Maths Preliminaries

Wei Wang @ CSE, UNSW

March 18, 2018

Introduction

- This review serves two purposes:
  - Recap relevant maths contents that you may have learned a long time ago (probably not in a CS course and rarely used in any CS course).
  - More importantly, present it in a way that is useful (i.e., giving semantics/motivations) for understanding maths behind Machine Learning.
- Contents
  - Linear Algebra

Note

- You've probability learned Linear Algebra from matrix/system of linear equations, etc. We will review key concepts in LA from the perspective of linear transformations (think of it as *function*s for now). This perspective provides semantics and intuition into most of the ML models and operations.
  - Here we emphasize more on intuitions; We deliberately skip many concepts and present some contents in an informal way.
- It is a great exercise for you to view related maths and ML models/operations in this perspective *throughout* this course!

A Common Trick in Maths I

Question

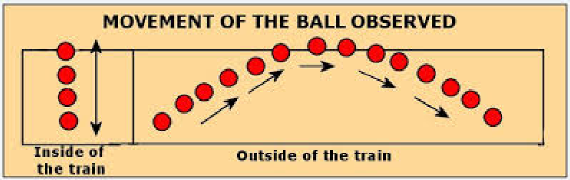
Calculate 2^{10}, 2^{-1}, $2^{\ln 5}$ and 2^{4-3i}?

- Properties:
  - $f_a(n) = f_a(n-1) * a$, for $n \geq 1$; $f_a(0) = 1$.
  - $f(u) * f(v) = f(u+v)$.
  - $f(x) = y \Leftrightarrow \ln(y) = x \ln(a) \Leftrightarrow f(x) = \exp\{x \ln a\}$.
  - $e^{ix} = cos(x) + i \cdot sin(x)$.
- The trick:
- Same in Linear algebra

Objects and Their Representations

Goal

- We need to study the objects
- On one side:
  - A good representation helps (a lot)!
- On the other side:
  - Properties of the objects should be independent of the representation!



Basic Concepts I

Algebra

- a set of objects
- two operations and their identity objects (aka. *identify element*):
  - addition ($+$); its identity is $\mathbf{0}$.
  - *scalar* multiplication ($\cdot$); its identity is $\mathbf{1}$.
- constraints:
  - Closed for both operations
  - Some nice properties of these operations:
    - Communicative: $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$.
    - Associative: $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$.
    - Distributive: $\lambda(\mathbf{a} + \mathbf{b}) = \lambda\mathbf{a} + \lambda\mathbf{b}$.

Basic Concepts II

Think: *What about substraction and division?*

Tips

Always use analogy from algebra on integers ($\mathbb{Z}$) *and* algebra on Polynomials ($\mathcal{P}$).

Why these constraints are natural and useful?

Basic Concepts III

Representation matters?

Consider even geometric vectors: $\mathbf{c} = \mathbf{a} + \mathbf{b}$
What if we represent vectors by a column of their coordinates?
What if by their polar coordinates?

Notes

- Informally, the objects we are concerned with in this course are (column) vectors.
- The set of all n-dimensional real vectors is called $\mathbb{R}^n$.

(Column) Vector

Vector

- A n-dimensional vector, $\mathbf{v}$, is a $n \times 1$ matrix. We can emphasize its shape by calling it a *column* vector.
    - A *row vector* is a transposed column vector: $\mathbf{v}^\top$.

Operations

- Addition: $\mathbf{v}_1 + \mathbf{v}_2 =$
- (Scalar) Multiplication: $\lambda \mathbf{v}_1 =$

Linearity I

Linear Combination: Generalization of Univariate Linear Functions

- Let $\lambda_i \in \mathbb{R}$, given a set of k vectors $\mathbf{v}_i$ ($i \in [k]$), a linear combination of them is
  $\lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \ldots + \lambda_k \mathbf{v}_k = \sum_{i \in [k]} \lambda_i \mathbf{v}_i$
  - Later, this is just $\mathbf{V}\boldsymbol{\lambda}$, where

$$\mathbf{V} = \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \ldots & v_k \\ | & | & | & | \end{bmatrix} \qquad \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_k \end{bmatrix}$$

- Span: All linear combination of a set of vectors is the *span* of them.
- Basis: The minimal set of vectors whose span is exactly the whole $\mathbb{R}^n$.

