CS 559: Machine Learning Fundamentals & Math Review I

Lecture 1

In Jang

ijang@stevens.edu

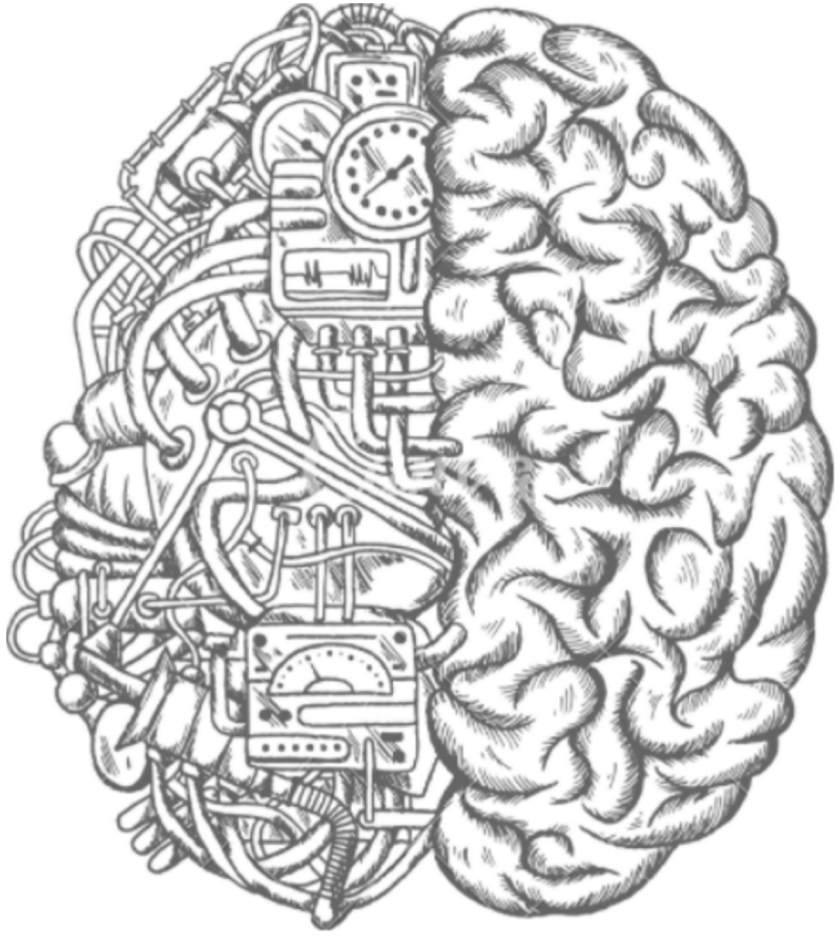
STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Announcement

- No Presentation Assignment for this lecture.
- The lecture will be posted on Tuesday as well as the assignments.
- The first assignment will be published next week.

Machine Learning

- ML is everywhere!
  - Computer Science
  - Healthcare
  - Retail
  - Manufacturing
  - Energy
  - Financial Service
  - …



What is Machine Learning?

A computer program is said to learn from ***experience*** *E* with respect to some class of ***tasks*** *T* and performance ***measure*** *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

What is Machine Learning?

- Machine Learning:
  - The term first coined in 1959, by Arthur Samuel from IBM
  - A branch of Artificial Intelligence (AI),
  - Focused on design and development of algorithm
  - Input: empirical data, such as that from sensors or databases,
  - Output: ***patterns*** or ***predictions*** thought to be features of the underlying mechanism that generated the data.
- Learner (the algorithm):
  - Takes advantage of ***data*** to capture ***characteristics of interest*** of their unknown underlying probability distribution.
- One fundamental difficulty:
  - **Generalization**: The set of all possible behaviors given all possible inputs ***is too large*** to be included in the set of observed examples (training data). Hence the learner must ***generalize*** from the given examples in order to produce a useful output in new cases.

ML from Other Aspects

- The Artificial Intelligence View:
  - Learning is central to **human** knowledge and intelligence, and, likewise, it is also essential for building **intelligent machines**.
  - Years of effort in AI has shown that trying to build intelligent computers by programming all the rules cannot be done; automatic learning is crucial.
  - For example, we humans are not born with the ability to understand language — we learn it — and it makes sense to try to have computers learn language instead of trying to program it all it.

