Background Linear Algebra[1]

David Barber

University College London

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Matrices

An $m \times n$ matrix $\mathbf{A}$ is a collection of scalar values arranged in a rectangle of m rows and n columns.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

The i, j element of matrix $\mathbf{A}$ can be written A_{ij} or more conventionally a_{ij}. Where more clarity is required, one may write $[\mathbf{A}]_{ij}$ (for example $\left[\mathbf{A}^{-1}\right]_{ij}$).

Matrix addition

For two matrix $\mathbf{A}$ and $\mathbf{B}$ of the same size,

$$[\mathbf{A} + \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} + [\mathbf{B}]_{ij}$$

Matrix multiplication

For an l by n matrix $\mathbf{A}$ and an n by m matrix B, the product $\mathbf{AB}$ is the l by m matrix with elements

$$[\mathbf{AB}]_{ik} = \sum_{j=1}^{n} [\mathbf{A}]_{ij} [\mathbf{B}]_{jk} \; ; \qquad i = 1, \ldots, l \quad k = 1, \ldots, m \, .$$

In general $\mathbf{BA} \neq \mathbf{AB}$. When $\mathbf{BA} = \mathbf{AB}$ we say they $\mathbf{A}$ and $\mathbf{B}$ commute.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} \end{pmatrix}$$

Identity

The matrix $\mathbf{I}$ is the identity matrix, necessarily square, with 1's on the diagonal and 0's everywhere else. For clarity we may also write $\mathbf{I}_m$ for a square $m \times m$ identity matrix. Then for an $m \times n$ matrix $\mathbf{A}$, $\mathbf{I}_m\mathbf{A} = \mathbf{A}\mathbf{I}_n = \mathbf{A}$. The identity matrix has elements $[\mathbf{I}]_{ij} = \delta_{ij}$ given by the Kronecker delta:

$$\delta_{ij} \equiv \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Transpose

The transpose $\mathbf{B}^{\mathsf{T}}$ of the n by m matrix $\mathbf{B}$ is the m by n matrix D with components

$$\left[\mathbf{B}^{\mathsf{T}}\right]_{kj} = \mathbf{B}_{jk}\,; \qquad k = 1, \ldots, m \quad j = 1, \ldots, n\,.$$

$\left(\mathbf{B}^{\mathsf{T}}\right)^{\mathsf{T}} = \mathbf{B}$ and $(\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$. If the shapes of the matrices $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$ are such that it makes sense to calculate the product $\mathbf{ABC}$, then

$$(\mathbf{ABC})^{\mathsf{T}} = \mathbf{C}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$$

Vector algebra

Vectors

Let $\mathbf{x}$ denote the n-dimensional column vector with components

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

A vector can be considered an $n \times 1$ matrix.

Addition

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

Scalar product

$$\mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{n} w_i x_i = \mathbf{w}^\mathsf{T} \mathbf{x}$$

The length of a vector is denoted $|\mathbf{x}|$, the squared length is given by

$$|\mathbf{x}|^2 = \mathbf{x}^\mathsf{T} \mathbf{x} = \mathbf{x}^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

A unit vector $\mathbf{x}$ has $\mathbf{x}^\mathsf{T} \mathbf{x} = 1$. The scalar product has a natural geometric interpretation as:

$$\mathbf{w} \cdot \mathbf{x} = |\mathbf{w}| \, |\mathbf{x}| \cos(\theta)$$

where θ is the angle between the two vectors. Thus if the lengths of two vectors are fixed their inner product is largest when $\theta = 0$, whereupon one vector is a constant multiple of the other. If the scalar product $\mathbf{x}^\mathsf{T} \mathbf{y} = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

Linear dependence

A set of vectors $\mathbf{x}^1, \ldots, \mathbf{x}^n$ is linearly dependent if there exists a vector $\mathbf{x}^j$ that can be expressed as a linear combination of the other vectors. If the only solution to

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}^i = \mathbf{0}$$

is for all $\alpha_i = 0, i = 1, \ldots, n$, the vectors $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are linearly independent.

