Georgia Tech College of Computing

CS4641 Spring 2025

Linear Algebra Revisit

Bo Dai
School of CSE, Georgia Tech
bodai@cc.gatech.edu

Basic / Prerequisites

- Probability
  - distributions, densities, marginalization, conditioning
- Statistics
  - mean, variance, maximum likelihood estimation
- Linear Algebra and Optimization
  - vector, matrix, multiplication, inversion, eigen-value decomposition
- Coding skills

Machine Learning for Apartment Hunting

- Suppose you are to move to Atlanta
- And you want to find the most reasonably priced apartment satisfying your needs:

$$\text{monthly rent} = \theta_1(\text{living area}) + \theta_2(\#\ \text{bedroom})$$

Living area (ft^2)	# bedroom	Monthly rent ($)
230	1	900
506	2	1800
433	2	1500
190	1	800
...		
150	1	?
270	1.5	?

Linear Regression Model

- Assume y is a linear function of x (features) plus noise ϵ

$$\text{monthly rent} = \theta_1(\text{living area}) + \theta_2(\#\ \text{bedroom})$$

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

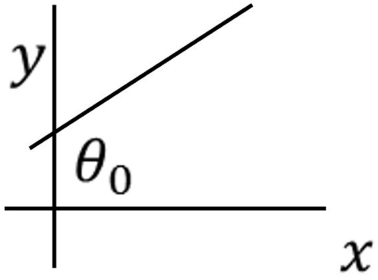
where ϵ is an error model as Gaussian $N(0, \sigma^2)$

Probability

- Let $\theta = (\theta_0, \theta_1, \ldots, \theta_n)^\top$, and augment data by one dimension

Linear algebra

$$x \leftarrow (1, x)^\top$$



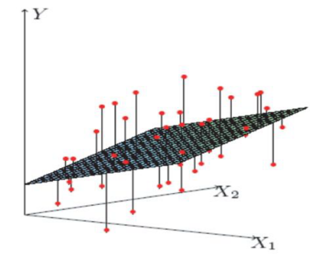
Then $y = \theta^\top x + \epsilon$

Linear algebra

Probabilistic Interpretation

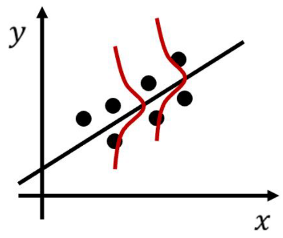
- Assume y is a linear in x plus noise ϵ

$$y = \theta^\top x + \epsilon$$



- Assume ϵ follows a Gaussian $N(0, \sigma)$

$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^i - \theta^\top x^i\right)^2}{2\sigma^2}\right)$$



- By independence assumption, likelihood is

$$L(\theta) = \prod_i^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{\sum_i^m \left(y^i - \theta^\top x^i\right)^2}{2\sigma^2}\right)$$

Probability

Probabilistic Interpretation

- Hence the log-likelihood is:

$$\log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i}^{m} \left(y^i - \theta^\top x^i\right)^2$$

- Least Mean Square (LMS)

$$LMS: \quad \frac{1}{m} \sum_{i}^{m} \left(y^i - \theta^\top x^i\right)^2$$

Statistics

- How to make it work in real data?

Algorithms
Programming

Revisit of Linear Algebra

- Basics
- Dot and Vector Products
- Identity, Diagonal and Orthogonal Matrices
- Trace
- Norms
- Inverse of a matrix
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Linear Algebra Basics - I

- Linear algebra provides a way of compactly representing and operating on sets of linear equations
  $4x_1 - 5x_2 = -13 \qquad\qquad - 2x_1 + 3x_2 = 9$
  can be written in the form of $Ax = b$
  $A = \begin{bmatrix} 4 & 5 \\ -2 & 3 \end{bmatrix}$ $b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$
- $A \in \mathbb{R}^{m \times n}$ denotes a matrix with m rows and n columns, where elements belong to real numbers.
- $x \in \mathbb{R}^n$ denotes a vector with n real entries. By convention an n dimensional vector is often thought as a matrix with n rows and 1 column.

Linear Algebra Basics - II

- Transpose of a matrix results from flipping the rows and columns. Given $A \in \mathbb{R}^{m \times n}$, transpose is $A^{\top} \in \mathbb{R}^{n \times m}$
- For each element of the matrix, the transpose can be written as $A^{\top}{}_{ij} = A_{ji}$
- The following properties of the transposes are easily verified

$$(A^{\top})^{\top} = A$$

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

$$(A + B)^{\top} = A^{\top} + B^{\top}$$

- A square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^{\top}$ and it is anti-symmetric if $A = -A^{\top}$. Thus each matrix can be written as a sum of symmetric and anti-symmetric matrices. $C = \frac{1}{2}(C + C^{\top}) + \frac{1}{2}(C - C^{\top})$