ML from Other Aspects

- The Software Engineering View:
  - Machine learning allows us to program computers by example, which can be easier than writing code in the traditional way.

- The Statistics View:
  - Machine learning is the marriage of computer science and statistics: computational techniques are applied to statistical problems.
  - Machine learning has been applied to a vast number of problems in many contexts, beyond the typical statistics problems.
  - Machine learning is often designed with different considerations than statistics (e.g., speed is often more important than accuracy).

Examples of ML

- Spam Filtering
- Goal: given an email, decide whether it is spam
- The learner learns from
  - Emails marked as spam
  - Emails not marked as spam (inbox)

Examples of ML

- Face Detection

Examples of ML

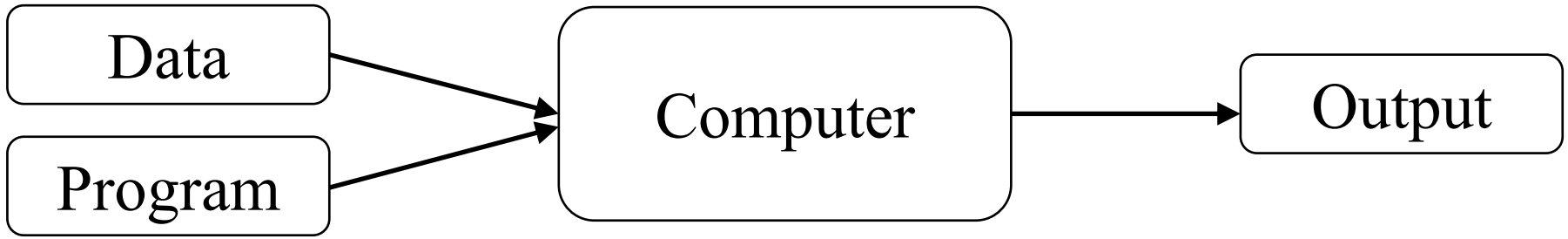
- Games

ML in Practice

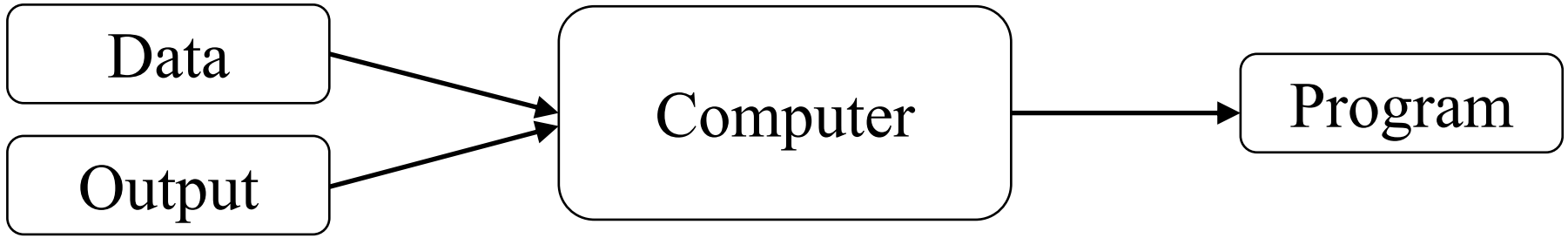
- ML is about:
  - Given a collections of examples, called “training data”
  - We want to predict something about novel examples, called “test data”
- What we usually do:
  - Build ***idealized models*** of the application area we are working in
  - Develop algorithms and implement in code
  - Use historical data to learn numeric parameters, and sometimes model structure
  - Use test data to validate the learned model, quantitatively measure its predictions
  - Assess errors and repeat…

ML vs Traditional Approach

- Traditional Programming



- Machine Learning



ML in a Nutshell

- Every machine learning algorithm has three components:
  - Representation / Model Class
  - Evaluation / Objective Function
  - Optimization

Representation / Model Class

- Decision trees
- Sets of rules / Logic programs
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles

