Mathematics for Machine Learning

Prof Willie Brink

Applied Mathematics, Stellenbosch University

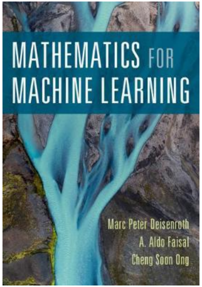
Lecture 1: Linear Algebra

Module information

Lecturer: Prof Willie Brink (wbrink@sun.ac.za)

Classes: Mondays and Wednesdays 9:30 to 12:30 in A403A

Textbook: https://mml-book.com



Assessment:

- Week 1: assignment 1 (15%), quiz 1 (15%)
- Week 2: assignment 2 (15%), quiz 2 (15%)
- Week 3: assignment 3 (15%), final test (25%)

SUNLearn: https://learn.sun.ac.za

Contents of the module

2.1 Systems of linear equations

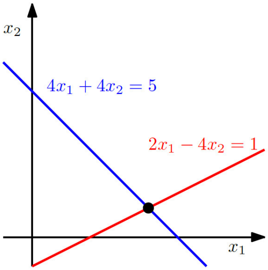
$$\begin{array}{ccccccccccc}
a_{1,1} & x_1 & + & a_{1,2} & x_2 & + & \dots & + & a_{1,n} & x_n & = & b_1 \\
a_{2,1} & x_1 & + & a_{2,2} & x_2 & + & \dots & + & a_{2,n} & x_n & = & b_2 \\
 & \vdots & & & & & & & & & \vdots & \\
a_{m,1} & x_1 & + & a_{m,2} & x_2 & + & \dots & + & a_{m,n} & x_n & = & b_m
\end{array}$$

$$\begin{bmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{m,1} \end{bmatrix} x_1 + \begin{bmatrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{m,2} \end{bmatrix} x_2 + \dots + \begin{bmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{m,n} \end{bmatrix} x_n = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \implies \boldsymbol{Ax} = \boldsymbol{b}$$

m equations, n unknowns

Can have no solution,
or exactly one solution,
or infinitely many.



2.2 Matrices

$\mathbb{R}^{m \times n}$ is the set of all real-valued matrices with m rows and n columns.

The sum of matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ is computed elementwise.

The product of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ is a matrix $\boldsymbol{C} \in \mathbb{R}^{m \times k}$ with $c_{i,j} = \sum_{\ell=1}^{n} a_{i,\ell} b_{\ell,j}$.

The Hadamard product of matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ is computed elementwise.

Matrix multiplication is associative and distributive, but in general not commutative ($\boldsymbol{AB} \neq \boldsymbol{BA}$).

With $\boldsymbol{I}_n$ the identity matrix in $\mathbb{R}^{n \times n}$, we have $\boldsymbol{I}_m \boldsymbol{A} = \boldsymbol{A}$ and $\boldsymbol{A} \boldsymbol{I}_n = \boldsymbol{A}$, $\forall \boldsymbol{A} \in \mathbb{R}^{m \times n}$.

The inverse of square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a matrix $\boldsymbol{B} = \boldsymbol{A}^{-1}$ such that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}_n$.

- if the inverse exists, $\boldsymbol{A}$ is called invertible / nonsingular (or regular)
- if the inverse doesn't exist, $\boldsymbol{A}$ is called noninvertible / singular

$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ and, in general, $(\boldsymbol{A} + \boldsymbol{B})^{-1} \neq \boldsymbol{A}^{-1} + \boldsymbol{B}^{-1}$.

The transpose of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is $\boldsymbol{B} = \boldsymbol{A}^{\mathsf{T}} \in \mathbb{R}^{n \times m}$, with $b_{i,j} = a_{j,i}$.

$(\boldsymbol{AB})^{\mathsf{T}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}$ and $(\boldsymbol{A} + \boldsymbol{B})^{\mathsf{T}} = \boldsymbol{A}^{\mathsf{T}} + \boldsymbol{B}^{\mathsf{T}}$.

A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric if $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}$.

If $\boldsymbol{A}$ is invertible, then so is $\boldsymbol{A}^{\mathsf{T}}$, and $(\boldsymbol{A}^{\mathsf{T}})^{-1} = (\boldsymbol{A}^{-1})^{\mathsf{T}} = \boldsymbol{A}^{-\mathsf{T}}$.

Scalar multiplication ($\lambda\boldsymbol{A}$) is calculated elementwise, and is associative and distributive.

