07-280 Notes
Math Background

Carnegie Mellon University
Machine Learning Department

Contents

1 Linear algebra

Most of the linear algebra listed here should be prerequisite material for you. The exceptions might be vector and matrix norms and any notational changes.

1.1 Notation

Matrix notation: $A \in \mathbb{R}^{N \times M}$:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}$$

Summation notation:

Matrix multiplication with $A \in \mathbb{R}^{M \times K}, B \in \mathbb{R}^{K \times N}, C \in \mathbb{R}^{M \times N}$. If $C = AB$, then $C_{i,j} = \sum_{k=1}^{K} a_{i,k} b_{k,j}$. The number of columns in A must match the number of rows in B. C will have

the same number of rows as A and the same number of columns as B.

Vector notation: $\mathbf{v} \in \mathbb{R}^N$, in this course, we'll assume all vectors are column vectors unless specified otherwise:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$$

Examples:

$\mathbf{u} \in \mathbb{R}^{M \times 1}$	Column vector of length M
$\mathbf{v} \in \mathbb{R}^{1 \times M}$	Row vector of length M
$\mathbf{z} \in \mathbb{R}^M$	Ambiguous. In this course, assume column vector unless stated otherwise

1.2 Linear systems of equations

Consider matrix $A \in \mathbb{R}^{N \times M}$ and vectors $\mathbf{v} \in \mathbb{R}^M$ and $\mathbf{u} \in \mathbb{R}^N$. The following is a linear system of equations with N equations and M unknowns, $v_1, \ldots, v_M$:

$$\begin{aligned} \mathbf{u} &= A\mathbf{v} \\ u_1 &= a_{1,1}v_1 + a_{1,2}v_2 + \cdots + a_{1,M}v_M \\ u_2 &= a_{2,1}v_1 + a_{2,2}v_2 + \cdots + a_{2,M}v_M \\ &\vdots \\ u_1 &= a_{N,1}v_1 + a_{N,2}v_2 + \cdots + a_{N,M}v_M \end{aligned}$$

underdetermined: If there are fewer equations than variables, the system is underdetermined and cannot have exactly 1 solution, it must have either infinitely many or no solutions.

overdetermined: A system with more equations than variables. An overdetermined system may have 1 solution, 0 solutions, or infinitely many solutions.

inconsistent: when the system of equations does not have a solution.

consistent: when the system of equations has at least one solution.

1.3 Vectors

dot product: $\mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + ... + a_Mb_M$ for two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^M$. It is the sum of the products of corresponding entries of the two vectors.

inner product (more general than dot product): is a way to multiply vectors, resulting in a scalar. Let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be vectors and let α be a scalar. Then, the inner product satisfies the following properties: $\setminus\ \langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle\ \setminus\ \langle \alpha\mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{v}, \mathbf{w} \rangle\ \setminus\ \langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle\ \setminus\ \langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if $\mathbf{v} = 0$

outer product: Outer product of vectors $\mathbf{u}$ and $\mathbf{v}$ is $Y = \mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T$, and $Y_{i,j} = u_i v_j$.

magnitude: Magnitude (length) of vector $\mathbf{u}$ is $|\mathbf{u}| = \|\mathbf{u}\|_2 = \left(\sum_i u_i^2\right)^{\frac{1}{2}}$.

L2 norm: Also known as Euclidean norm $\|\mathbf{v}\|_2 = \left(\sum_i v_i^2\right)^{\frac{1}{2}} = \left(\mathbf{v}^T\mathbf{v}\right)^{\frac{1}{2}}$

L1 norm: The sum of absolute values of the entries of the vector $\|\mathbf{v}\|_1 = \sum_i |v_i|$

L0 "norm": Number of non-zero entries in a vector (not technically a norm) $\|\mathbf{v}\|_0 = \sum_i |v_i|^0$, where 0^0 is defined as being equal to zero.

p-norm: $\|\mathbf{v}\|_p = \left(\sum_i |v_i|^p\right)^{\frac{1}{p}}$ (Only a norm for $p \geq 1$).

span: Set of all linear combinations of a set of vectors. For example, given a set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}, \mathbf{v_i} \in \mathbb{R}^M$, $span(\mathcal{S}) = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 \mid \alpha_i \in \mathbb{R}\}$. Span is an example of a vector space.

vector space: Span of a set of vectors is an example of a vector space. Vector space is a more general term for the result of combining a set of vectors with addition and scalar multiplication. For example, if we changed span to include multiplication by complex scalars, that would be a different vector space.

linearly dependent: A vector, $\mathbf{u}$, is linearly dependent with a set of vectors, $\mathcal{S} = \{\mathbf{v}_k\}_{k=1}^K$, if it is possible to represent $\mathbf{u}$ as a linear combination of vectors in $\mathcal{S}$. For example, if $\mathbf{u} = 3\mathbf{v}_1 - 3.5\mathbf{v}_3$, then $\mathbf{u}$ is linearly dependent with $\mathcal{S}$.

A set of vectors in linearly dependent if any of the vectors in the set can be represented by a linear combination of the remaining vectors in the set.

linearly independent: A vector, $\mathbf{u}$ is linearly independent from a set of vectors, $\mathcal{S} = \{\mathbf{v}_k\}_{k=1}^K$, if it is not possible to represent $\mathbf{u}$ as a linear combination of vectors in $\mathcal{S}$.

A set of vectors is linearly independent if no single vector in the set can be represented by a linear combination of the remaining vectors in the set.