Evaluation / Objective Function

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence

Optimization

- Discrete optimization
  - Minimal Spanning Tree
  - Shortest Path

- Continuous Optimization
  - Gradient Descent
  - Linear Programming

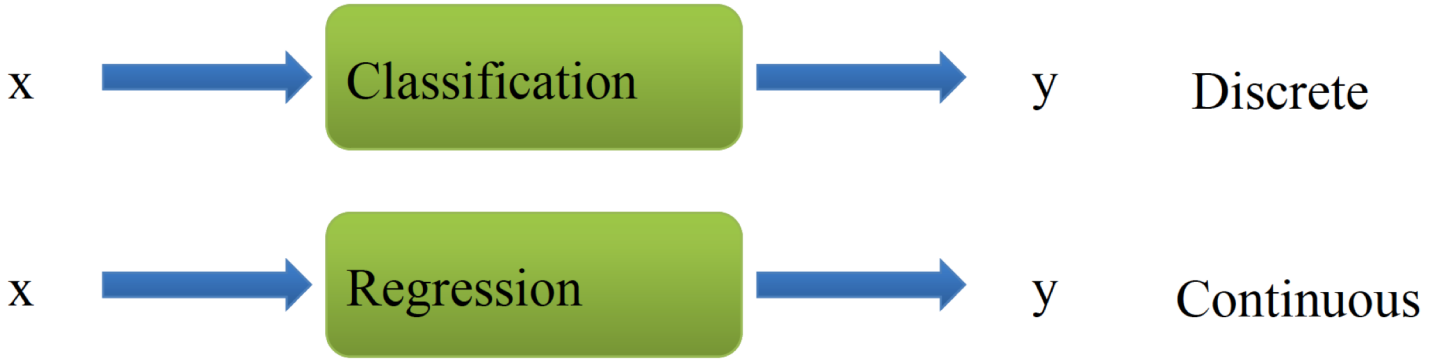
Types of ML

- Supervised Learning
  - Training data include desired outputs
  - Test data only have features, must predict outputs

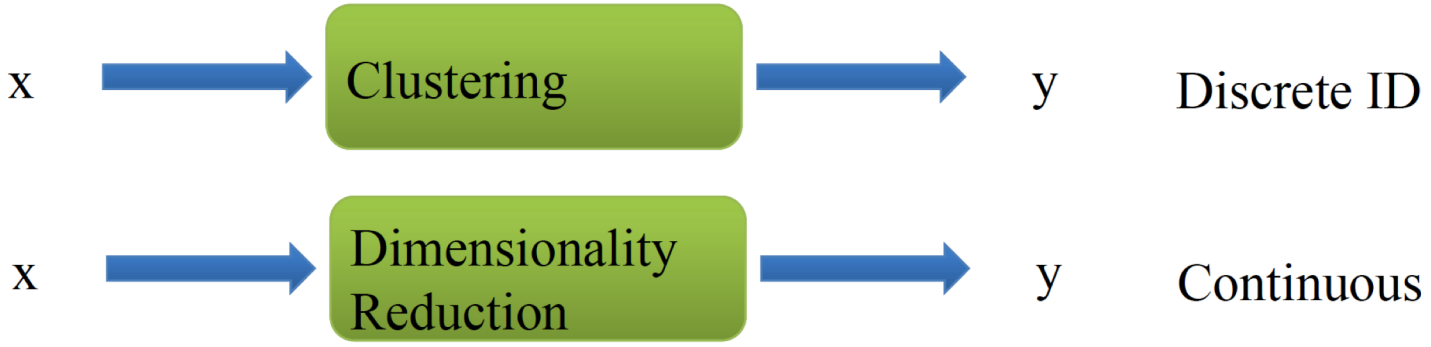
- Unsupervised learning
  - Training data do not include desired outputs

Types of ML

Supervised Learning



Unsupervised Learning



Intro to Linear Algebra

Vector Space

- A vector space is a collection of objects called ***vectors***, which may be added together and multiplied ("scaled") by numbers, called ***scalars***.
- A vector space over a field F (such real numbers) is a set V together with **two** operations that satisfy the **eight** axioms listed in the next slide.
- Two Operations
  - **Vector addition** or simply **addition** $+ : V \times V \rightarrow V$, takes any two vectors **v** and **w** and assigns to them a third vector which is commonly written as **v** + **w**, and called the sum of these two vectors. (Note that the resultant vector is also an element of the set V).
  - **Scalar multiplication** $\cdot : F \times V \rightarrow V$, takes any scalar a and any vector **v** and gives another vector a**v**. (Similarly, the vector a**v** is an element of the set V).

