Mathematics Essentials

Additional reading can be found from relevant math textbooks, e.g. [4], web resources (e.g., Wikipedia). Appendixes of the recommended textbooks [1] and [2]
Outline

• Vector and Matrix Notation
• Vectors
• Matrices
• Vector Spaces
• Linear Transformations
• Eigenvectors and Eigenvalues
• Random Vectors
• Matrix Calculus
• Optimization Basics
• Machine Learning Principle
Vector and Matrix Notation

- A $d$-dimensional (column) vector $\mathbf{x}$ and its transpose are written as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad \text{and} \quad \mathbf{x}^T = [x_1 x_2 \ldots x_d]$$

- An $n \times d$ (rectangular) matrix and its transpose are written as:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1d} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nd} \end{bmatrix} \quad \text{and} \quad \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ a_{13} & a_{23} & \cdots & a_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1d} & a_{2d} & \cdots & a_{nd} \end{bmatrix}$$
Vectors

- The **inner product** (a.k.a. *dot product* or *scalar product*) of two vectors is defined by

\[
\langle x, y \rangle = x^T y = y^T x = \sum_{k=1}^{d} x_k y_k
\]

- The **magnitude** of a vector is

\[
|x| = \sqrt{x^T x} = \left[ \sum_{k=1}^{d} x_k x_k \right]^{1/2}
\]

- The **orthogonal projection** of vector \( y \) onto vector \( x \) is

\[
(y^T u_x) u_x
\]

- where vector \( u_x \) has unit magnitude and the same direction as \( x \)

- The **angle** between vectors \( x \) and \( y \) is

\[
\cos \theta = \frac{x^T y}{|x| \||y|}
\]
Vectors

- Two vectors $x$ and $y$ are said to be
  - orthogonal if $x^Ty = 0$
  - orthonormal if $x^Ty = 0$ and $|x| = |y| = 1$

- A set of vectors $x_1, x_2, \ldots, x_n$ are said to be linearly dependent if there exists a set of coefficients $a_1, a_2, \ldots, a_n$ (at least one different than zero) such that
  $$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$$

- Alternatively, a set of vectors $x_1, x_2, \ldots, x_n$ are said to be linearly independent if
  $$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0 \Rightarrow a_k = 0 \quad \forall k$$
Matrices

- The product of two matrices is

\[
AB = \begin{bmatrix}
a_{11} & a_{12} & a_{13} & \cdots & a_{1d} \\
a_{21} & a_{22} & a_{23} & \cdots & a_{2d} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & a_{m3} & \cdots & a_{md}
\end{bmatrix}
\begin{bmatrix}
b_{11} & b_{12} & \cdots & b_{1n} \\
b_{21} & b_{22} & \cdots & b_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
b_{d1} & b_{d2} & \cdots & b_{dn}
\end{bmatrix}
= \begin{bmatrix}
c_{11} & c_{12} & \cdots & c_{1n} \\
c_{21} & c_{22} & \cdots & c_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
c_{m1} & c_{m2} & c_{m3} & \cdots & c_{mn}
\end{bmatrix}
\]

where \( c_{ij} = \sum_{k=1}^{d} a_{ik}b_{kj} \)

- Properties for matrix multiplication

\[ AB \neq BA \]

\[ ABC = (AB)C = A(BC) \]
Square Matrices

- The determinant of a square matrix $A_{d \times d}$ is

$$|A| = \sum_{k=1}^{d} a_{ik} |A_{ik}| (-1)^{k+i}$$

- where $A_{ik}$ is the minor matrix formed by removing the $i^{th}$ row and the $k^{th}$ column of $A$
- NOTE: the determinant of a square matrix and its transpose is the same: $|A| = |A^T|$

- The trace of a square matrix $A_{d \times d}$ is the sum of its diagonal elements

$$\text{tr}(A) = \sum_{k=1}^{d} a_{kk}$$

- The rank of a matrix is the number of linearly independent rows (or columns)
- A square matrix is said to be non-singular if and only if its rank equals the number of rows (or columns)
  - A non-singular matrix has a non-zero determinant
- A square matrix is said to be orthonormal if $AA^T = A^T A = I$

- The inverse of a square matrix $A$ is denoted by $A^{-1}$ and is such that $AA^{-1} = A^{-1} A = I$
  - The inverse $A^{-1}$ of a matrix $A$ exists if and only if $A$ is non-singular
Vector Space

- The n-dimensional space in which all the n-dimensional vectors reside is called a vector space.