2.3 Solving systems of linear equations

General approach to solve $\boldsymbol{Ax} = \boldsymbol{b}$:

1. find a particular solution to $\boldsymbol{Ax} = \boldsymbol{b}$
2. find all solutions to $\boldsymbol{Ax} = \boldsymbol{0}$
3. combine the solutions from steps 1 and 2

Gaussian elimination

Use elementary transformations that do not change the solution (row exchange, multiplying a row with nonzero constant, adding rows) to find a row-echelon form.

pivot: first nonzero element in a row from the left

staricase structure: every pivot is strictly to the right of the pivot above it

The reduced row echelon form, where every pivot is 1 and is the only nonzero entry in its column, eases steps 1 and 2 above.

1. Finding a particular solution to $\boldsymbol{Ax} = \boldsymbol{b}$:

   Write $[\,\boldsymbol{A} \mid \boldsymbol{b}\,]$ in reduced row-echelon form (RREF).

   Set free variables (not corresponding to pivots) to zero.

   Easily solve for the basic variables (corresponding to pivots).

2. Finding a general solution to $\boldsymbol{Ax} = \boldsymbol{0}$:

   Augment the RREF of $\boldsymbol{A}$ with rows of the form $[0 \cdots 0 \; -1 \; 0 \cdots 0]$ so that we have 1 or -1 on the diagonal.

   General solution: all linear combinations of the columns with -1 on the diagonal.

3. A general solution to $\boldsymbol{Ax} = \boldsymbol{b}$ will be the sum of steps 1 and 2.

Calculating the inverse of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$: the RREF of $[\,\boldsymbol{A} \mid \boldsymbol{I}_n\,]$ will be $[\,\boldsymbol{I}_n \mid \boldsymbol{A}^{-1}\,]$.

2.4 Vector spaces

A real-valued vector space $V = (\mathcal{V}, +, \cdot)$ consists of a set $\mathcal{V}$ and two operations

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V} \quad \textit{vector addition}$$

$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V} \quad \textit{scalar multiplication}$$

where $(\mathcal{V}, +)$ is an Abelian group* with neutral element $\mathbf{0}$

and $\forall \lambda, \psi \in \mathbb{R},\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$:

$$\lambda \cdot (\boldsymbol{x} + \boldsymbol{y}) = \lambda \cdot \boldsymbol{x} + \lambda \cdot \boldsymbol{y}$$
$$(\lambda + \psi) \cdot \boldsymbol{x} = \lambda \cdot \boldsymbol{x} + \psi \cdot \boldsymbol{x}$$
$$\lambda \cdot (\psi \cdot \boldsymbol{x}) = (\lambda\psi) \cdot \boldsymbol{x}$$

and the neutral element w.r.t. scalar multiplication is 1, such that $\forall \boldsymbol{x} \in \mathcal{V} : 1 \cdot \boldsymbol{x} = \boldsymbol{x}$.

* closed, associative, commutative, $\forall \boldsymbol{x} \in \mathcal{V} : \boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$, $\forall \boldsymbol{x} \in \mathcal{V}\ \exists \boldsymbol{y} \in \mathcal{V} : \boldsymbol{x} + \boldsymbol{y} = \mathbf{0}$

We will denote a vector space $(\mathcal{V}, +, \cdot)$ by V, and assume $+$ and $\cdot$ are the standard vector addition and scalar multiplication.

We'll often write $\boldsymbol{x} \in V$ to simplify notation.

We also often omit the dot in scalar multiplication: $\lambda \boldsymbol{x} = \lambda \cdot \boldsymbol{x}$

Vector subspaces

Let $V = (\mathcal{V}, +, \cdot)$ be a vector space, and $\mathcal{U} \subseteq \mathcal{V}$ with $\mathcal{U} \neq \emptyset$.