Vector Space

The eight axioms

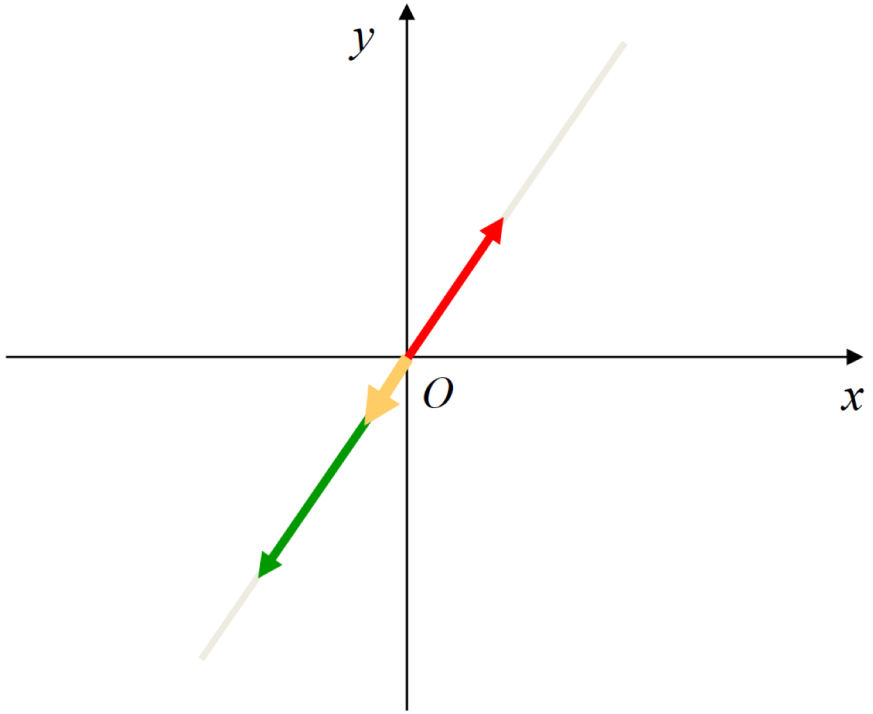
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
Identity element of addition	There exists an element $0 \in V$, called the zero vector, such that $v + 0 = v$ for all $v \in V$.
Inverse elements of addition	For every $v \in V$, there exists an element $-v \in V$, called the additive inverse of v, such that $v + (-v) = 0$.
Compatibility of scalar multiplication with field multiplication	$a(bv) = (ab)v$
Identity element of scalar multiplication	$1v = v$, where 1 denotes the multiplicative identity in F.
Distributivity of scalar multiplication with respect to vector addition	$a(u + v) = au + av$
Distributivity of scalar multiplication with respect to field addition	$(a + b)v = av + bv$

Subspace

- Let F be a field, V be a vector space over F, and let W be a subset of V. Then W is a **subspace** if:
  - The zero vector, $\mathbf{0}$, is in W.
  - If $\mathbf{u}$ and $\mathbf{v}$ are elements of W, then the sum $\mathbf{u} + \mathbf{v}$ is an element of W.
  - If $\mathbf{u}$ is an element of W and c is a scalar from K, then the scalar product $c\mathbf{u}$ is an element of W.