- A set of vectors \( \{u_1, u_2, \ldots, u_n\} \) is said to form a basis for a vector space if any arbitrary vector \( x \) can be represented by a linear combination of the \( \{u_i\} \)

\[
x = a_1 u_1 + a_2 u_2 + \cdots + a_n u_n
\]

- The coefficients \( \{a_1, a_2, \ldots, a_n\} \) are called the components of vector \( x \) with respect to the basis \( \{u_i\} \).

- In order to form a basis, it is necessary and sufficient that the \( \{u_i\} \) vectors be linearly independent.

- A basis \( \{u_i\} \) is said to be orthogonal if

\[
 u_i^T u_j = \begin{cases} 
 0 & i = j \\
 1 & i \neq j 
\end{cases}
\]

- A basis \( \{u_i\} \) is said to be orthonormal if

\[
 u_i^T u_j = \begin{cases} 
 1 & i = j \\
 0 & i \neq j 
\end{cases}
\]

- As an example, the Cartesian coordinate base is an orthonormal base.
A linear transformation is a mapping from a vector space $X^N$ onto a vector space $Y^M$, and is represented by a matrix

- Given vector $x \in X^N$, the corresponding vector $y$ on $Y^M$ is computed as

$$
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_M
\end{bmatrix} =
\begin{bmatrix}
  a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\
  a_{21} & a_{22} & a_{23} & \cdots & a_{2N} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  a_{M1} & a_{M2} & a_{M3} & \cdots & a_{MN}
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_N
\end{bmatrix}
$$

- Notice that the dimensionality of the two spaces does not need to be the same.
- For pattern recognition we typically have $M < N$ (project onto a lower-dimensionality space)
Eigenvectors and Eigenvalues

Given a matrix \( A_{N \times N} \), we say that \( \mathbf{v} \) is an eigenvector\(^*\) if there exists a scalar \( \lambda \) (the eigenvalue) such that

\[
A\mathbf{v} = \lambda \mathbf{v} \iff \begin{cases} 
\mathbf{v} \text{ is an eigenvector} \\
\lambda \text{ is the corresponding eigenvalue}
\end{cases}
\]

Properties

- If \( A \) is non-singular
  - All eigenvalues are non-zero
- If \( A \) is real and symmetric
  - All eigenvalues are real
  - The eigenvectors associated with distinct eigenvalues are orthogonal

\[
|A| = \prod_{i=1}^{N} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_N
\]
Random Vectors

- A random vector refers to a multidimensional generalization of the concept of random variable.
- A random vector can be described with measures similar to those defined for scalar random variables.

Mean vector

\[ \mathbf{\mu} = [\mu_1 \mu_2 \cdots \mu_d]^T, \quad \mu_i = \frac{1}{N} \sum_{n=1}^{N} x_{in}. \]

Covariance matrix

\[ \text{Cov}[X] = \begin{bmatrix} C_{11} & \cdots & C_{1d} \\ \vdots & \ddots & \vdots \\ C_{d1} & \cdots & C_{dd} \end{bmatrix} \]

\[ c_{ij} = \frac{1}{N} \sum_{n=1}^{N} (x_{in} - \mu_i)(x_{jn} - \mu_j). \]
Random Vectors

- The covariance terms can be expressed as:
  \[ c_{ii} = \sigma_i^2 \quad \text{and} \quad c_{ik} = \rho_{ik} \sigma_i \sigma_k \]
  
