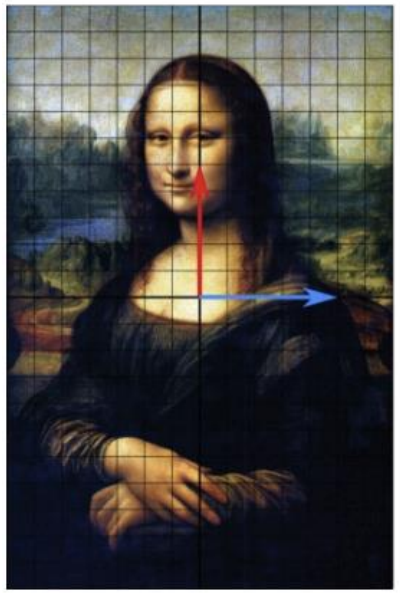
Eigen values and eigen vectors

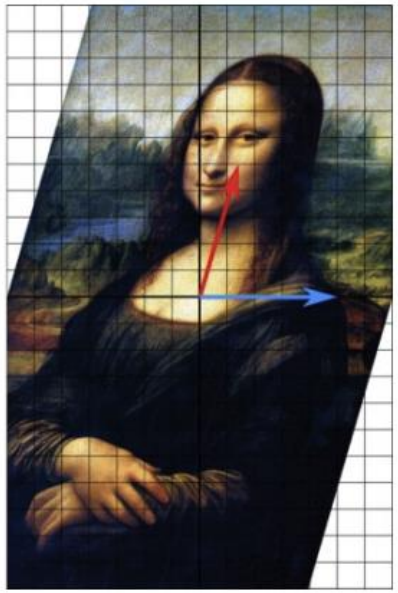
Given a square matrix $A \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector*** if

$$Ax = \lambda x, \quad x \neq 0.$$

$$\begin{pmatrix} I & M \\ 0 & I \end{pmatrix}.$$

Intuition :
Vector Ax is parallel to x.
x is unchanged by A other than a scale





In this shear mapping the red arrow changes direction, but the blue arrow does not. The blue arrow is an eigenvector of this shear mapping because it does not change direction, and since its length is unchanged, its eigenvalue is 1.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x + my \\ y \end{pmatrix} = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

[a,0] is an eigenvector for any value of a.

https://www.cs.unc.edu/~ronisen/teaching/spring_2023/web_materials/lecture2_maths.pdf

https://www.youtube.com/watch?v=cdZnhQjJu4I

https://math.mit.edu/~gs/linearalgebra/ila5/linearalgebra5_6-1.pdf

Solve for eigen values

$Ax = \lambda x$

For non-zero x

$(A-I\lambda)x = 0$

$(A-I\lambda)$ is singular

Why?

$Det(A-I\lambda) = 0$

Suppose $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

Compute $A - \lambda I = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}$.

Calculate the determinant:

$$\det(A - \lambda I) = (2-\lambda)(2-\lambda) - 1 \cdot 1$$
$$= \lambda^2 - 4\lambda + 3$$

Solve: $\lambda^2 - 4\lambda + 3 = 0$.

$$(\lambda - 1)(\lambda - 3) = 0$$

Eigenvalues: $\lambda = 1, 3$

Eigen-Decomposition

$$\mathbf{Au} = \lambda \mathbf{u} . \quad (1)$$

when rewritten, the equation becomes:

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{u} = \mathbf{0} , \quad (2)$$

where λ is a scalar called the *eigenvalue* associated to the *eigenvector*.

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \quad (3)$$

has the eigenvectors:

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \text{with eigenvalue} \quad \lambda_1 = 4 \quad (4)$$

and

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{with eigenvalue} \quad \lambda_2 = -1 \quad (5)$$

Traditionally, we put together the set of eigenvectors of **A** in a matrix denoted **U**. Each column of **U** is an eigenvector of **A**. The eigenvalues are stored in a diagonal matrix (denoted **Λ**), where the diagonal elements gives the eigenvalues (and all the other values are zeros). We can rewrite the first equation as:

$$\mathbf{AU} = \mathbf{U\Lambda} ;$$

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1} .$$

$$\mathbf{A} = \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} .2 & .2 \\ -.4 & .6 \end{bmatrix}$$

