CSI 436/536
Introduction to Machine Learning

Review of Linear Algebra (1)

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York
Importance of linear algebra

• Linear algebra
  • provides superior notations (algebra)
    • many topics can be understood better with vector-matrix-space idea (e.g., Fourier)
  • has a consistent intuition (geometry)
    • what is true for low dimensional space is usually also true for high dimensional space
      • not usually the case in general
  • computes efficiently (numerical algorithms)
    • Almost all numerical computation requires support of linear algebra
    • LAPACK is the backbone of Matlab, NumPy, R
Algebra

- Fourier transform

\[
DFT(FFT):
X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j \frac{2\pi}{N} nk} \quad (k = 0, 1, \ldots, N - 1)
\]

\[
IDFT(IFFT):
x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \cdot e^{j \frac{2\pi}{N} nk} \quad (n = 0, 1, \ldots, N - 1)
\]

- FT provide eigenvectors for circulant matrix (discrete case) or LTI operator (continuous case)

- proof of the fundamental convolution theorem (and its continuous version) becomes very easy
curse of dimensionality

• sphere inscribed in cube

\[ \text{as } d \to \infty. \]

• Gaussian distribution in high dimension

\[ \text{as } d \to \infty. \]
Numerical linear algebra

- NLA is behind the majority of numerical procedures for machine learning
  - The majority of ML algorithms are optimization problems [there is a small fraction is about integration instead of optimization]
  - All optimization problems are practically solved as a sequence of quadratic optimization problems
  - All quadratic optimization problems are solved as linear equations or eigenvalues
Overview

- Objects in linear algebra
  - vectors, linear spaces, matrices, linear transforms
- Problems in linear algebra
  - linear equation $Ax = b$
  - eigenvalue equation $Ax = \lambda x$
- Techniques in linear algebra
  - Matrix factorizations: LU decomposition, eigen decomposition, QR decomposition, etc
- Mostly we will work with
  - Symmetric positive (semi)definite matrices
vectors, space and transforms

- Vectors are list of numbers over a field (real space)
  - Geometrically correspond to points
  - we use column vector by default
  - vector can add/subtract/scale

- Linear space is the set of vectors closed under addition and scalar product

- Subspace is a subset of a space including zero

- A space can be **spanned** by a set of vectors
  \[
  \alpha_1, \cdots, \alpha_k \in \mathbb{R}, \quad \sum_{i=1}^k \alpha_i \mathbf{x}_i
  \]

- A linear transform is a mapping between points in two spaces that keeps linearity

\[
x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}
\]
linear independence of vectors

• A set of vectors is linear independent if for any $\vec{u}_i$ is not in $\text{span}(\vec{u}_1, \cdots, \vec{u}_{i-1}, \vec{u}_{i+1}, \cdots, \vec{u}_n)$

• A set of bases of a space $V$ is a set of independent vectors that also span it
  • Canonical basis is the basis that are orthonormal
  • Coordinates are coefficients on basis
  • the $\text{max}$ number of vectors that are linearly independent in a space is its dimension
  • Dimension of a space may not be the same as the dimension of an individual vector in it
additional structures of space

• distance between two vectors: **metric**
  • metric space
• length of a vector: **norm**
  • norm space
• angle between two vectors: **inner product**
  • inner product space (Hilbert space)
• parallelogram by two vectors: **exterior product**
  • Grassmann space
Vector metrics (distance)

- L2 (Euclidean) metric
- L1 (Manhattan) metric
- $L_\infty$ metric
- Lp metric ($p \geq 1$)
- All metrics satisfy
  - symmetric: $d(x,y) = d(y,x)$
  - non-negativity: $d(x,x) \geq 0$
  - triangle inequality: $d(x,y) + d(y,z) \geq d(x,z)$
Norms

- L2 (Euclidean) norm, L1 (Manhattan) norm, L∞ norm, Lp norm (p ≥ 1)
- All norms satisfy
  - non-negativity: |x| ≥ 0
  - triangle inequality: |x| + |y| ≥ |x+y|
- Normalization to unit vectors (w.r.t. to a norm)
  - Projections onto unit spheres (w.r.t. to a norm)
- Given a norm, we can define metric (distance) as the norm of the different vector
- due norm: ||x||_{p^*} = \max\{s^Tx | \|s\|_p \leq 1\}, L2 is self-dual, L1 is dual of L∞
Vector products

