10-701/15-781 Recitation : Linear Algebra Review (based on notes written by Jing Xiang)

Manojit Nandi

Outline

Linear Algebra

General Properties

A Vector Space V is a space $\mathcal{X}$ defined over some field $\mathcal{F}$ that satisfies the following properties.

For all $\mathbf{x}, \mathbf{y} \in V$ and for all $\alpha \in \mathbb{R}$

1. $\mathbf{x} + \mathbf{y} \in V$ (Closure under addition)

2. $\alpha\mathbf{x} \in V$ (Closure under scalar multiplication)

A vector space V defined over the field $\mathbb{R}^{mxn}$ is an mXn matrix with m rows and n columns.

Examples:

3 x 3 matrix:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

2 x 3 matrix:

$$\begin{pmatrix} 1 & 2 & 3 \\ 12 & 17 & 23 \end{pmatrix}$$

Matrix Addition: If A is a mXn matrix, and B is a mXn matrix, then $C = A + B$ is a mXn matrix s.t.,

$$C_{ij} = A_{ij} + B_{ij}$$

Matrix Multiplication: If A is a mXn matrix, and B is a nXp matrix, then $C = AB$ is a mXp matrix s.t.,

$$C_{ij} = \sum_{k=1}^{n} A_{ik} * B_{kp}$$

Matrix Multiplication has the following properties:

- Associativity: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

Matrix multiplication is not commutative. $\mathbf{AB} \neq \mathbf{BA}$, except in certain cases.

Matrix Operations

The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{mxn}$, is $\mathbf{A}^T \in \mathbb{R}^{nxm}$ where the entries of $\mathbf{A}^T$ are given by:

$$(\mathbf{A}^T)_{ij} = \mathbf{A}_{ji}$$

Properties of the transpose operator:

- $(\mathbf{A} + \mathbf{B})^T = \mathbf{B}^T + \mathbf{A}^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$
- $(\mathbf{A}^T)^T = \mathbf{A}$

A set of vectors $x_1, x_2, ...x_n \subset \mathbb{R}^m$ are linearly independent if no vector can be represented as a linear combination of the remaining vectors. The rank of a matrix is the cardinality of the largest subset of the columns of some matrix A that is a linearly independent set.

A matrix $\mathbf{A} \in \mathbb{R}^{mxn}$ is full rank, if $rank(\mathbf{A}) = min(m, n)$.

A norm of a vector $||\mathbf{x}||$ reflects the magnitude of the vector.
Commonly used norms fall into a family of norms called the L_p norm: $|||\mathbf{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

Examples:

- Eucldiean norm (L_2): $|||\mathbf{x}||_2 = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}}$
- L_1 norm: $|||\mathbf{x}||_1 = (\sum_{i=1}^{n} |x_i|^1)^{\frac{1}{1}} = \sum_{i=1}^{n} |x_i|$
- L_0 norm: $|||\mathbf{x}||_0 = \sum_{i=1}^{n} \mathbf{1}(x_i \neq 0)$
- L_∞ norm: $|||\mathbf{x}||_\infty = \max_i |x_i|$

Suppose we have a vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, and we want to represent it using a single point $\mathbf{s} = (s, s, ..., s)$ such that $s \in \mathbb{R}$. Therefore we want to choose the $\mathbf{s}$ that minimizes $||\mathbf{x} - \mathbf{s}|$.
However, which we chooses effects the value of $\mathbf{s}$.

- Eucldiean norm (L_2): $|||\mathbf{x} - \mathbf{s}||_2 = (\sum_{i=1}^{n} |x_i - s|^2)^{\frac{1}{2}}$
- L_1 norm: $|||\mathbf{x} - \mathbf{s}||_1 = (\sum_{i=1}^{n} |x_i|^1)^{\frac{1}{1}} = \sum_{i=1}^{n} |x_i - s|$
- L_0 norm: $|||\mathbf{x} - \mathbf{s}||_0 = \sum_{i=1}^{n} \mathbf{1}(x_i - s \neq 0)$

Suppose we have a vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, and we want to represent it using a single point $\mathbf{s} = (s, s, ..., s)$ such that $s \in \mathbb{R}$. Therefore we want to choose the $\mathbf{s}$ that minimizes $||\mathbf{x} - \mathbf{s}|$.
However, which we chooses effects the value of $\mathbf{s}$.

