Advanced Neural Networks (CY4C9) — Mathematical Foundations

Dr Xia Hong, Dr Slawek Nasuto
School of Systems Engineering, UoR

Spring term, 2010

Topics covered in the course:

- Mathematical Foundations (XH)
  - Matrix Computations
  - Probability and Information Theory
  - Optimization Methods and Estimation Theory
- Selected Advances in Neurocomputation (SJN)
  - Support Vector Machines
  - Ensembles of Classifiers
  - Independent Component Analysis
  - Hidden Markov Models
  - Markov Decision Processes and Reinforcement Learning

Reference:

- S. Haykin, "Neural Networks. A comprehensive foundation," Prentice Hall, 2nd edition.

Matrix Computations

Basics

Matrix $\mathbf{A}$: $m \times n$ is a matrix with m rows and n columns.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}$$

Vector $\mathbf{b}$: $m \times 1$ a column vector with m rows.

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = [b_1, b_2, b_3, b_4]^T$$

where ^{T} denotes transpose.

Matrix operations:

1. Multiplication: $\mathbf{C} = \mathbf{AB}$ such that $c_{ij} = \sum_k a_{ik} b_{kj}$ (The number of columns of $\mathbf{A}$ must be the same as number of rows of $\mathbf{B}$)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \tag{1}$$

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \tag{2}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} \\ a_{41}b_{11} + a_{42}b_{21} + a_{43}b_{31} & a_{41}b_{12} + a_{42}b_{22} + a_{43}b_{32} \end{bmatrix}$$

2. Addition: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ such that $c_{ij} = a_{ij} + b_{ij}$ ($\mathbf{A}$ and $\mathbf{B}$ must be the same size)

3. Transpose: $\mathbf{C} = \mathbf{A}^T$ such that $c_{ij} = a_{ji}$

4. Symmetric matrix: $\mathbf{C} = \mathbf{C}^T$ such that $c_{ij} = c_{ji}$ (If $\mathbf{C}$ is symmetric it must be square)

5. Identity matrix: The identity matrix $\mathbf{I}$ is square and has 1s on the major diagonal, elsewhere 0s.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}$$

6. The determinant: The determinant of a scalar given by itself. The determinant of $\mathbf{A} \in \Re^{n \times n}$ is defined in terms of order $(n-1)$ determinants:

$$det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{j+1} a_{1,j} det(A_{1j}).$$

where A_{1j} is an $(n-1)$ by $(n-1)$ matrix obtained by deleting the first row and jth column of $\mathbf{A}$.

Properties of determinant rules:

$$det(\mathbf{AB}) = det(\mathbf{A})det(\mathbf{B}),$$

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

$$det(c\mathbf{A}) = c^n det(\mathbf{A})$$

$$det(\mathbf{A}) \neq 0 \quad \Longleftrightarrow \quad \mathbf{A} \quad \text{is nonsingular}$$

in which $\mathbf{A}, \mathbf{B} \in \Re^{n \times n}$, c is a real scalar.

7. Inverse matrix: $\mathbf{AA}^{-1} = \mathbf{I}$; $\mathbf{A}^{-1}$ exists only if $\mathbf{A}$ is square and not singular. $\mathbf{A}$ is singular if $det(\mathbf{A}) = 0$.

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{|\mathbf{A}|} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{ad - bc}$$

8. Orthogonal matrix: a matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}^T\mathbf{A}$ is diagonal (orthonormal if $\mathbf{A}^T\mathbf{A} = \mathbf{I}$).

9. Matrix trace: The trace of $\mathbf{A} \in \Re^{n\times n}$ is given by

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{i,i}$$

i.e., the sum of the diagonal elements.

Properties:

$$tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$$

$$tr(\alpha\mathbf{A}) = \alpha tr(\mathbf{A})$$

$$tr(\mathbf{A}) = tr(\mathbf{A}^T)$$

$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$

where $\mathbf{A},\mathbf{B} \in \Re^{n\times n}$ and α is a real scalar.

Algebra rules :

$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$ (Addition is commutative)

$\mathbf{AB}\neq\mathbf{BA}$ (Multiplication not commutative)

$\mathbf{A}(\mathbf{BC})=(\mathbf{AB})\mathbf{C}$ (Associative)

$\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$ (Associative)

$\mathbf{AI}=\mathbf{IA}=\mathbf{A}$

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

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

Vector Norms : Norms on vector spaces, denoted by $\|.\|$, furnish a measure of distance. Let $\mathbf{x} = [x_1, \ldots, x_n]^T$. A useful class of vector norms are the p-norms defined by

$$\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

For example, the 2-norm is most common.

$$\|\mathbf{x}\| = (|x_1|^2 + \cdots + |x_n|^2)^{\frac{1}{2}} = (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}$$

A unit vector is a vector $\mathbf{x}$ that satisfies $\|\mathbf{x}\| = 1$.

