10-605 Machine Learning With Large Datasets

Linear Algebra Review

Fall 2022

Carnegie Mellon University

Linear Equations

- Set of linear equations (two equations, two unknowns)

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \\ -2x_1 + 3x_2 &= 9 \end{aligned}$$

- Can represent compactly in matrix notation

$$Ax = b$$

with

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

Basic notation

- A matrix with real-valued entires, m rows and n columns:

$$A \in \mathbb{R}^{m \times n}$$

  A_{ij} denotes the entry in the ith row and jth column

- A (column) vector with n real-valued entries

$$x \in \mathbb{R}^n$$

  x_i denotes the ith entry

Transpose

- The transpose operator $A^\top$ switches the rows and columns of a matrix:
$$A_{ij} = (A^\top)_{ji}$$
- For a vector $x \in \mathbb{R}^{n}, x^\top \in \mathbb{R}^{1 \times n}$ represents a row vector
- Properties
    - $(A^\top)^\top = A$
    - $(A + B)^\top = A^\top + B^\top$
    - $(AB)^\top = B^\top A^\top$
    - $(A^{-1})^\top = (A^\top)^{-1}$

Elements of a matrix

- Can write a matrix in terms of its columns:

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

- Note: a_i here corresponds to an entire vector $a_i \in \mathbb{R}^m$, not an element of a vector

Elements of a matrix

- Similarly, can write in terms of rows:

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

- Note: $a_i \in \mathbb{R}^n$ here and $a_i \in \mathbb{R}^m$ on previous slide are not the same vector

Matrix addition

- For two matrices of the same size and type, $A, B \in \mathbb{R}^{m\times n}$, addition is just the sum of the corresponding elements:

$$A + B = C \in \mathbb{R}^{m\times n} \iff C_{ij} = A_{ij} + B_{ij}$$

- Addition is undefined for matrices of different sizes

Matrix multiplication

- For two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}$, their product is:

$$AB = C \in R^{m\times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- Multiplication is undefined with the number of columns in A doesn't equal the number of rows in B (unless in case: cA where $c \in \mathbb{R}$ is a scalar)
- Special cases:
    - Inner product: $x, y \in \mathbb{R}^n, \quad x^\top y \in \mathbb{R} = \sum_{i=1}^{n} x_i y_i$
    - Matrix-vector product: $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n \iff Ax \in \mathbb{R}^m$

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix}, \, Ax \in \mathbb{R}^m = \sum_{i=1}^{n} a_i x_i$$

Important properties

- Associative: $A(BC) = (AB)C$
- Distributive: $A(B + C) = AB + AC$
- *Not* Commutative: $AB \neq BA$
- Transpose: $(AB)^\top = B^\top A^\top$

Special matrices

- **Identity matrix**:

$$I_n \in \mathbb{R}^{n \times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

  Has the property that for any $A \in \mathbb{R}^{m \times n}$

$$AI_n = A = I_m A$$

- **Ones vector**: $1 \in \mathbb{R}^n = \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^\top$. Useful, e.g., to represent sums: $a \in \mathbb{R}^n$, $1^\top a = \sum_{i=1}^n a_i$
- **Symmetric matrix**: $A \in \mathbb{R}^{n \times n}$ where $A = A^\top$
- **Diagonal matrix**: $diag(d) \in \mathbb{R}^{n \times n} = dI_n$

Norms

- A vector norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ with
  - $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$
  - $f(ax) = |a| f(x)$ for $a \in \mathbb{R}$
  - $f(x + y) \leq f(x) + f(y)$
- e.g., ℓ_2 norm: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$
- e.g., ℓ_1 norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$

Matrix Inverse

- The *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that:

$$AA^{-1} = A^{-1}A = I_n$$

- If A^{-1} exists, then A is called invertible or non-singular
- Otherwise, A is called singular
- A matrix A is invertible iff $\det(A) \neq 0$

Eigenvalues and Eigenvectors

- For $A \in \mathbb{R}^{n\times n}, \lambda$ is an eigenvalue and $x \neq 0$ is an eigenvector if:

$$Ax = \lambda x$$

- $\det(A - \lambda I_n)$ is called the **characteristic equation** of the matrix A
- Eigenvalues of A are the roots of the characteristic equation
- Associated eigenvectors of A are non-zero solutions to the equation $(A - \lambda I_n)x = 0$.

Singular value decomposition (SVD)

Every matrix has the following decomposition:

SVD

Let $X \in \mathbb{R}^{n \times m}$ then

$$X = U \Sigma V^\top,$$

where $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{m \times m}$ are orthogonal matrices (*i.e.* $U^\top = U^{-1}$) and $\Sigma \in \mathbb{R}^{n \times m}$ is a diagonal matrix with *singular values* of X denoted by σ_i appearing by non-increasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$.

- The square singular values of X are the eigenvalues of the matrix $XX^\top$ or $X^\top X$, i.e., $\sigma_i(X) = \sqrt{\lambda_i(XX^\top)} = \sqrt{\lambda_i(X^\top X)}$

PCA by Covariance Matrix

Steps to perform PCA using covariance matrix

- X is $n \times k$ raw data
- $Z = X\ P$ is $n \times r$ (reduced representation)
- P is $k \times r$ (columns contain r principal components)
- $C_X = U \Lambda U^T$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

  where Eigen values of C_X are ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$
- U is $k \times k$ (columns are eigenvectors)
- Choose P as the first r columns of U
- This captures the r directions of maximum variance

PCA by SVD

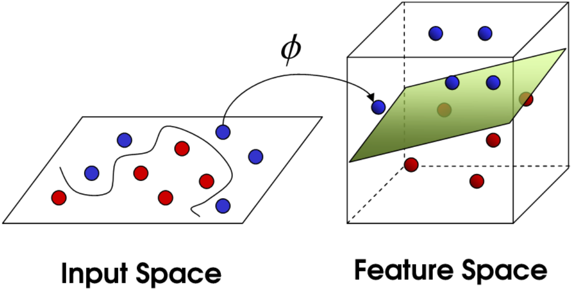
Proof:

- $C_X = \frac{1}{n}\mathrm{X}^T X = \frac{1}{n}\mathrm{WDW}^T$ (covariance matrix is symmetric so it can be diagonalized)
- Apply SVD on X, $X = U\Sigma V^T$
- Then we have $\frac{1}{n}\mathrm{X}^T X = \frac{1}{n}(\mathrm{U}\Sigma V^T)^T(\mathrm{U}\ \Sigma V^T) = \frac{1}{n}(\mathrm{V}\Sigma U^T)(U\Sigma V^T)$
- Since U is orthogonal matrix ($U^T U = I$) then we have
- $C_X = \frac{1}{n}\mathrm{X}^T X = \frac{1}{n}\mathrm{V}\ \Sigma^2 V^T$

SVD

The singular values are related to the eigenvalues of covariance matrix via $\lambda_i = \sigma_i^2$. Columns $U\Sigma$ are reduced representation ("scores").

Kernel



- A Kernel maps an input vector in a lower dimension to a feature vector in higher dimensional space
- This is often used in SVMs as well as other machine learning methods to make the data linearly separable