- Eucldiean norm (L_2): $|||\mathbf{x} - \mathbf{s}||_2 = (\sum_{i=1}^{n} |x_i - s|^2)^{\frac{1}{2}} =$ Mean
- L_1 norm: $|||\mathbf{x} - s||_1 = \sum_{i=1}^{n} |x_i - s| =$ Median
- L_0 norm: $|||\mathbf{x} - s||_0 = \sum_{i=1}^{n} \mathbf{1}(x_i - s \neq 0) =$ Mode

A vector is said to be normalized if $||\mathbf{x}|| = 1$.

The trace of a square matrix, Tr($\mathbf{A}$) $\mathbf{A} \in \mathbb{R}^{nxn}$ is the sum of the diagonal elements.

$$Tr(\mathbf{A}) = \sum_{i=1}^{n} \mathbf{A}_{ii}$$

Properties of the Trace:

- $\mathbf{A} \in \mathbb{R}^{nxn}, Tr(\mathbf{A}) = Tr(\mathbf{A}^T)$
- $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{nxn}, Tr(\mathbf{A} + \mathbf{B}) = Tr(\mathbf{A}) + Tr(\mathbf{B})$
- $\mathbf{A} \in \mathbb{R}^{nxn}, c \in \mathbb{R}, Tr(c\mathbf{A}) = cTr(\mathbf{A})$
- $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{nxn}, Tr(\mathbf{A}^T\mathbf{B}) = Tr(\mathbf{A}\mathbf{B}^T)$
- $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{nxn}, Tr(\mathbf{ABC}) = Tr(\mathbf{BCA}) = Tr(\mathbf{CBA})$

The inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{nxn}$, written $\mathbf{A}^{-1}$ is defined such that:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

If $\mathbf{A}^{-1}$ exists,t he matrix is said to be nonsingular, otherwise it is singular.

Properties of the Inverse:

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^{T} = (\mathbf{A}^{T})^{-1}$

Inner Products and Orthogonal Matrices

An Inner Product Space is a vector space V equipped with a operation called the inner product, denoted $< \cdot, \cdot >$, that maps $< \cdot, \cdot >: VxV \rightarrow \mathcal{F}$.

For $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ (when V is defined over the real numbers) and a constant $a, b \in \mathbb{R}$, the inner product satisfies the following properties:

- $< \mathbf{x}, \mathbf{y} >=< \mathbf{y}, \mathbf{x} >$ [Symmetry]
- $< a\mathbf{x} + b\mathbf{z}, \mathbf{y} >= a < \mathbf{x}, \mathbf{y} > + b < \mathbf{z}, \mathbf{y} >$ [Linearity]
- $< \mathbf{x}, \mathbf{x} >\geq 0$ [Positive Semi-Definite]

For the vector space V defined over $\mathbb{R}^n$, given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the inner product is defined as: $< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y}$.

As a result,

$$< \mathbf{x}, \mathbf{x} >= \sum_{i=1}^{n} x_i x_i = \sum_{i=1}^{n} x_i^2 = ||\mathbf{x}||_2$$

Two vectors $\mathbf{x}, \mathbf{y}$ are said to orthogonal if $\mathbf{x}^T\mathbf{y} = 0$.

A square matrix $\mathbf{A} \in \mathbb{R}^{nxn}$ is orthogonal if all its normalized colums are orthogonal to one another.

If $\mathbf{U}$ is an orthogonal matrix, $\mathbf{U}^T = \mathbf{U}^{-1}$, then $\mathbf{U}^T\mathbf{U} = I$.

If $\mathbf{U}$ is a square matrix, then $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = I$

Eigenvalue and Eigenvectors

Given a square matrix $\mathbf{A} \in \mathbb{R}^{nxn}, \lambda \in \mathbf{C}, \mathbf{x} \in \mathbf{C}$, λ and $\mathbf{x}$ are the eigenvalue and eigenvector of $\mathbf{A}$ respectively iff

$$\mathbf{Ax} = \lambda\mathbf{x}, \mathbf{x} \neq 0 \Leftrightarrow (\mathbf{X} - \lambda\mathbf{I})\mathbf{x} = 0$$

We often use the determinant to expand this expression into the characteristic polynomial in terms of λ and then find the roots of the characteristic polynomial to find the eigenvalues.

Now that we have eigenvalues and eigenvectors, some important properties of matrices relating to eigenvalues are:

- $Tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$
- $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$
- $Rank(\mathbf{A}) =$ number of non-zero eigenvalues of $\mathbf{A}$

A square matrix $\mathbf{A}$ is diagonalizable if there exists an invertible matrix $\mathbf{X}$ and a diagonal matrix Λ such that $\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1}$. When the columns of $\mathbf{X}$ are the eigenvectors of $\mathbf{A}$, and the eigenvalues of $\mathbf{A}$ form the diagonal entries of Λ, this corresponds to the eigendecomposition.