Vector and Matrix Multiplication - I

- The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is given by $C \in \mathbb{R}^{m \times p}$, where $C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$
- Given two vectors $\mathrm{x}, \mathrm{y} \in \mathbb{R}^{\mathrm{n}}$, the term $x^{\top} y$ (also $x \cdot y$) is called the *inner product* or *dot product* of the vectors, and is a real number given by $\sum_{i=1}^{n} x_i y_i$. For example,

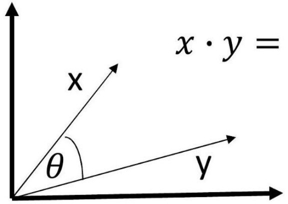
$$x^{\top} y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \sum_{i=1}^{3} x_i y_i$$

- Given two vectors $x \in \mathbb{R}^{n}, y \in \mathbb{R}^{m}$, the term $x y^{\top}$ is called the *outer product* of the vectors, and is a matrix given by $\left(x_i y_j\right)^{\top} = x_i y_j$. For example,

Vector and Matrix Multiplication - II

$$xy^\top = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} [y_1 \quad y_2 \quad y_3] = \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix}$$

- The dot product also has a geometrical interpretation, for vectors in $x, y \in \mathbb{R}^2$ with angle θ between them



$$x \cdot y = |x||y| \cos\theta$$

which leads to use of dot product for testing orthogonality, getting the Euclidean norm of a vector, and scalar projections.

Norms - I

- Norm of a vector $\|x\|$ is informally a measure of the "length" of a vector
- More formally, a norm is any function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies
  - For all $x \in \mathbb{R}^n, f(x) \geq 0$ (non-negativity)
  - $f(x) = 0$ is and only if $x = 0$ (definiteness)
  - For $x \in \mathbb{R}^n, t \in \mathbb{R}, f(tx) = |t| f(x)$ (homogeneity)
  - For all $x, y \in \mathbb{R}^n, f(x + y) \leq f(x) + f(y)$ (triangle inequality)

Norms - II

- Common norms used in machine learning are
  - ℓ_2 norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
  - ℓ_1 norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
  - ℓ_∞ norm: $\|x\|_\infty = \max_i |x_i|$
- All norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$:

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

- Norms can be defined for matrices, such as the Frobenius norm.

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)}$$

Trace of a Matrix

- The trace of a matrix $A \in \mathbb{R}^{n \times n}$, denoted as $\mathbf{tr}(\boldsymbol{A})$, is the sum of the diagonal elements in the matrix

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

- The trace has the following properties
  - For $A \in \mathbb{R}^{n \times n}, \operatorname{tr}(A) = \operatorname{tr} A^{\top}$
  - For $A, B \in \mathbb{R}^{n \times n}, \operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B)$
  - For $A \in \mathbb{R}^{n \times n}, \mathrm{t} \in \mathbb{R}, \operatorname{tr}(tA) = t \cdot \operatorname{tr}(A)$
  - For A, B, C such that ABC is a square matrix $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$
- The trace of a matrix helps us easily compute norms and eigenvalues of matrices as we will see later

Identity, Diagonal and Orthogonal Matrices

- The identity matrix, denoted by $I \in \mathbb{R}^{n \times n}$ is a square matrix with ones on the diagonal and zeros everywhere else
- A diagonal matrix is matrix where all non-diagonal matrices are 0 . This is typically denoted as $\mathrm{D} = \mathbf{diag}(d_1, d_2, d_3, \ldots, d_n)$
- Two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x.y = 0$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized
- It follows from orthogonality and normality that
  - $U^\top U = \mathrm{I} = U U^\top$
  - $\|Ux\|_2 = \|x\|_2$

Inverse of a Matrix

- The inverse of a square matrix $A \in \mathbb{R}^{n\times n}$ is denoted A^{-1} and is the unique matrix such that $A^{-1}A = I = AA^{-1}$
- For some square matrices A^{-1} may not exist, and we say that A is ***singular or non-invertible***. In order for A to have an inverse, A must be ***full rank***.
- For non-square matrices the inverse, denoted by A^{+},is given by $A^{+} = (A^{\top}A)^{-1}\, \mathrm{A}^{\top}$ called the ***pseudo inverse***

Eigenvalues and Eigenvectors - I

- 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 an eigenvector if

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

- Intuitively this means that upon multiplying the matrix A with a vector x , we get the same vector, but scaled by a parameter λ
- Geometrically, we are transforming the matrix A from its original orthonormal basis/co-ordinates to a new set of orthonormal basis x with magnitude as λ

Eigenvalues and Eigenvectors - II

- All the eigenvectors can be written together as $AX = X\Lambda$ where the diagonals of X are the eigenvectors of A, and Λ is a diagonal matrix whose elements are eigenvalues of A
- If the eigenvectors of A are invertible, then $A = X\Lambda X^{-1}$
- There are several properties of eigenvalues and eigenvectors
  - $\mathrm{Tr}(A) = \sum_{i=1}^{n} \lambda_i$
  - $|A| = \prod_{i=1}^{n} \lambda_i$
  - Rank of A is the number of non-zero eigenvalues of A
  - If A is non-singular then $\frac{1}{\lambda_i}$ are the eigenvalues of A^{-1}
  - The eigenvalues of a diagonal matrix are the diagonal elements of the matrix itself!