Subspace Example

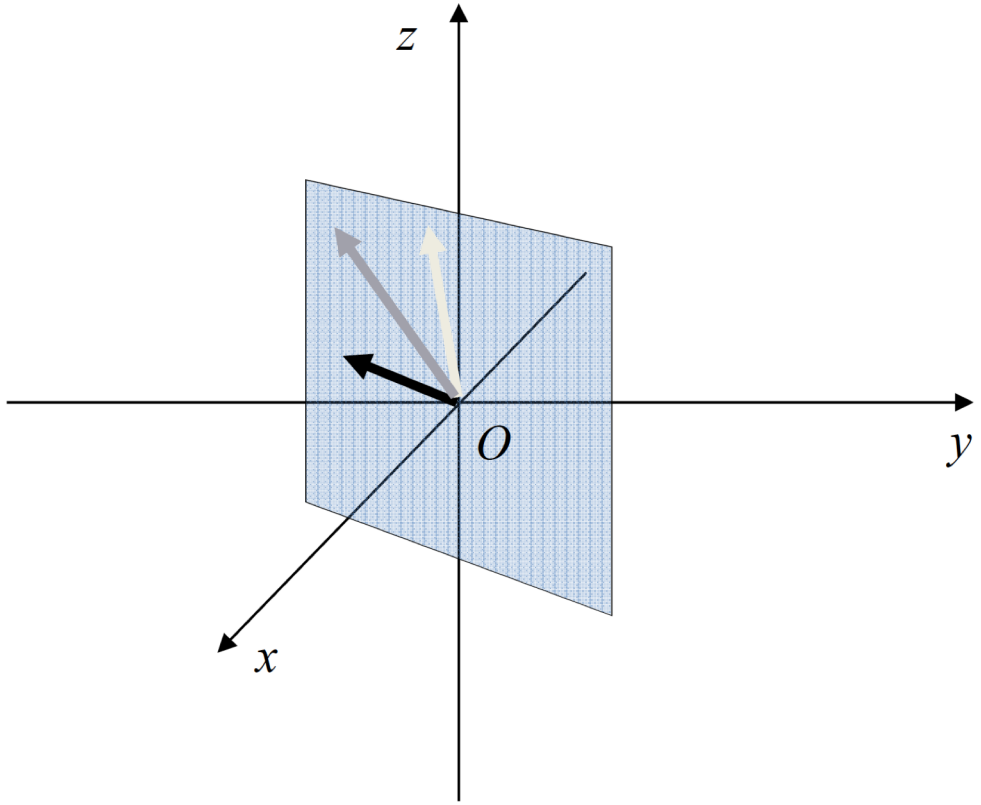
- Let l be a 2D line though the origin
- $L = \{\mathbf{p} - O \mid \mathbf{p} \in l\}$ is a linear subspace of R^2



O. Sorkine, 2006

Subspace Example

- Let π be a plane through the origin in 3D
- $V = \{\mathbf{p} - O \mid \mathbf{p} \in \pi\}$ is a linear subspace of R^3



Linear Independence

- The vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ are a linearly independent set if:

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k = \mathbf{0} \Leftrightarrow \alpha_i = 0\ \forall i$$

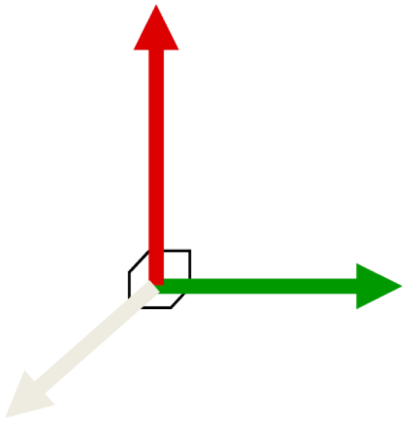
- It means that none of the vectors can be obtained as a linear combination of the others.

Linear independence - example

- Parallel vectors are always dependent:

$$\mathbf{v} = 2.4\,\mathbf{w} \quad \Rightarrow \quad \mathbf{v} + (-2.4)\mathbf{w} = 0$$

- Orthogonal vectors are always linearly independent:



Basis of Vector Space

- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are **linear independent**
- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ **span** the whole vector space $\mathbf{V}$

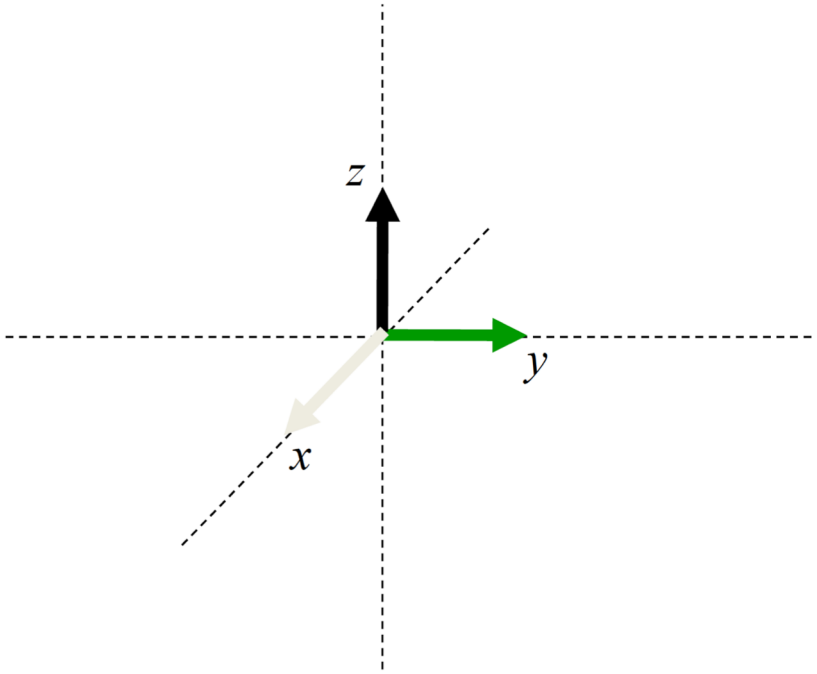
$$\mathbf{V} = \{\alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k | \alpha_1 \in R\}$$