Then $U = (\mathcal{U}, +, \cdot)$ is a vector subspace of V if:

- $\mathcal{U}$ contains the neutral element w.r.t. vector addition: $\mathbf{0} \in \mathcal{U}$
- U is closed w.r.t. vector addition: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{U} \;:\; \boldsymbol{x} + \boldsymbol{y} \in \mathcal{U}$
- U is closed w.r.t. scalar multiplication: $\forall \lambda \in \mathbb{R},\ \boldsymbol{x} \in \mathcal{U} \;:\; \lambda \boldsymbol{x} \in \mathcal{U}$

2.5 Linear independence

A linear combination of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ in vector space V is any vector $\boldsymbol{v} \in V$ of the form

$$\boldsymbol{v} = \lambda_1 \boldsymbol{x}_1 + \ldots + \lambda_k \boldsymbol{x}_k \quad \text{with } \lambda_1, \ldots, \lambda_k \in \mathbb{R}.$$

A set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$ is linearly dependent if there is a non-trivial linear combination

$$\lambda_1 \boldsymbol{x}_1 + \ldots + \lambda_k \boldsymbol{x}_k = \boldsymbol{0} \quad \text{with at least one } \lambda_i \neq 0.$$

If the only way to form $\boldsymbol{0}$ is with $\lambda_1, \ldots, \lambda_k = 0$, the set is linearly independent.

Linear independence implies no vector in the set can be written as a linear combination of the others (no redundancy).

In row-echelon form, non-pivot columns can be expressed as linear combinations of pivot columns on their left. So columns of $\boldsymbol{A}$ are linearly independent if and only if all columns in the REF of $\boldsymbol{A}$ are pivot columns.

2.6 Basis and rank

The span of $\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ in vector space V is all possible linear combinations of the vectors in $\mathcal{A}$. If $\text{span}[\mathcal{A}] = V$, then $\mathcal{A}$ is a generating set of V.

If the vectors in generating set $\mathcal{A}$ are linearly independent, then $\mathcal{A}$ is a basis of V.

The canonical / standard basis of $\mathbb{R}^n$ consists of the columns of $\boldsymbol{I}_n$.

Every basis of vector space V has the same number of vectors; the dimension of V.

If $U \subseteq V$, $\dim(U) \leq \dim(V)$, and $\dim(U) = \dim(V)$ if and only if $U = V$.

Finding a basis of $U = \text{span}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$:

1. write the spanning vectors as columns of matrix $\boldsymbol{A}$
2. determine the row-echelon form of $\boldsymbol{A}$
3. the spanning vectors associated with pivot columns are a basis of U

The rank of matrix $\boldsymbol{A}$, written as $\text{rk}(\boldsymbol{A})$, is the number of linearly independent columns (or rows) of $\boldsymbol{A}$. Note: $\text{rk}(\boldsymbol{A}) = \text{rk}(\boldsymbol{A}^\top)$.

If U is the subspace spanned by the columns of $\boldsymbol{A}$, then $\dim(U) = \text{rk}(\boldsymbol{A})$.
Later we'll call this U the *image* or *range* of $\boldsymbol{A}$.

If W is the subspace spanned by the rows of $\boldsymbol{A}$, then $\dim(W) = \text{rk}(\boldsymbol{A})$.

Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is invertible if and only if $\text{rk}(\boldsymbol{A}) = n$.

For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the subspace of solutions for $\boldsymbol{Ax} = \boldsymbol{b}$ has dimension $n - \text{rk}(\boldsymbol{A})$.
Later we'll call this subspace the *kernel* or *null space* of $\boldsymbol{A}$.

Matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has full rank if $\text{rk}(\boldsymbol{A}) = \min\{m, n\}$. Otherwise, $\boldsymbol{A}$ is rank deficient.

2.7 Linear mappings

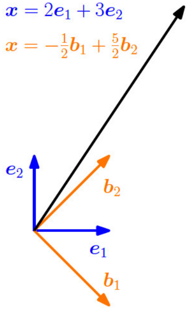
For vector spaces V and W, the mapping $\Phi : V \rightarrow W$ is a linear mapping if

$$\forall \lambda, \psi \in \mathbb{R},\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}\ :\ \Phi(\lambda \boldsymbol{x} + \psi \boldsymbol{y}) = \lambda \Phi(\boldsymbol{x}) + \psi \Phi(\boldsymbol{y})$$

If Φ is bijective*, there exists an inverse mapping $\Psi : W \rightarrow V$ such that $\Psi(\Phi(\boldsymbol{x})) = \boldsymbol{x}$.

Identity mapping in V: $\mathrm{id}_V : V \rightarrow V$, with $\mathrm{id}_V(\boldsymbol{x}) = \boldsymbol{x}$.

Let $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ be an ordered basis of vector space V. Any $\boldsymbol{x} \in V$ can be written as $\boldsymbol{x} = \alpha_1 \boldsymbol{b}_1 + \ldots + \alpha_n \boldsymbol{b}_n$ and we call $\alpha_1, \ldots, \alpha_n$ the coordinates of $\boldsymbol{x}$ w.r.t. B.