1.4 Matrices

identity matrix: A matrix with all ones on the diagonal and zeros elsewhere. Represented as I, or more specifically I_N, where the N indicates that it is an $N \times N$ identity matrix. $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

matrix inverse: The inverse of matrix $A \in \mathbb{R}^{N \times N}$ is denoted A^{-1}. A^{-1} is also an $N \times N$ matrix. The inverse of a square, $N \times N$ matrix will exist if the matrix is full rank, i.e., the column rank and row rank is N. If the inverse of a square matrix exists then $A^{-1}A = AA^{-1} = I$, where I is the $N \times N$ identity matrix.

column rank of a matrix: the maximal number of linearly independent vectors among the column vectors in a given matrix. This is also the dimensionality of the vector space spanned by the column vectors of the matrix.

row rank of a matrix: the maximal number of linearly independent vectors among the row vectors in a given matrix. This is also the dimensionality of the vector space spanned by the row vectors of

the matrix.

rank: the row rank and column rank of a matrix are always equal, so there are often just referred to as matrix "rank".

full rank: A matrix is full rank is the rank is equal to the minimum of its number of rows and number of columns. If a matrix is square and full rank then the inverse of that matrix exists.

singular matrix: A square matrix is singular if it is not full rank and thus it's inverse doesn't exist.

Frobenius norm of a matrix: Basically the L2 norm if we were to flatten the matrix into a vector. For matrix $A \in \mathbb{R}^{N \times M}$,

$$\|A\|_F = \left(\sum_{i=1}^{N} \sum_{j=1}^{M} {a_{i,j}}^2 \right)^{\frac{1}{2}}$$

$$\|A\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} {a_{i,j}}^2$$

2 Multivariate calculus

While you may not have explicitly learned multivariate calculus, if very much just builds on top of normal (scalar) calculus. In multivariate calculus, we are just shifting to working with functions that have multiple inputs variables and potentially multiple outputs.

2.1 Partial derivatives

A **partial derivative** is when we take the derivative of a function f with respect to one of its many input variables. Notation-wise, you'll see it written as $\frac{\partial}{\partial z} f(x, z)$ or $\frac{\partial f}{\partial z}$. (It could also be written as $f_z(x, z)$, but we won't use that in this course.)

When we take the partial derivative with respect to one variable, we just hold all other variables constant.

For example:

$$f(x, z) = 2x^3 z^5 \tag{1}$$
$$\frac{\partial f}{\partial x} = 6x^2 z^5 \tag{2}$$
$$\frac{\partial f}{\partial z} = 10x^3 z^4 \tag{3}$$
$$\tag{4}$$

You can think of linear algebra as having many individual variables. Take, for example, the L2 norm squared of $\mathbf{x} \in \mathbb{R}^3$:

$$f(\mathbf{x}) = \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_i {x_i}^2 = {x_1}^2 + {x_2}^2 + {x_3}^2 \tag{5}$$
$$f(x_1, x_2, x_3) = \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_i {x_i}^2 = {x_1}^2 + {x_2}^2 + {x_3}^2 \tag{6}$$
$$\frac{\partial f}{\partial x_1} = 2x_1 \tag{7}$$
$$\frac{\partial f}{\partial x_2} = 2x_2 \tag{8}$$
$$\frac{\partial f}{\partial x_3} = 2x_3 \tag{9}$$
$$\tag{10}$$

2.2 Gradients

Given a scalar function with vector input, $f : \mathbb{R}^M \to \mathbb{R}$, $f(\mathbf{x}) = f(x_1, ..., x_M)$, the **gradient** is a column vector where the i-th entry is the partial derivative of the function with respect to the i-th input entry in the input vector.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_M} \end{bmatrix}$$

We call this the gradient of f with respect to $\mathbf{x}$. The $\mathbf{x}$ in the subscript is redundant if $\mathbf{x}$ is the only argument to f and would typically be dropped, $\nabla f(\mathbf{x})$.

Using the same example from above, the L2 norm squared of $\mathbf{x} \in \mathbb{R}^3$:

$$f(\mathbf{x}) = \|\mathbf{x}\|^2 = \begin{bmatrix} {x_1}^2 \\ {x_2}^2 \\ {x_3}^2 \end{bmatrix} \tag{11}$$

$$\frac{\partial f}{\partial x_1} = 2x_1 \tag{12}$$

$$\frac{\partial f}{\partial x_2} = 2x_2 \tag{13}$$

$$\frac{\partial f}{\partial x_3} = 2x_3 \tag{14}$$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix} = 2\mathbf{x} \tag{15}$$

3 Optimization notation

We can formalize an optimization problem with the following form:

$$y^* = \min_{x \in \mathcal{X}} f(x)$$

where

- $f : \mathbb{R}^K \to \mathbb{R}$ is known as the objective function that we are trying to minimize
- $\mathcal{X} \subset \mathbb{R}^K$ is the set of feasible inputs that we are trying to minimize f over
- y^* is the smallest value of $f(x)$ for all possible values $x \in \mathcal{X}$ (which is why we have a min in the formulation)

For all possible values x in the set $\mathcal{X}$ and return **the** x **corresponding to** the output of $f(x)$ that has the smallest value (i.e. return the argument, not the value of the function):

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$$

For example, suppose that we wanted to minimize the objective function $f(x) = 3(x-5)^2 - 200$:

$$y^* = \min_{x \in \mathbb{R}} 3(x-5)^2 - 200 \tag{16}$$

$$= -200 \tag{17}$$

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}} 3(x-5)^2 - 200 \tag{18}$$

$$= 5 \tag{19}$$

$$\tag{20}$$

Plot for the above example:

```
# Plot for above example
def f(x):
    return 3*(x-5)**2 - 200
    x_grid = np.linspace(-10, 10, 100)
    y_grid = f(x_grid)
    plt.plot(x_grid, y_grid, '-');
```