- inner (scalar) product: \((\mathbf{v}, \mathbf{v}) \rightarrow\) a number \(\mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i\)
  - geometrically related with angles
  - Cauchy-Schwartz inequality
  \[ |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \| \mathbf{u} \| \| \mathbf{v} \|\]
  - \(\langle \mathbf{u}, \mathbf{v} \rangle = 0\) iff \(\mathbf{u}\) and \(\mathbf{v}\) orthogonal
- rect (Dirac) product: \((\mathbf{v}, \mathbf{v}) \rightarrow\) a vector
- exterior (cross/wedge) product: \((\mathbf{v}, \mathbf{v}) \rightarrow\) a vector
- outer (tensor) product: \((\mathbf{v}, \mathbf{v}) \rightarrow\) a matrix (actually a tensor)

\[
\mathbf{x} \odot \mathbf{y} = \begin{pmatrix}
  x_1 \cdot y_1 \\
  x_2 \cdot y_2 \\
  \vdots \\
  x_n \cdot y_n
\end{pmatrix}
\]

\[
\mathbf{x} \mathbf{y}^T = \begin{pmatrix}
  x_1 \cdot y_1 & x_1 \cdot y_2 & \cdots & x_1 \cdot y_m \\
  x_2 \cdot y_1 & x_2 \cdot y_2 & \cdots & x_2 \cdot y_m \\
  \vdots & \vdots & \ddots & \vdots \\
  x_n \cdot y_1 & x_n \cdot y_2 & \cdots & x_n \cdot y_m
\end{pmatrix}
\]
matrix

- matrix is 2D table of numbers
  - all matrices of the same dim form a vector space
- the transpose of a matrix $A$, denoted $A^T$, is the matrix whose $(i,j)$ entry equals the $(j,i)$ entry of $A$
- Matrix multiplication
- non communicative multiplication, $AB \neq BA$ usually
  - as outer product of “inner products”
  - as inner product of “outer products”
Some special matrices

- Square and rectangular matrices
- Diagonal and identity matrices
  \[
  \begin{pmatrix}
  1 & 0 \\
  0 & 4
  \end{pmatrix}
  \quad
  \begin{pmatrix}
  1 & 0 \\
  0 & 1
  \end{pmatrix}
  \]
- Upper and lower triangular matrices
  \[
  \begin{pmatrix}
  1 & 0 \\
  2 & 4
  \end{pmatrix}
  \quad
  \begin{pmatrix}
  1 & 2 \\
  0 & 4
  \end{pmatrix}
  \]
- Symmetric matrices $A^T = A$
- skew-symmetric matrices $A^T = -A$
- Matrix inverse $A^{-1}A = AA^{-1} = I$
- orthogonal matrices: $A^TA = AA^T = I$, or $A^T = A^{-1}$
Solving linear equations

• The most important problem in LA is solving the linear equation: \( Ax = b \), \( b \) is a known vector (dim n), \( x \) is unknown vector (dim m)

• \( A \) is a matrix (dim n x m): collection of m vectors

\[
A = \begin{pmatrix}
| & | & \cdots & | \\
| a_1 | a_2 | \cdots | a_m |
| & | & \cdots & | \\
| & | & \cdots & | \\
| & | & \cdots & |
\end{pmatrix} = \text{col}(a_1a_2\cdots, a_m)
\]

• \( Ax \) represents all vectors in the column space of \( A \)

\[
Ax = x_1a_1 + x_2a_2 + \cdots + x_ma_m
\]

• \( Ax = 0 \) is the null space of \( A \), with \( x = 0 \) always in it

• Column space determines the existence of the solution, null space determines the uniqueness of the solution
Geometric interpretation

• To solve $Ax = b$ is equivalent to find a representation of $b$ in the column space of $A$

$$Ax = x_1a_1 + x_2a_2 + \cdots + x_ma_m = b$$

• If $b$ is in $\text{col}(A)$, solution exists

• If $\text{null}(A) = \{0\}$, solution is unique
Solve $Ax = b$

- **case 1**: matrix $A$ is square and full ranked  
  $n = m$, # of equations = # of unknowns  
  $\Rightarrow$ complete problem $\Rightarrow$ **unique** solution

- **case 2**: matrix $X$ is tall & thin  
  $n > m$, # of equations > # of unknowns  
  $\Rightarrow$ over-complete problem $\Rightarrow$ **no** solution

- **case 3**: matrix $A$ is short & fat  
  $n < m$, # of equations < # of unknowns  
  $\Rightarrow$ under-complete problem $\Rightarrow$ **non-unique** solution
matrix inverse

• for **square matrix** $A$, if $\det(A) \neq 0$, then $A^{-1}$ is defined as the matrix satisfying $A^{-1}A = AA^{-1} = I$

• matrix $A$ is invertible, otherwise, it is singular

• For a $2 \times 2$ matrix, inverses can be computed as

$$
B = \begin{bmatrix}
A & B \\
C & D
\end{bmatrix} \quad \text{If } AD - BC \neq 0, \text{ then } B
\text{ has an inverse, denoted } B^{-1}
$$

$$
B^{-1} = \frac{1}{AD - BC} \begin{bmatrix}
D & -B \\
-C & A
\end{bmatrix}
$$

• for **rectangular matrix** $A$

• its **left** Moore-Penrose pseudo inverse $(A^TA)^{-1}A^T$

• its **right** Moore-Penrose pseudo inverse $A^T(AA^T)^{-1}$
Matrix trace & determinant