Eigenvalues and Eigenvectors - III

- For a symmetric matrix A, it can be shown that eigenvalues are real and the eigenvectors are orthonormal. Thus it can be represented as $U\Lambda U^\top$
- Considering quadratic form of A ,

$$x^\top Ax = x^\top U\Lambda U^\top x = y^\top \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \qquad \text{(where } y = U^\top x \text{)}$$

- Since y_i^2 is always positive the sign of the expression always depends on λ_i. If $\lambda_i > \mathbf{0}$ then the matrix A is positive definite, if $\lambda_i \geq \mathbf{0}$ then the matrix A is positive semidefinite

Singular Value Decomposition

- Singular value decomposition, known as SVD, is a factorization of a real matrix with applications in calculating pseudo-inverse, rank, solving linear equations, and many others.
- For a matrix $M \in \mathbb{R}^{m \times n}$ assume $n \leq m$
  - $M = U\Sigma V^\top$ where $U \in \mathbb{R}^{m \times m}, V^\top \in \mathbb{R}^{n \times n}, \Sigma \in \mathbb{R}^{m \times n}$
  - The m columns of U, and the n columns of V are called the left and right singular vectors of M. The diagonal elements of Σ, Σ_{ii} are known as the singular values of M.
  - Let v be the i^{th} column of V, and u be the i^{th} column of U, and σ be the i^{th} diagonal element of Σ

$$Mv = \sigma u \text{ and } M^\top u = \sigma v$$

Singular Value Decomposition - II

- Singular value decomposition is related to eigenvalue decomposition
  - Suppose $X = [x_1 - u \quad x_2 - u \ldots \quad x_m - u] \in \mathbb{R}^{m \times n}$
  - Then covariance matrix is $C = \frac{1}{m} X X^\top$
  - Starting from singular vector pair
    - $M^\top u = \sigma v$
      $\Rightarrow M M^\top \mathrm{u} = \sigma M v$
      $\Rightarrow M M^\top \mathrm{u} = \sigma^2 u$
      $\Rightarrow C u = \lambda u$

Matrix Calculus

- For a vector $x, b \in \mathbb{R}^n$, let $f(x) = b^\top x$, then $\nabla_x b^\top x$ is equal to b
  - $\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$
- Now for a quadratic function, $f(x) = x^\top A x$, with $A \in \mathbb{S}^n$, $\frac{\partial f(x)}{\partial x_k} = 2Ax$
  - $$\begin{aligned} \frac{\partial f(x)}{\partial xk} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{i=1}^{n} A_{ij} x_i x_j \\ &= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k \\ &= 2 \sum_{i=1}^{n} A_{ki} x_i \end{aligned}$$
- Let $f(X) = X^{-1}$, then $\partial(X^{-1}) = -X^{-1}(\partial X)X^{-1}$

References for self study

Resources for review of material

- Linear Algebra Review and Reference by Zico Kotler
- Matrix Cookbook by KB Peterson

Back to Apartment Hunting

- Given m data points, find θ that minimizes the mean square error

$$\hat{\theta} = argmin_{\theta}\, L(\theta) = \frac{1}{m}\sum_{i}^{m}\left(y^i - \theta^{\top}x^i\right)^2$$

Optimization

Statistics

- Set gradient to 0 and find parameter

Optimization

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m}\sum_{i}^{m}\left(y^i - \theta^{\top}x^i\right)x^i = 0$$

$$\Leftrightarrow -\frac{2}{m}\sum_{i}^{m} y^i x^i + \frac{2}{m}\sum_{i}^{m} x^i {x^i}^{\top}\theta = 0$$

Linear algebra

Statistics

Statistics

Optimization for LMS

- Define $X = (x^1, x^2, \dots x^m), y = (y^1, y^2, \dots, y^m)^\top$, gradient becomes

Linear algebra

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} Xy + \frac{2}{m} XX^\top \theta$$

Linear algebra

$$\Rightarrow \hat{\theta} = (XX^\top)^{-1} Xy$$

Algorithms Programming

- Matrix inversion in $\hat{\theta} = (XX^\top)^{-1} Xy$ expensive to compute
  - Gradient descent

$$\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{m} \sum_i^m \left(y^i - \hat{\theta}^{t\top} x^i \right) x^i$$

Optimization

Registration

- Friday is the registration deadline.

- If you decide to drop the course, please do so ASAP so that other people on the waitlist have time to register!

- See you next week!

Q&A