One nice property of this is:
$\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1}$
$\mathbf{A}^2 = (\mathbf{X}\Lambda\mathbf{X}^{-1})(\mathbf{X}\Lambda\mathbf{X}^{-1}) = \mathbf{X}\Lambda\mathbf{I}\Lambda\mathbf{X}^{-1} = \mathbf{X}\Lambda^2\mathbf{X}^{-1}$

Furthermore, the eigenvectors of $\mathbf{A}$ are orthonormal, so $\mathbf{X}$ is orthonormal, meaning $\mathbf{X}^{-1} = \mathbf{X}^T$. Therefore:
$\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^T$

Matrix Calculus

Gradients

Let $f : \mathbb{R}^{mxn} \rightarrow \mathbb{R}$, the gradient of f with respect to the matrix $\mathbf{A} \in \mathbb{R}^{mxn}$ is an mxn matrix of the partial derivatives of f with respect to $\mathbf{A}$.

$$(\nabla_A f(\mathbf{A})) = \begin{pmatrix} \frac{\partial f(\mathbf{A})}{\partial A_{11}} & \frac{\partial f(\mathbf{A})}{\partial A_{12}} & \frac{\partial f(\mathbf{A})}{\partial A_{13}} \\ \frac{\partial f(\mathbf{A})}{\partial A_{21}} & \frac{\partial f(\mathbf{A})}{\partial A_{22}} & \frac{\partial f(\mathbf{A})}{\partial A_{23}} \\ \frac{\partial f(\mathbf{A})}{\partial A_{31}} & \frac{\partial f(\mathbf{A})}{\partial A_{32}} & \frac{\partial f(\mathbf{A})}{\partial A_{33}} \end{pmatrix}$$

As a collorary, when we have a vector **x** rather than a matrix **A**, the gradient of f with respect to the vector is:

$$(\nabla_x f(\mathbf{x})) = (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n})$$

Example, if $f(x, z) = x \sin(z) e^{-x}$, then the gradient of f with respect to the standard basis vectors $\mathbf{x} = (x, z)$ is

$$(\nabla_x f(\mathbf{x})) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial z}) = ((1 - x) \sin(z) e^{-x}, x e^{-x} \cos(z))$$

For vectors, the gradient represents the direction of steepest ascent.

Hessians

Let $f : \mathbb{R}^n \to \mathbb{R}$ (Note: This function takes is vector inputs, not matrix inputs). The Hessian of f with respect to a vector $\mathbf{x} \in \mathbb{R}^n$ is a nxn matrix of partial derivatives.

$$(\nabla_x^2 f(\mathbf{x}))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

$$(\nabla_x f(\mathbf{x})) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial^2 x_1 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_3} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_3} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_3 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_3 x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_3 x_3} \end{pmatrix}$$

The geometric interpretation of the Hessian is the curvature of a surface.

Derivatives

Let $\mathbf{x} \in \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^n$, and $\mathbf{y} = f(\mathbf{x})$. The derivative of $\mathbf{y}$ with respect to the vector $\mathbf{x}$ is an mxn matrix:

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \begin{pmatrix} \frac{dy_1}{dx_1} & \frac{dy_2}{dx_1} & .... & \frac{dy_n}{dx_1} \\ \frac{dy_1}{dx_2} & \frac{dy_2}{dx_2} & .... & \frac{dy_n}{dx_2} \\ ... & .... & .... & .... \\ \frac{dy_1}{dx_m} & \frac{dy_2}{dx_m} & .... & \frac{dy_n}{dx_m} \end{pmatrix}$$

Let $\mathbf{a}$ be some arbitrary vector. Some common matrix derivatives are:

- $\frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial\mathbf{x}} = \mathbf{a}$
- $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = \mathbf{A}\mathbf{x}\frac{\partial\mathbf{x}^T}{\partial\mathbf{x}} + \mathbf{x}^T\mathbf{A}\frac{\partial\mathbf{x}}{\partial\mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$
- $\frac{\partial(\mathbf{a}^T\mathbf{X}\mathbf{b})}{\partial\mathbf{X}} = \mathbf{a}\mathbf{b}^T$

References

Zico Kolter's Linear Algebra Reiw and Reference:
htpp://cs229.stanford.edu/section/cs229-linalg.pdf

The Matrix Cookbook:
http://orion.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf

Randall J. Barne's Matrix Differentiation Notes:
http://www.atmos.washington.edu/ dennis/MatrixCalculus.pdf

Questions?