- **trace**
  - property: \( \text{tr}(AB) = \text{tr}(BA^T) \)

- **determinant**
  - computation involves Levi-Civita tensor
    \[
    \text{det} A = (a_1 b_2 c_3 + b_1 c_2 a_3 + c_1 a_2 b_3) - (a_3 b_2 c_1 + b_3 c_2 a_1 + c_3 a_2 b_1)
    \]
  - \( \text{det}(aA) = a^n \text{det}(A) \), \( \text{det}(AB) = \text{det}(A)\text{det}(B) \)
  - \( \text{det}(A^{-1}) = \frac{1}{\text{det}(A)} \)
  - A not invertible, then \( \text{det}(A) = 0 \), and vice versa
Solve $Ax = b$ using matrix inverse

• for square matrix $A$, if $\det(A) \neq 0$, then $A^{-1}$ is defined as the matrix satisfying $A^{-1}A = AA^{-1} = I$
  
• matrix $A$ is invertible, otherwise, it is singular

$$
\begin{align*}
2x + 3y &= 6 \\
4x + 9y &= 15
\end{align*}
\Rightarrow
\begin{bmatrix}
2 & 3 \\
4 & 9
\end{bmatrix}
\begin{bmatrix}
x \\
y
\end{bmatrix}
= 
\begin{bmatrix}
6 \\
15
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
x \\
y
\end{bmatrix}
= A^{-1}
\begin{bmatrix}
6 \\
15
\end{bmatrix}
$$

• Why is this not a good way to solve linear equation
  
• Running time is $O(n^3)$
  
• Numerically unstable
  
\[
\begin{pmatrix}
1 & 0 \\
0 & \epsilon
\end{pmatrix}^{-1}
= \begin{pmatrix}
1 & 0 \\
0 & \frac{1}{\epsilon}
\end{pmatrix}
\]

• Lose of good structure in $A$, e.g., sparsity

• On modern computers, for matrix smaller than 1000 dimension, direct inverse is feasible.
Solve $Ax = b$ using decomposition

- We can decompose a square matrix $A = LDU$, where $L$ and $U$ are a lower triangular and upper matrices with diagonal 1, and $D$ is a diagonal matrix with pivots
- If $A$ is not invertible, then one of the pivot is zero
- Solving $Ax = b$ becomes $LDUx = b$, then two steps $Ly = b$ (forward elimination), $DUx = y$ (backward elimination)
  - This is known as Gaussian elimination
  - Solution time is $O(n^2)$, and numerically it is very stable (caveat: if the pivots are chosen right)
  - It is numerically stable (only divide by pivot)
Projection

- for \( \text{col}(X) \) as a 2D subspace of the 3D space
- least squares problem is equivalent to finding the projection of vector \( y \) in \( \text{col}(X) \)

The transform \( \Pi_X(y) \) is known as the *projection* of \( y \) on \( X \). The geometrical interpretation of \( \Pi_X(y) \) is that it is the vector in \( \text{col}(X) \) that has the minimum \( \ell_2 \) distance to \( y \).

\[
\Pi_X(y) = X(X^TX)^{-1}X^Ty
\]

- idempotent \( \Pi_X(x) = x \), for \( x \in \text{col}(X) \)
- orthogonality \( y - \Pi_X(y) \perp X \)
- Householder transform mirror reflection
  \( H(y) = 2 \Pi_X(y) - y \)
Positive definite matrix

- A is a square matrix, for any \( x \neq 0 \), we form a quadratic form using A and x, \( x^T Ax \), then if
  - \( x^T Ax > 0 \), A is a positive definite matrix
  - \( x^T Ax < 0 \), A is a negative definite matrix
  - \( x^T Ax \geq 0 \), A is a positive semi-definite matrix
  - \( x^T Ax \leq 0 \), A is a negative semi-definite matrix
  - otherwise, A is indefinite

- Geometrical interpretation
- Symmetric positive (semi)definite matrices play a very important role in machine learning and optimization
Matrix inversion lemma

- Woodsbury identity: when A and D are invertible
  \[(A + BDC^T)^{-1} = A^{-1} - A^{-1}C(D^{-1} + CA^{-1}B^T)^{-1}B^TA^{-1}\]
- Proof: multiply the matrix on both sides
- important special case
  - B=C=z, a vector, D=I
    \[(A + zz^T)^{-1} = A^{-1} - (A^{-1}zz^TA^{-1})/(1 + z^TA^{-1}z)\]
  - B=-C=z, a vector, D=I
    \[(A - zz^T)^{-1} = A^{-1} + (A^{-1}zz^TA^{-1})/(1 - z^TA^{-1}z)\]
- caching \(A^{-1}\) and computing the inversion recursively, typical inversion will take \(O(n^3)\), while this special case it is \(O(n)\)
CSI 436/536
Introduction to Machine Learning

Review of multivariate calculus (1)

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York