Projections

Suppose we wish to resolve the vector $\mathbf{a}$ into its components along the orthogonal directions specified by the unit vectors $\mathbf{e}$ and $\mathbf{e}^*$. That is $|\boldsymbol{e}| = |\boldsymbol{e}|^* = 1$ and $\mathbf{e} \cdot \mathbf{e^*} = 0$. We are required to find the scalar values α and β such that

$$\mathbf{a} = \alpha\mathbf{e} + \beta\mathbf{e}^*$$

$$\mathbf{a} \cdot \mathbf{e} = \alpha\mathbf{e} \cdot \mathbf{e} + \beta\mathbf{e^*} \cdot \mathbf{e}, \qquad \mathbf{a} \cdot \mathbf{e^*} = \alpha\mathbf{e} \cdot \mathbf{e^*} + \beta\mathbf{e^*} \cdot \mathbf{e^*}$$

From orthogonality and unit lengths of the vectors $\mathbf{e}$ and $\mathbf{e}^*$, this becomes

$$\mathbf{a} \cdot \mathbf{e} = \alpha, \qquad \mathbf{a} \cdot \mathbf{e^*} = \beta$$

Hence

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{e})\,\mathbf{e} + (\mathbf{a} \cdot \mathbf{e}^*)\,\mathbf{e}^*$$

The projection of a vector $\mathbf{a}$ onto a direction specified by general $\mathbf{f}$ is $\frac{\mathbf{a} \cdot \mathbf{f}}{|\boldsymbol{f}|^2}\mathbf{f}$.

Determinant

For a square matrix $\mathbf{A}$, the determinant is the volume of the transformation of the matrix $\mathbf{A}$ (up to a sign change). That is, we take a hypercube of unit volume and map each vertex under the transformation. The volume of the resulting object is defined as the determinant. Writing $[\mathbf{A}]_{ij} = a_{ij}$,

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

The determinant in the (3×3) case has the form

$$a_{11}\det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{12}\det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} + a_{13}\det \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

More generally, the determinant can be computed recursively as an expansion along the top row of determinants of reduced matrices.
The absolute value of the determinant is the volume of the transformation.

$$\det\left(\mathbf{A}^{\mathsf{T}}\right) = \det\left(\mathbf{A}\right)$$

For square matrices $\mathbf{A}$ and $\mathbf{B}$ of equal dimensions,

$$\det\left(\mathbf{AB}\right) = \det\left(\mathbf{A}\right)\det\left(\mathbf{B}\right), \qquad \det\left(\mathbf{I}\right) = 1 \Rightarrow \det\left(\mathbf{A}^{-1}\right) = 1/\det\left(\mathbf{A}\right)$$

Matrix inversion

For a square matrix $\mathbf{A}$, its inverse satisfies

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

It is not always possible to find a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, in which case $\mathbf{A}$ singular. Geometrically, singular matrices correspond to projections: if we transform each of the vertices $\mathbf{v}$ of a binary hypercube using $\mathbf{A}\mathbf{v}$, the volume of the transformed hypercube is zero ($\mathbf{A}$ has determinant zero). Given a vector $\mathbf{y}$ and a singular transformation, $\mathbf{A}$, one cannot uniquely identify a vector $\mathbf{x}$ for which $\mathbf{y} = \mathbf{A}\mathbf{x}$. Provided the inverses exist

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

Pseudo inverse

For a non-square matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ is invertible,

$$\mathbf{A}^{\dagger} = \mathbf{A}^{\mathsf{T}}\left(\mathbf{A}\mathbf{A}^{\mathsf{T}}\right)^{-1}$$

satisfies $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$.

Solving Linear Systems

Problem

Given square $N \times N$ matrix $\mathbf{A}$ and vector $\mathbf{b}$, find the vector $\mathbf{x}$ that satisfies

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Solution

Algebraically, we have the inverse:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

In practice, we solve solve for $\mathbf{x}$ numerically using Gaussian Elimination—more stable numerically.

Complexity

Solving a linear system is $O\left(N^3\right)$—can be very expensive for large N.
Approximate methods include conjugate gradient and related approaches.

Matrix rank

For an $m \times n$ matrix $\mathbf{X}$ with n columns, each written as an m-vector:

$$\mathbf{X} = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$$

the rank of $\mathbf{X}$ is the maximum number of linearly independent columns (or equivalently rows).

Full rank

An $n \times n$ square matrix is full rank if the rank is n, in which case the matrix is must be non-singular. Otherwise the matrix is reduced rank and is singular.

Trace and Det

$$\operatorname{trace}(\mathbf{A}) = \sum_i A_{ii} = \sum_i \lambda_i$$

where λ_i are the eigenvalues of $\mathbf{A}$.

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

A matrix is singular if it has a zero eigenvalue.