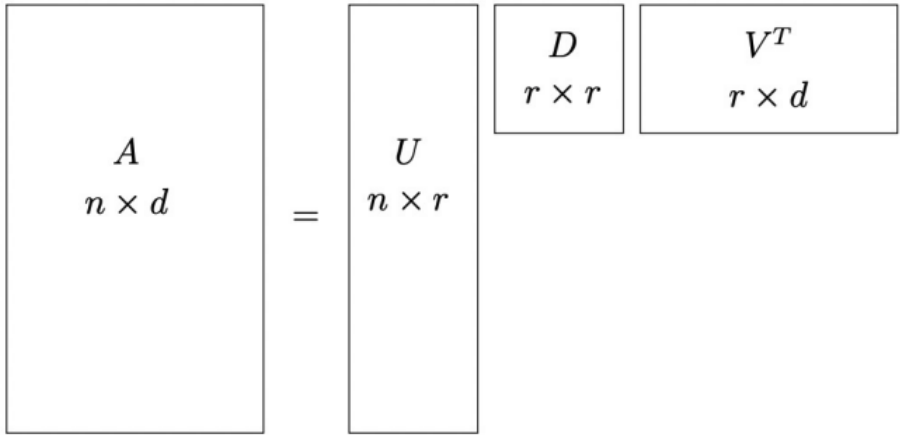
$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} .$$

Can you calculate this yourself?

Singular Value Decomposition

General form of ED

$$A = UDV^T \qquad A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$



U and V are orthonormal matrices, i.e. $U^TU=I$ and $V^TV=I$

D is a diagonal matrix, where each diagonal element is known as singular values. $D_{ii}=\sigma_i$

r is the rank of the matrix
r <= min (n,d)

http://www.juyang.co/everything-you-need-to-know-about-matrix-in-machine-learning-ii-eigendecomposition-and-singular-value-decomposition/

https://www.cs.unc.edu/~ronisen/teaching/spring_2023/web_materials/lecture2_maths.pdf

ED vs SVD

$A = P.D.P^{-1}$ $A = U.D.V^{T}$

- The vectors in the eigen-decomposition matrix P are not necessarily orthogonal, so the change of basis isn't a simple rotation. On the other hand, the vectors in the matrices U and V in the SVD are orthonormal, so they do represent rotations (and possibly flips).

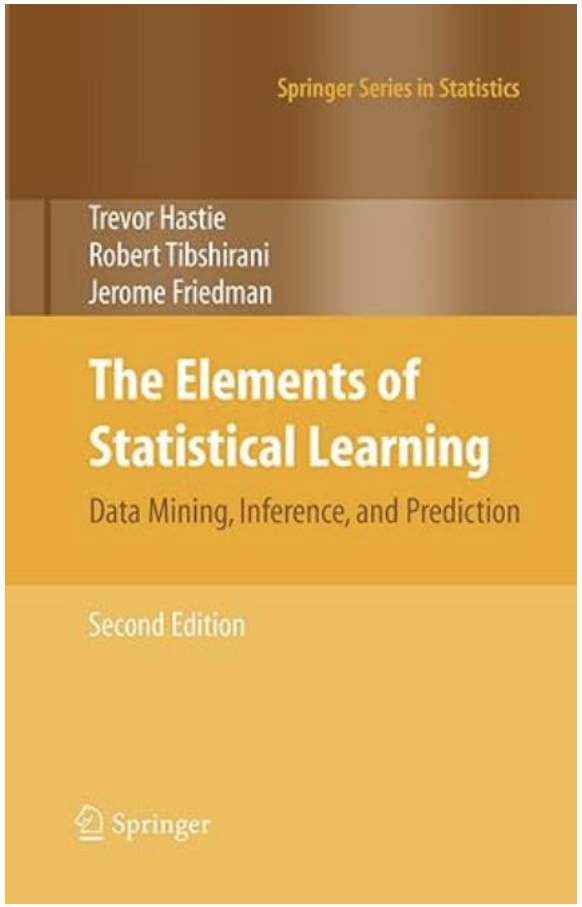
- In the SVD, the nondiagonal matrices U and V are not necessarily the inverse of one another. They are usually not related to each other at all. In the eigen decomposition the nondiagonal matrices P and P^{-1} are inverses of each other.