Linearity II

- Benefit: every vector has a unique decomposition into basis.
  Think: *Why uniqueness is desirable?*

Examples

- Span of $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ is $\mathbb{R}^2$. They are also the basis.
- Span of $\begin{bmatrix}1\\0\end{bmatrix}$, $\begin{bmatrix}0\\1\end{bmatrix}$, $\begin{bmatrix}2\\3\end{bmatrix}$ is $\mathbb{R}^2$. But one of them is *redundant*.
  Think: *Who?*
  - Decompose $\begin{bmatrix}4\\6\end{bmatrix}$

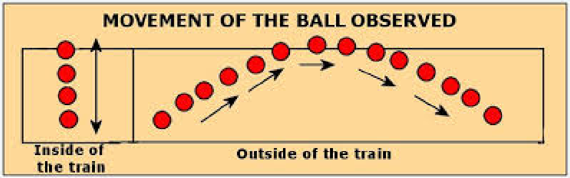
Linearity III

Exercises

- What are the (natural) basis of all (univariate) Polynomials of degrees up to d?
- Decompose $3x^2 + 4x - 8$ into *the* linear combination of 2, $2x - 3$, $x^2 + 1$.

$$3x^2 + 4x - 7 = \textcolor{red}{3}(x^2 + 1) + \textcolor{red}{2}(2x - 3) + \textcolor{red}{(-2)}(2).$$

- The "same" polynomial is mapped to two different vectors under two different bases. **Think:** *Any analogy?*



Matrix I

Linear Transformation

- is a "nice" linear function that maps a vector in $\mathbb{R}^n$ to another vector in $\mathbb{R}^m$.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow{f} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

- The general form:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow{f} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \implies \begin{aligned} y_1 &= M_{11}x_1 + M_{12}x_2 \\ y_2 &= M_{21}x_1 + M_{22}x_2 \\ y_3 &= M_{31}x_1 + M_{32}x_2 \end{aligned}$$

Matrix II

Nonexample

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow{f} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \implies \begin{aligned} y_1 &= \alpha x_1^2 + \beta x_2 \\ y_2 &= \gamma x_1^2 + \theta x_1 + \tau x_2 \\ y_3 &= cos(x_1) + e^{x_2} \end{aligned}$$

Why Only Linear Transformation?

- Simple and nice properties:
  - $(f_1 + f_2)(x) = f_1(x) + f_2(x)$
  - $(\lambda f)(x) = \lambda \cdot f(x)$
  - What about $f(g(x))$?
- Useful

Matrix I

Definition

- A $m \times n$ matrix *corresponds to* a *linear transformation* from $\mathbb{R}^n$ to $\mathbb{R}^m$
  - $f(\mathbf{x}) = \mathbf{y} \implies \mathbf{M}\,\mathbf{x} = \mathbf{y}$, where matrix-vector multiplication is defined as: $y_i = \sum_k M_{ik} \cdot x_k$
  - $\mathbf{M}_{\text{outDim} \times \text{inDim}}$
  - *Transformation* or *Mapping* emphasizes more on the mapping between two sets, rather than the detailed specification of the mapping; the latter is more or less the *elementary* understanding of a *function*. These are all specific instances of *morphism* in category theory.

Semantic Interpretation

Matrix II

- Linear combination of columns of $\mathbf{M}$:

$$\begin{bmatrix} | & | & | & | \\ M_1 & M_2 & \dots & M_n \\ | & | & | & | \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ M_1 & M_2 & \dots & M_n \\ | & | & | & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$\mathbf{y} = x_1 \mathbf{M}_{\bullet 1} + \dots + x_n \mathbf{M}_{\bullet n}$$

- Example:

$$\begin{bmatrix} 1 & 2 \\ -4 & 9 \\ 25 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ -4 \\ 25 \end{bmatrix} + 10 \begin{bmatrix} 2 \\ 9 \\ 1 \end{bmatrix} = \begin{bmatrix} 21 \\ 86 \\ 35 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ -4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ -4 \end{bmatrix} + 10 \begin{bmatrix} 2 \\ 9 \end{bmatrix} = \begin{bmatrix} 21 \\ 86 \end{bmatrix}$$

Matrix III

Think: *What does* $\mathbf{M}$ *do for the last example?*

- Rotation and scaling

- When $\mathbf{x}$ is also a matrix,

$$\begin{bmatrix} 1 & 2 \\ -4 & 9 \\ 25 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 10 & 20 \end{bmatrix} = \begin{bmatrix} 21 & 42 \\ 86 & 172 \\ 35 & 70 \end{bmatrix}$$