- Any vector in $\mathbf{V}$ is a **unique** linear combination of the basis

- The number of basis vectors is called the **dimension** of $\mathbf{V}$

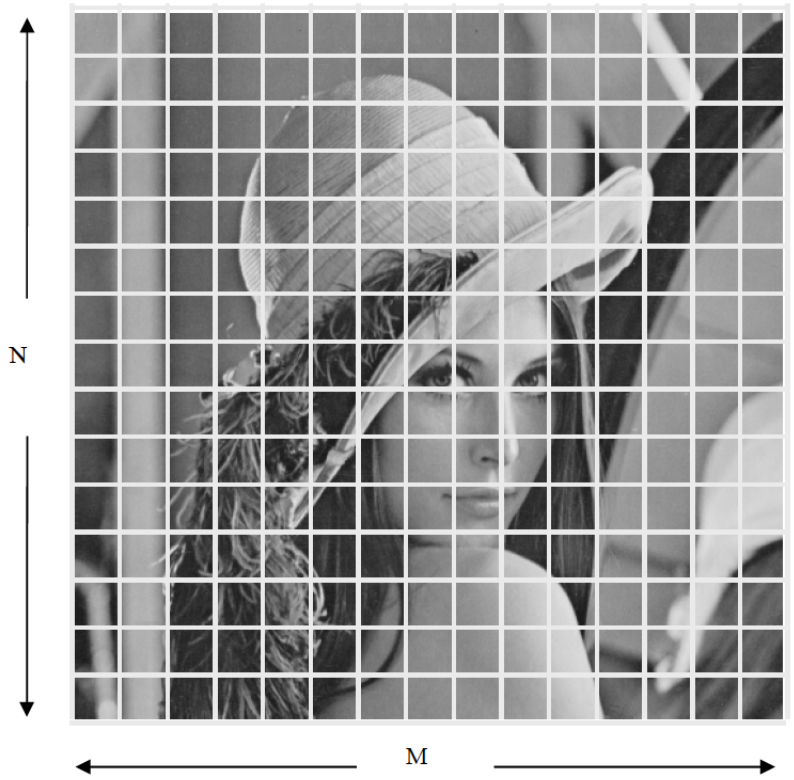
Basis Example

- The standard basis of R^3



Basis – Another Example

- Grayscale NxM images:
  - Each pixel has value between 0 (black) and 1 (white)
  - The image can be interpreted as a vector $\in R^2$



Matrix Representation

- Let $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ be a basis
- Every $\mathbf{v}$ can be uniquely represented as:

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k$$

- Denote $\mathbf{v}$ by the column-vector: $\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$
- Denote the basis vectors as: $\begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1} \end{pmatrix}$