- The SVD always exists for any sort of rectangular or square matrix, whereas the eigen decomposition can only exists for square matrices, and even among square matrices sometimes it doesn't exist (eigen vectors need to be linearly independent).

Resources

https://www.cs.ubc.ca/~schmidtm/Documents/2009_Notes_LinearAlgebra.pdf

Probability basics



Good to have

Random Variable

A **random variable** is a variable that takes on different values determined by chance. In other words, it is a numerical quantity that varies at random.

Discrete Random Variable
When the random variable can assume only a countable, sometimes infinite, number of values.

Continuous Random Variable
When the random variable can assume an uncountable number of values in a line interval.

Note on notation! We use capitalized letters to represent the random variables and lowercase for the specific values of the variable.

https://online.stat.psu.edu/stat500/book/export/html/470

PMF and PDF

PMF

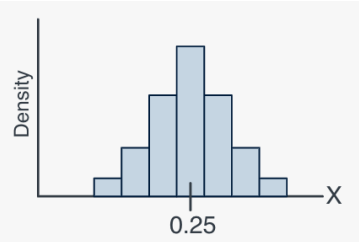
If the random variable is a **discrete random variable**, the probability function is usually called the **probability mass function (PMF)**. If X is discrete, then $f(x) = P(X = x)$. In other words, the PMF for a constant, x, is the probability that the random variable X is equal to x. The PMF can be in the form of an equation or it can be in the form of a table.

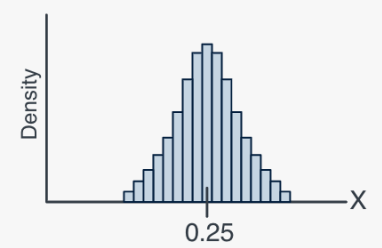
Properties of probability mass functions:

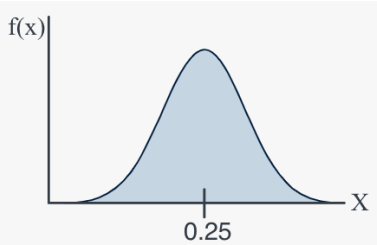
1. $f(x) > 0$, for x in the sample space and 0 otherwise.
2. $\sum_x f(x) = 1$. In other words, the sum of all the probabilities of all the possible outcomes of an experiment is equal to 1.

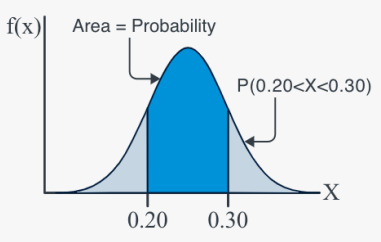
CMF?

PDF









A probability density function f_X of a random variable X is a mapping $f_X : \Omega \to \mathbb{R}$, with the property that

- Non-negativity: $f_X(x) \geq 0$ for all $x \in \Omega$
- Unity: $\int_\Omega f_X(x)dx = 1$
- Measure of a set: $\mathbb{P}[\{x \in A\}] = \int_A f_X(x)dx$

Let X be a continuous random variable. The probability density functic (PDF) of X is a function $f_X : \Omega \to \mathbb{R}$, when integrated over an interval $[a, b]$, yields the probability of obtaining $a \leq X \leq b$:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x)dx.$$

https://probability4datascience.com/slides/Slide_4_01.pdf

Mean and Variance

- Variance:

$$Var(X) = E((X-\mu)^2)$$

$$Var(X) = E(X^2) - \mu^2$$

  - Discrete RVs:

$$V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \, \mathrm{P}(\mathrm{X} = v_i)$$

  - Continuous RVs:

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

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

Do this at home

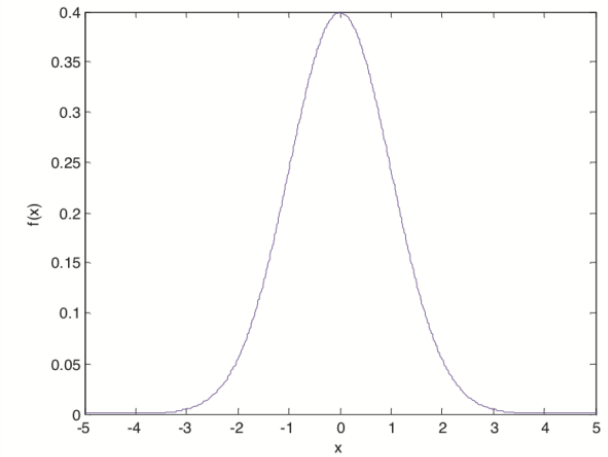
$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

https://isl.stanford.edu/~abbas/ee178/lect02-2.pdf

https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/descriptive-statistics/variance-and-standard-deviation.html