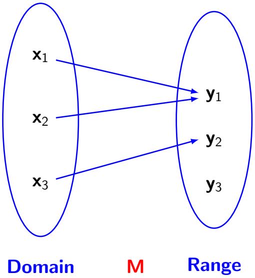
System of Linear Equations I

$$\begin{aligned} y_1 &= M_{11}x_1 + M_{12}x_2 \\ y_2 &= M_{21}x_1 + M_{22}x_2 \\ y_3 &= M_{31}x_1 + M_{32}x_2 \end{aligned} \quad \Longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \\ M_{31} & M_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{M}\mathbf{x}$$

- Interpretation: find a vector in $\mathbb{R}^2$ such that its image (under $\mathbf{M}$) is exactly the given vector $\mathbf{y} \in \mathbb{R}^3$.
- How to solve it?

System of Linear Equations II



The above transformation is *injective*, but not *surjective*.

A Matrix Also Specifies a (Generalized) Coordinate System I

Yet another interpretation

- $\mathbf{y} = \mathbf{M}\mathbf{x} \implies \mathbf{I}\mathbf{y} = \mathbf{M}\mathbf{x}$.
- The vector $\mathbf{y}$ wrt standard coordinate system, $\mathbf{I}$, is the same as $\mathbf{x}$ wrt the coordinate system defined by column vectors of $\mathbf{M}$. **Think:** *why columns of* $\mathbf{M}$*?*

A Matrix Also Specifies a (Generalized) Coordinate System II

Example for polynomials

$$\mathbf{I}: \begin{array}{r} \text{for } 1 \\ \text{for } x \\ \text{for } x^2 \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies \mathbf{M}: \begin{array}{c} \text{for } 3 \\ \text{for } x-1 \\ \text{for } 2x^2+5x-4 \end{array} \begin{bmatrix} 3 & -1 & -4 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\text{Let } \mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} \implies \mathbf{Mx} = \mathbf{I} \begin{bmatrix} -7 \\ 13 \\ 6 \end{bmatrix}$$

Exercise I

- What if $\mathbf{y}$ is given in the above example?
- What does the following mean?

$$\begin{bmatrix} 3 & -1 & -4 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & -1 & -4 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix}$$

- Think about representing polynomials using the basis: $(x-1)^2$, x^2-1, x^2+1.

Inner Product

THE binary operator – some kind of “similarity”

- Type signature: vector $\times$ vector $\to$ scalar: $\langle \mathbf{x}, \mathbf{y} \rangle$.
  - In $\mathbb{R}^n$, usually called *dot product*: $\mathbf{x} \cdot \mathbf{y} \stackrel{\text{def}}{=} \mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i$.
  - For certain functions, $\langle f, g \rangle = \int_a^b f(t) g(t) \, \mathrm{d}t$. $\Rightarrow$ leads to the Hilbert Space
- Properties / definitions for $\mathbb{R}$:
  - conjugate symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
  - linearity in the first argument: $\langle a\mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = a \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$
  - positive definitiveness: $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$; $\langle \mathbf{x}, \mathbf{x} \rangle \Leftrightarrow \mathbf{x} = \mathbf{0}$;
- Generalizes many geometric concepts to vector spaces: angle (orthogonal), projection, norm
  - $\langle \sin nt, \sin mt \rangle = 0$ within $[-\pi, \pi]$ $(m \neq n)$ $\Rightarrow$ they are orthogonal to each other.
- $\mathbf{C} = \mathbf{A}^\top \mathbf{B}$: $C_{ij} = \langle A_i, B_j \rangle$
  - Special case: $\mathbf{A}^\top \mathbf{A}$.

Eigenvalues/vectors and Eigen Decomposition

"Eigen" means "characteristic of" (German)

- A (right) eigenvector of a square matrix $\mathbf{A}$ is $\mathbf{u}$ such that $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$.
- Not all matrices have eigenvalues. Here, we only consider "good" matrices. Not all eigenvalues need to be distinct.
- Traditionally, we normalize $\mathbf{u}$ (such that $\mathbf{u}^\top\mathbf{u} = 1$).
- We can use all eigenvectors of $\mathbf{A}$ to construct a matrix $\mathbf{U}$ (as columns). Then $\mathbf{A}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}$, or equivalently, $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$. This is the Eigen Decomposition.
    - We can interpret $\mathbf{U}$ as a transformation between two coordinate systems. Note that vectors in $\mathbf{U}$ are not necessarily orthogonal.
    - $\boldsymbol{\Lambda}$ as the scaling on each of the directions in the "new" coordinate system.