Linear Operators

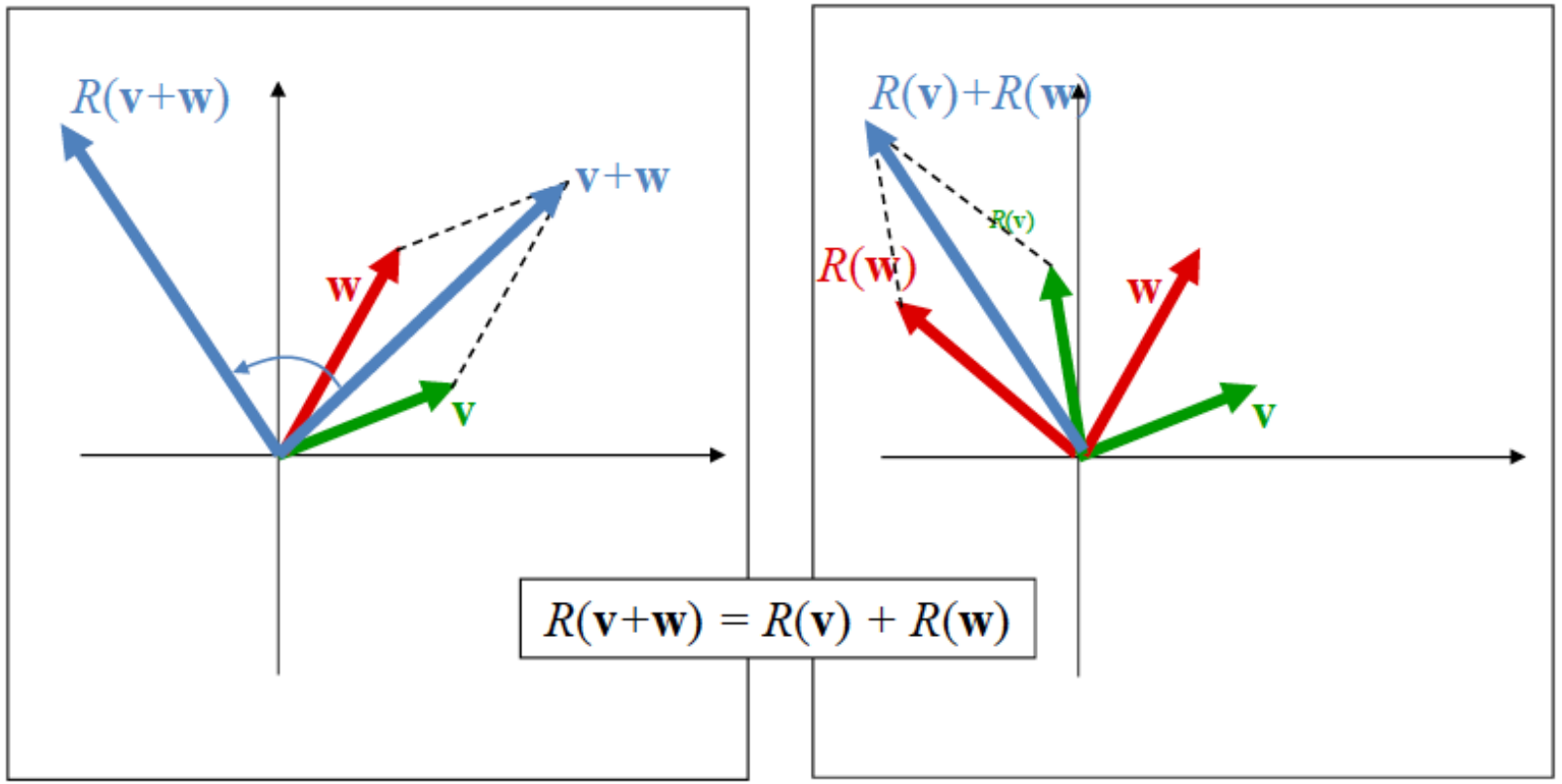
- $A : \mathrm{V} \rightarrow \mathrm{W}$ is called linear operator if:
  - $A(\mathbf{v} + \mathbf{w}) = A(\mathbf{v}) + A(\mathbf{w})$
  - $A(\alpha\mathbf{v}) = \alpha A(\mathbf{v})$

- In particular, $A(\mathbf{0}) = 0$

- Are the following operators linear?
  - Scaling
  - Rotation
  - Translation

Linear operators - illustration

- Rotation is a linear operator:



Matrix Operations

- Addition, subtraction, scalar multiplication
- Multiplication of matrix by column vector:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \sum_i a_{1i} b_i \\ \vdots \\ \sum_i a_{mi} b_i \end{pmatrix} = \begin{pmatrix} < \text{row}_1, \mathbf{b} > \\ \vdots \\ < \text{row}_m, \mathbf{b} > \end{pmatrix}$$