* injective: if $\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{y})$ then $\boldsymbol{x} = \boldsymbol{y}$
surjective: $\Phi(V) = W$
bijective: injective and surjective

Let $B = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ and $C = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_m)$ be bases of vectors spaces V and W.
Consider a linear mapping $\Phi : V \to W$, such that $\Phi(\boldsymbol{b}_j) = \alpha_{1,j}\boldsymbol{c}_1 + \dots + \alpha_{m,j}\boldsymbol{c}_m$.
The matrix $\boldsymbol{A}$ with elements $\alpha_{i,j}$ is the transformation matrix of Φ (w.r.t. B and C).

If $\hat{\boldsymbol{x}}$ is the coordinate vector of $\boldsymbol{x}$, and $\hat{\boldsymbol{y}}$ that of $\boldsymbol{y} = \Phi(\boldsymbol{x})$, then $\hat{\boldsymbol{y}} = \boldsymbol{A}\,\hat{\boldsymbol{x}}$.

Basis change

Consider two ordered bases $B = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ and $\tilde{B} = (\tilde{\boldsymbol{b}}_1, \dots, \tilde{\boldsymbol{b}}_n)$ of V,
and two ordered bases $C = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_n)$ and $\tilde{C} = (\tilde{\boldsymbol{c}}_1, \dots, \tilde{\boldsymbol{c}}_n)$ of W,

Let $\boldsymbol{A}_\Phi \in \mathbb{R}^{m \times n}$ be the transformation matrix of $\Phi : V \to W$ w.r.t. bases B and C,
and $\tilde{\boldsymbol{A}}_\Phi \in \mathbb{R}^{m \times n}$ the corresponding transformation matrix w.r.t. bases $\tilde{B}$ and $\tilde{C}$

Then $\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1}\boldsymbol{A}_\Phi\,\boldsymbol{S}$ with $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ the t.m. of id_V that maps coords w.r.t. $\tilde{B}$ to B,
and $\boldsymbol{T} \in \mathbb{R}^{m \times m}$ the t.m. of id_W that maps coords from $\tilde{C}$ to C.

Image and kernel

The image/range of $\Phi : V \to W$ is $\mathrm{Im}(\Phi) = \Phi(V) = \{\boldsymbol{w} \in W \mid \exists \boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \boldsymbol{w}\}$.

The kernel/null space of $\Phi : V \to W$ is $\ker(\Phi) = \Phi^{-1}(\mathbf{0}) = \{\boldsymbol{v} \in V : \Phi(v) = \mathbf{0}\}$.

$\mathrm{Im}(\Phi)$ is a subspace of W, and $\ker(\Phi)$ is a subspace of V.

For the mapping $\Phi(\boldsymbol{x}) = \boldsymbol{Ax}$, $\mathrm{Im}(\Phi)$ is the column space of $\boldsymbol{A}$ (span[columns of $\boldsymbol{A}$]), and $\ker(\Phi)$ is all solutions to $\boldsymbol{Ax} = \mathbf{0}$.

Rank-nullity theorem: $\dim(V) = \dim(\mathrm{Im}(\Phi)) + \dim(\ker(\Phi))$

2.8 Affine subspaces

Let V be a vector space, $\boldsymbol{x}_0 \in V$ and $U \subseteq V$ a subspace of V. The subset L, with $L = \{\boldsymbol{x}_0 + \boldsymbol{u} : \boldsymbol{u} \in U\}$, is called an affine subspace (or linear manifold, or hyperplane).

U is the direction space, and $\boldsymbol{x}_0$ is the support point.

If $\boldsymbol{x}_0 \notin U$, the affine subspace is not a vector subspace because it won't contain $\mathbf{0}$.

If $(\boldsymbol{b}_1, \dots, \boldsymbol{b}_k)$ is a basis of U, then any element $\boldsymbol{x} \in L$ can be written as

$$\boldsymbol{x} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{b}_1 + \ldots + \lambda_k \boldsymbol{b}_k.$$

An affine mapping from V to W has the form $\phi(\boldsymbol{x}) = \boldsymbol{a} + \Phi(\boldsymbol{x})$, where $\Phi : V \to W$ is a linear mapping, and $\boldsymbol{a}$ is a translation vector.