We can generate a unit vector $\mathbf{e}$ in the direction of $\mathbf{x}$ by $\mathbf{e} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$

Orthogonality: A set of vectors $\{\mathbf{x}_1, ..., \mathbf{x}_p\}$ is orthogonal if $\mathbf{x}_i^T\mathbf{x}_j = 0$ whenever $i \neq j$. These are orthonormal if the set of vectors $\{\mathbf{x}_1, ..., \mathbf{x}_p\}$ are unit vectors.

Linear Independence: A set of vectors $\{\mathbf{x}_1, ..., \mathbf{x}_p\}$ in linearly independent if $\sum_{j=1}^{n} \alpha_j \mathbf{x}_j = \mathbf{0}$ implies $\alpha_j = 0, \forall j$. Otherwise it is said to be linearly dependent.

Inner Product: Inner product between two vectors $\mathbf{x}, \mathbf{y}$ is defined as the product of their magnitude (2-norm) and a cosine of an angle θ between them.

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos \theta$$

Properties:

- linearity:

$$\langle a\mathbf{x}, \mathbf{y} \rangle = a \langle \mathbf{x}, \mathbf{y} \rangle$$

$$\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$$

- $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

- Let $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$. Then

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$$

(sometimes used as a definition of the inner product)

- if $\mathbf{x}$ and $\mathbf{y}$ are both nonzero and $\langle \mathbf{x}, \mathbf{y} \rangle = 0 \Rightarrow \mathbf{x}$ and $\mathbf{y}$ are orthogonal.

Cauchy Schwarz Inequality: For vectors $\mathbf{x}, \mathbf{y}$

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

with equality if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. This is one of the most important inequalities in mathematics.

Eigenvalue and Eigenvectors: Let a linear transformation be represented by a square matrix $\mathbf{A}$. If there is a nonzero vector $\mathbf{x}$ such that

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

for some scalar λ, then λ is called the eigenvalue of with corresponding eigenvector $\mathbf{x}$.

Characteristic Equation can be used to find the eigenvalues and the associated eigenvectors. From

$$\mathbf{A}\mathbf{x} - \lambda\mathbf{x} = \mathbf{0}$$

we obtain the characteristic equation

$$\text{det}\{\mathbf{A} - \lambda\mathbf{I}\} = 0.$$

Example: Find the eigenvalues and eigenvectors for

$$\mathbf{A} = \begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}.$$

$$\det\{\mathbf{A} - \lambda\mathbf{I}\} = \det\{\begin{bmatrix} 4-\lambda & 3 \\ 1 & 2-\lambda \end{bmatrix}\} = 0.$$

$$(4-\lambda)(2-\lambda) - 3 = 0$$

or

$$\lambda^2 - 6\lambda + 5 = 0, \text{ or } (\lambda - 1)(\lambda - 5) = 0$$

so the eigenvalues are $\lambda = 1$, and $\lambda = 5$ respectively.

Plug back the eigenvalues into

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

to find eigenvectors. e.g. For $\lambda = 5$,

$$\begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}\mathbf{x} = 5\mathbf{x}.$$

or

$$\begin{bmatrix} -1 & 3 \\ 1 & -3 \end{bmatrix}\mathbf{x} = 0.$$

Use any of the equations, e.g. $-x_1 + 3x_2 = 0$, and fix x_1 to obtain x_2. If $x_1 = 3$, then $x_2 = 1$, so the eigenvector is $\mathbf{x} = [3 \quad 1]^T$. The unit eigenvector can be obtained as $\mathbf{x} = [\frac{3}{\sqrt{10}} \quad \frac{1}{\sqrt{10}}]^T$.

For $\lambda = 1$,

$$\begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix} \mathbf{x} = \mathbf{x}.$$

or

$$\begin{bmatrix} 3 & 3 \\ 1 & 1 \end{bmatrix} \mathbf{x} = 0.$$

Use any of the equations, e.g. $x_1 + x_2 = 0$, and also fix x_1 to obtain x_2. If $x_1 = 1$, then $x_2 = -1$, so the eigenvector is $\mathbf{x} = [1 \quad -1]^T$. The unit eigenvector can be obtained as $\mathbf{x} = [\frac{1}{\sqrt{2}} \quad -\frac{1}{\sqrt{2}}]^T$.