Most common distribution

- Normal X $N(\mu, \sigma^2)$
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
  - E.g. the height of a population



Joint Probability Distribution

- P(X= x, Y= y) = P(x, y)

$$\sum_x \sum_y P(X = x, Y = y) = 1$$

$$\int_x \int_y f_{X,Y}(x, y)\,dx\,dy = 1$$

$$p(x) = P(X = x) = \sum_y p(x, y); \quad p(y) = P(Y = y) = \sum_x p(x, y)$$

are respectively called the **marginal** distributions of X and Y.

$$P_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n)\,dx_2 \cdots dx_n.$$

Example 1: The joint distribution of $p(x, y)$ of X (number of cars) and Y (the number of buses) per signal cycle at a traffic signal is given by

		y		
p(x,y)		0	1	2
x	0	0.025	0.015	0.010
	1	0.050	0.030	0.020
	2	0.125	0.075	0.050
	3	0.150	0.090	0.060
	4	0.100	0.060	0.040
	5	0.050	0.030	0.020

(a) Find $P(X = Y)$.
(b) Find the marginal distribution of X and Y.

(a) The number of cars equals the number of buses if $X = Y$. Hence, $P(X = Y) = p(0, 0) + p(1, 1) + p(2, 2) = .025 + .030 + .050 = .105$. That is, about 10.5% of the time.

(b) Adding the row values yields the marginal distribution of the x values:

x	0	1	2	3	4	5
p(x):	0.05	0.1	0.25	0.3	0.2	0.1

Similarly, adding the column values yields the marginal distribution of the y as:

y:	0	1	2
p(y):	0.50	0.30	0.20

https://stt.msu.edu/Academics/ClassPages/uploads/US19/351-201/Lecture-08.pdf

https://www.stat.uchicago.edu/~yibi/teaching/stat234/2022/L09.pdf

Joint Probability Distribution

Let X, Y and Z be three jointly continuous random variables with joint PDF

$$f_{XYZ}(x,y,z) = \begin{cases} c(x+2y+3z) & 0 \le x,y,z \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the constant c.
2. Find the marginal PDF of X.

https://www.probabilitycourse.com/chapter6/6_1_1_joint_distributions_independence.php

Joint Probability Distribution

Let X, Y and Z be three jointly continuous random variables with joint PDF

$$f_{XYZ}(x,y,z) = \begin{cases} c(x+2y+3z) & 0 \le x,y,z \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the constant c.
2. Find the marginal PDF of X.

1.

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XYZ}(x,y,z)dxdydz \\ &= \int_0^1\int_0^1\int_0^1 c(x+2y+3z)\ dxdydz \\ &= \int_0^1\int_0^1 c\left(\frac{1}{2}+2y+3z\right)\ dydz \\ &= \int_0^1 c\left(\frac{3}{2}+3z\right)\ dz \\ &= 3c. \end{aligned}$$

Thus, $c = \frac{1}{3}$.

$$f_X(x) = \begin{cases} \frac{1}{3}\left(x+\frac{5}{2}\right) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Mean and Variance

- Variance:

$$Var(X) = E((X-\mu)^2)$$

$$Var(X) = E(X^2) - \mu^2$$

  - Discrete RVs:

$$V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \, \mathrm{P}(\mathrm{X} = v_i)$$

  - Continuous RVs:

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

- Covariance:

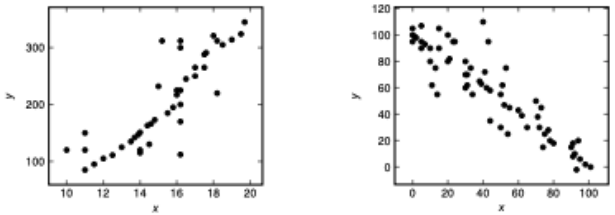
$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x \mu_y$$