Similar Matrices

Similar Matrix

- Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ matrix. $\mathbf{A}$ is similar to $\mathbf{B}$ (denoted as $\mathbf{A} \sim \mathbf{B}$) if there exists an invertible $n \times n$ matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$.
- **Think:** *What does this mean?*
  - Think of $\mathbf{P}$ as a *change of basis* transformation.
  - Relationship with the Eigen decomposition.
- Similar matrices have the same value wrt many properties (e.g., rank, trace, eigenvalues, determinant, etc.)

SVD

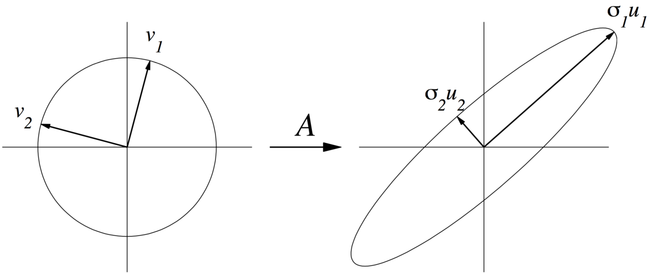
Singular Vector Decomposition

- Let $\mathbf{M}$ be $n \times d$ ($n \geq d$).
- Reduced SVD: $\mathbf{M} = \hat{\mathbf{U}}\hat{\Sigma}\mathbf{V}^\top$ exists for any $\mathbf{M}$, such that
  - $\hat{\Sigma}$ is a diagonal matrix with diagonal elements σ_i (called *singular values*) in decreasing order
  - $\hat{\mathbf{U}}$ consists of an (incomplete) set of basis vectors $\mathbf{u}_i$ (*left singular vectors* in $\mathbb{R}^n$) ($n \times d$: original space as $\mathbf{M}$)
  - $\hat{\mathbf{V}}$ consists of a set of basis vectors $\mathbf{v}_i$ (*right singular vectors* in $\mathbb{R}^d$) ($d \times d$: reduced space)
- Full SVD: $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^\top$:
  - Add the remaining $(n - d)$ basis vectors to $\hat{\mathbf{U}}$ (thus becomes $n \times n$).
  - Add the $n - d$ rows of $\mathbf{0}$ to $\hat{\Lambda}$ (thus becomes $n \times d$).

Geometric Illustration of SVD

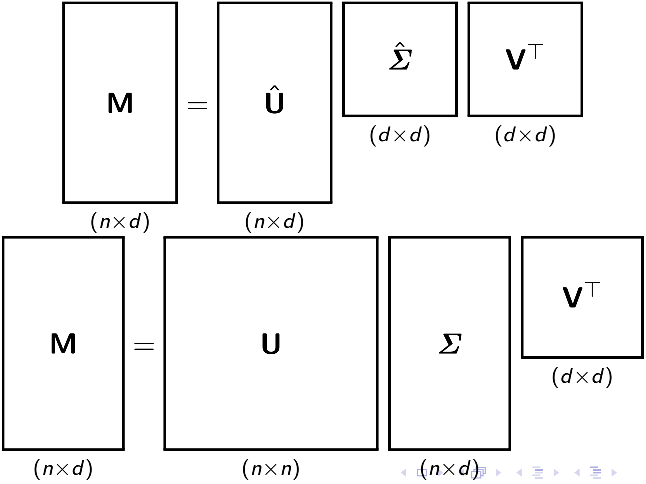
Geometric Meaning

- $\mathbf{M} v_i = \sigma_i u_i$



Graphical Illustration of SVD I

Figure: Reduced SVD vs Full SVD



Graphical Illustration of SVD II

Meaning:

- Columns of $\mathbf{U}$ are the basis of $\mathbb{R}^n$
- Rows of $\mathbf{V}^\top$ are the basis of $\mathbb{R}^d$

SVD Applications I

Relationship between Singular Values and Eigenvalues

- What are the eigenvalues of $\mathbf{M}^\top\mathbf{M}$?
- Hint: $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ and $\mathbf{U}$ and $\mathbf{V}$ are unitary (i.e., rotations)

- Related to *PCA (Principle Component Analysis)*

References and Further Reading I

- Gaussian Quadrature:
  https://www.youtube.com/watch?v=k-yUdqRXijo
- Linear Algebra Review and Reference.
  http://cs229.stanford.edu/section/cs229-linalg.pdf
- Scipy LA tutorial. https://docs.scipy.org/doc/scipy/reference/tutorial/linalg.html
- We Recommend a Singular Value Decomposition.
  http://www.ams.org/samplings/feature-column/fcarc-svd