Trace-Log formula

For a positive definite matrix $\mathbf{A}$,

$$\operatorname{trace}(\log \mathbf{A}) \equiv \log \det(\mathbf{A})$$

The above logarithm of a matrix is not the element-wise logarithm. In general for an analytic function $f(x)$, $f(\mathbf{M})$ is defined via the power-series expansion of the function. On the right, since $\det(\mathbf{A})$ is a scalar, the logarithm is the standard logarithm of a scalar.

Orthogonal matrix

A square matrix $\mathbf{A}$ is orthogonal if

$$\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$$

From the properties of the determinant, we see therefore that an orthogonal matrix has determinant ± 1 and hence corresponds to a volume preserving transformation.

Linear transformations

Cartesian coordinate system

Define $\mathbf{u}_i$ to be the vector with zeros everywhere expect for the i^{th} entry, then a vector can be expressed as $\mathbf{x} = \sum_i x_i \mathbf{u}_i$.

Linear transformation

A linear transformation of $\mathbf{x}$ is given by matrix multiplication by some matrix $\mathbf{A}$

$$\mathbf{A}\mathbf{x} = \sum_i x_i \mathbf{A}\mathbf{u}_i = \sum_i x_i \mathbf{a}_i$$

where $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$.

Eigenvalues and eigenvectors

For an $n \times n$ square matrix $\mathbf{A}$, $\mathbf{e}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ if

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$$

For an $(n \times n)$ dimensional matrix, there are (including repetitions) n eigenvalues, each with a corresponding eigenvector. We can write

$$\underbrace{(\mathbf{A} - \lambda\mathbf{I})}_{\mathbf{B}}\,\mathbf{e} = \mathbf{0}$$

If $\mathbf{B}$ has an inverse, then the only solution is $\mathbf{e} = \mathbf{B}^{-1}\mathbf{0} = \mathbf{0}$, which trivially satisfies the eigen-equation. For any non-trivial solution we therefore need $\mathbf{B}$ to be non-invertible. Hence λ is an eigenvalue of $\mathbf{A}$ if

$$\det\left(\mathbf{A} - \lambda\mathbf{I}\right) = 0$$

It may be that for an eigenvalue λ the eigenvector is not unique and there is a space of corresponding vectors.

Spectral decomposition

A real symmetric matrix $N \times N$ $\mathbf{A}$ has an eigen-decomposition

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathsf{T}}$$

where λ_i is the eigenvalue of eigenvector $\mathbf{e}_i$ and the eigenvectors form an orthogonal set,

$$\left(\mathbf{e}^i\right)^{\mathsf{T}} \mathbf{e}^j = \delta_{ij} \left(\mathbf{e}^i\right)^{\mathsf{T}} \mathbf{e}^i$$

In matrix notation

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathsf{T}}$$

where $\mathbf{E}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ the corresponding diagonal eigenvalue matrix. More generally, for a square non-symmetric 'diagonalisable' $\mathbf{A}$ we can write

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}$$

Computational Complexity

It generally takes $O\left(N^3\right)$ time to compute the eigen-decomposition.

Singular Value Decomposition

The SVD decomposition of a $n \times p$ matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$$

where $\dim \mathbf{U} = n \times n$ with $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_n$. Also $\dim \mathbf{V} = p \times p$ with $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}_p$. The matrix $\mathbf{S}$ has $\dim \mathbf{S} = n \times p$ with zeros everywhere except on the diagonal entries. The singular values are the diagonal entries $[\mathbf{S}]_{ii}$ and are positive. The singular values are ordered so that the upper left diagonal element of $\mathbf{S}$ contains the largest singular value.

Computational Complexity

It generally takes $O\left(\max(n,p)\left(\min(n,p)\right)^2\right)$ time to compute the SVD-decomposition.

Positive definite matrix

A symmetric matrix $\mathbf{A}$ with the property that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0$ for any vector $\mathbf{x}$ is called positive semidefinite. A symmetric matrix $\mathbf{A}$, with the property that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for any vector $\mathbf{x} \neq 0$ is called positive definite. A positive definite matrix has full rank and is thus invertible.

Eigen-decomposition

Using the eigen-decomposition of $\mathbf{A}$,

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \sum_i \lambda_i \mathbf{x}^\mathsf{T}\mathbf{e}^i(\mathbf{e}^i)^\mathsf{T}\mathbf{x} = \sum_i \lambda_i \left(\mathbf{x}^\mathsf{T}\mathbf{e}^i\right)^2$$

which is greater than zero if and only if all the eigenvalues are positive. Hence $\mathbf{A}$ is positive definite if and only if all its eigenvalues are positive.