A $\quad$ $\mathbf{b}$

$$\begin{bmatrix} -4 & 4 & 6 \\ 2 & 3 & -1 \end{bmatrix}\begin{bmatrix} 0 & -4 & -2 & 5 \\ 1 & 6 & -1 & 7 \\ 8 & 2 & 4 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} -4\times0 + 4\times1 + 6\times8 & -4\times-4 + 4\times6 + 6\times2 & -4\times-2 + 4\times-1 + 6\times4 & -4\times5 + 4\times7 + 6\times3 \\ 2\times0 + 3\times1 + -1\times8 & 2\times-4 + 3\times6 + -1\times2 & 2\times-2 + 3\times-1 + -1\times4 & 2\times5 + 3\times7 + -1\times3 \end{bmatrix}$$

$$= \begin{bmatrix} 52 & 52 & 28 & 26 \\ -5 & 8 & -11 & 28 \end{bmatrix}$$

Matrix operations

- Transposition: make the rows to be the columns

- $(AB)^T = B^T A^T$

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

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

Matrix properties

- Matrix $A_{(n \times n)}$ is **non-singular** if $\exists B,\ AB = BA = I$
- $B = A^{-1}$ is called the **inverse** of A
- A is non-singular $\Leftrightarrow$ $det\ A \neq 0$

- If A is non-singular then the equation
  - $A\mathbf{x}=\mathbf{b}$ has one unique solution for each $\mathbf{b}$
  - the rows of A are linearly independent (and so are the columns)

Orthogonal matrices

- Matrix $A_{(n\times n)}$ is orthogonal if $A^{-1} = A^T$
- Follows: $AA^T = A^TA = I$
- The rows of A are orthonormal vectors!

  Proof:

$$I = A^TA = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \\ \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & & \mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \mathbf{v}_i^T\mathbf{v}_j \end{pmatrix} = \begin{pmatrix} \delta_{ij} \end{pmatrix}$$

$$\Rightarrow \langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1 \quad \Rightarrow \quad \|\mathbf{v}_i\| = 1; \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$$

Trace

- The trace of a square matrix denoted by tr(A) is sum of the diagonal elements

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

- $tr\left(\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}\right) = A_{11} + A_{22} = 1 + 1 = 2$

Determinant

- For a square matrix A, the determinant is denoted by $|A|$ or $\det(A)$

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

Determinant

- $|A| = |A^{\mathrm{T}}|$
- $|AB| = |A|\,|B|$
- $|A| = 0$, if and only if A is singular
  - Else, $|A^{-1}| = 1/|A|$

Eigenvalues and Eigenvectors

- For an n×n **square** matrix A, e is an eigenvector with eigenvalue λ if

$$Ae = \lambda e$$

- Or

$$(A - \lambda \mathbf{I})e = 0$$

- If $(A - \lambda \mathbf{I})$ is invertible, the only solution is e=0 (trivial)

Eigenvalues and Eigenvectors

$$(A- \lambda \mathbf{I})e=0$$

- For non-trivial solutions:

$$\det(A- \lambda \mathbf{I})=0$$

- Above equation is called the “characteristic polynomial”
- Solutions are not unique
  - If e is an eigenvector αe is also an eigenvector

Simple Example

- For a 2×2 matrix

$$\det[\mathbf{A}-\lambda\mathbf{I}]=\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix}=(a_{11}-\lambda)(a_{22}-\lambda)-a_{12}a_{21}=0$$

$$0=a_{11}a_{22}-a_{12}a_{21}-\lambda(a_{11}+a_{22})+\lambda^2$$

Simple Example

$$0 = a_{11}a_{22} - a_{12}a_{21} - (a_{11} + a_{22})\lambda + \lambda^2$$

$$0 = 1 \cdot 4 - 2 \cdot 2 - (1 + 4)\lambda + \lambda^2$$

$$(1 + 4)\lambda = \lambda^2$$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- The solutions are λ=0 and λ=5
- The eigenvector for the first eigenvalue, λ=0 is:

$$\mathbf{Ax} = \lambda\mathbf{x}, \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

$$\left[\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right] \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1x + 2y \\ 2x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is x=2, y=-1

Simple Example

- The second eigenvalue is λ=5

$$\left[\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}-\begin{bmatrix}5 & 0\\ 0 & 5\end{bmatrix}\right]\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4 & 2\\ 2 & -1\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4x+2y\\ 2x-1y\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$-4x+2y=0,\ \text{and}\ 2x-y=0,\ \text{so},\ x=1, y=2$$

Properties

- The product of the eigenvalues = |A|
- The sum of the eigenvalues = trace(A)
- The eigenvectors are pairwise orthogonal