Covariance

Covariance

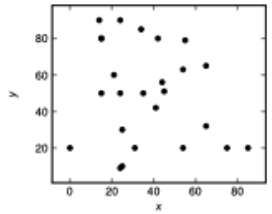
Definition: The covariance between two random variables X and Y is

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] \\ &= E(XY)-\mu_x\mu_y \\ &= \begin{cases} \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y), & \text{if } X \text{ \& } Y \text{ are discrete} \\ \int\int (x-\mu_x)(y-\mu_y)f(x,y)dxdy, & \text{if } X \text{ \& } Y \text{ are continuous} \end{cases} \end{aligned}$$



(a) Positive correlation

(b) Negative correlation



(c) No correlation

Correlation

Definition: The correlation coefficient between X and Y, denoted Corr(X,Y) or $\rho_{X,Y}$ or simply ρ is defined as (here σ_x denotes the SD of X)

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

https://stt.msu.edu/Academics/ClassPages/uploads/US19/351-201/Lecture-08.pdf

Conditional Probability

I roll a fair die. Let A be the event that the outcome is an odd number, i.e., $A = \{1, 3, 5\}$. Also let B be the event that the outcome is less than or equal to 3, i.e., $B = \{1, 2, 3\}$. What is the probability of A, $P(A)$? What is the probability of A given B, $P(A|B)$?

Solution

This is a finite sample space, so

$$P(A) = \frac{|A|}{|S|} = \frac{|\{1, 3, 5\}|}{6} = \frac{1}{2}.$$

Now, let's find the conditional probability of A given that B occurred. If we know B has occurred, the outcome must be among $\{1, 2, 3\}$. For A to also happen the outcome must be in $A \cap B = \{1, 3\}$. Since all die rolls are equally likely, we argue that $P(A|B)$ must be equal to

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{2}{3}.$$

Now let's see how we can generalize the above example. We can rewrite the calculation by dividing the numerator and denominator by $|S|$ in the following way

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} = \frac{P(A \cap B)}{P(B)}.$$

If A and B are two events in a sample space S, then the **conditional probability of A given B** is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0.$$

https://www.probabilitycourse.com/chapter1/1_4_1_independence.php

Independence

Let A be the event that it rains tomorrow, and suppose that $P(A) = \frac{1}{3}$. Also suppose that I toss a fair coin; let B be the event that it lands heads up. We have $P(B) = \frac{1}{2}$.

Now I ask you, what is $P(A|B)$? What is your guess? You probably guessed that $P(A|B) = P(A) = \frac{1}{3}$. You are right! The result of my coin toss does not have anything to do with tomorrow's weather. Thus, no matter if B happens or not, the probability of A should not change. This is an example of two **independent** events. Two events are independent if one does not convey any information about the other. Let us now provide a formal definition of independence.

> Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

Now, let's first reconcile this definition with what we mentioned earlier, $P(A|B) = P(A)$. If two events are independent, then $P(A \cap B) = P(A)P(B)$, so

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)P(B)}{P(B)} \\ &= P(A). \end{aligned}$$

https://www.probabilitycourse.com/chapter1/1_4_1_independence.php

		y			
$p(x,y)$		1	2	3	$f_X(x)$
	1	0.04	0.12	0.04	0.2
x	2	0.12	0.36	0.12	0.6
	3	0.04	0.12	0.04	0.2
$f_Y(y)$		0.2	0.6	0.2	

Find joint pmf from marginal and vice versa

Walk through the example yourself

https://www.stat.uchicago.edu/~yibi/teaching/stat234/2022/L09.pdf

Bayes Theorem

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

Dividing by $P(A)$, we obtain

prior

posterior

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

Find the conditional probability of an event, if we know the reverse conditioning

A desk lamp produced by The Luminar Company was found to be defective (D). There are three factories (A, B, C) where such desk lamps are manufactured. A Quality Control Manager (QCM) is responsible for investigating the source of found defects. This is what the QCM knows about the company's desk lamp production and the possible source of defects:

Factory	% of total production	Probability of defective lamps
A	$0.35 = P(A)$	$0.015 = P(D\|A)$
B	$0.35 = P(B)$	$0.010 = P(D\|B)$
C	$0.30 = P(C)$	$0.020 = P(D\|C)$

Disease / Symptom ?

P(C|D) = ?

P(D) = P(D∩A) + P(D∩A) + P(D∩A)

Till next time!