  - where \( \rho_{ik} \) is called the **correlation coefficient**

- The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to **co-vary**

- The covariance has several important properties:
  
  - If \( x_i \) and \( x_k \) tend to increase together, then \( c_{ik} > 0 \)
  
  - If \( x_i \) tends to decrease when \( x_k \) increases, then \( c_{ik} < 0 \)
  
  - If \( x_i \) and \( x_k \) are **uncorrelated**, then \( c_{ik} = 0 \)
  
  - \( |c_{ik}| \leq \sigma_i \sigma_k \), where \( \sigma_i \) is the standard deviation of \( x_i \)
  
  - \( c_{ii} = \sigma_i^2 = \text{VAR}(x_i) \)
Random Vectors

- Given the following samples from a 3-dimensional distribution
  - Compute the covariance matrix
  - Generate scatter plots for every pair of variables
    - Can you observe any relationships between the covariance and the scatter plots?

- You may work your solution in the templates below

<table>
<thead>
<tr>
<th>Examples</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>
Matrix Calculus

- Matrix derivatives: three-step procedure
  - Step 1: know the dimensions
  - Step 2: element-wise calculations
  - Step 3: put into the vector/matrix form

<table>
<thead>
<tr>
<th>$\mathbf{X}$</th>
<th>$\mathbf{Y}$</th>
<th>Scalar</th>
<th>Vector</th>
<th>Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalar</td>
<td>$\frac{dy}{dx}$</td>
<td>$\frac{dy}{dx} = \left[ \frac{\partial y_i}{\partial x} \right]$</td>
<td>$\frac{d\mathbf{Y}}{dx} = \left[ \frac{\partial y_{ij}}{\partial x} \right]$</td>
<td></td>
</tr>
<tr>
<td>Vector</td>
<td>$\frac{dy}{dx} = \left[ \frac{\partial y_j}{\partial x_j} \right]$</td>
<td>$\frac{dy}{dx} = \left[ \frac{\partial y_i}{\partial x_j} \right]$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matrix</td>
<td>$\frac{dy}{d\mathbf{X}} = \left[ \frac{\partial y}{\partial x_{ji}} \right]$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Matrix Calculus

• Example 1:

\[
\frac{\partial a^T x}{\partial x}
\]

• Dimensions? Element-wise? Vector/matrix form?

\[
\frac{\partial a^T x}{\partial x} = a
\]
Matrix Calculus

- Commonly Used Derivatives

\[ f = x^T A x + b^T x \]

\[ \nabla_x f = \frac{\partial f}{\partial x} = (A + A^T)x + b \]

\[ \frac{\partial^2 f}{\partial x \partial x^T} = A + A^T \]

\[ \frac{\partial a^T X b}{\partial X} = ab^T \]

\[ \frac{\partial a^T X^T b}{\partial X} = ba^T \]
Optimization Basics

- **Optimality condition**
  
  For a differentiable function $f(x), \ x = (x_1, \ldots, x_n)$, optimum point $x'$ must satisfy the following condition:
  
  $$\nabla f(x') = 0 \quad \text{or} \quad \frac{\partial f(x')}{\partial x'_1} = \cdots = \frac{\partial f(x')}{\partial x'_n} = 0.$$  

- **Constrained optimization**
  
  A constrained optimization often has the following form:
  
  minimize $f(x)$  
  subject to $c_i(x) = 0$  

  objective function  
  constraint functions
Optimization Basics

• Lagrangian multipliers
  - Convert a constrained optimization problem into an unconstrained counterpart.
  - Introduce parameters to form a unconstrained function:

    \[
    \text{minimize} \quad L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) = f(x) + \lambda^T c(x).
    \]

  - The constant of proportionality is called the Lagrangian multipliers.
  - In general, the Lagrangian function has the property that when its gradient is zero at the optimum; i.e.,

    \[
    \frac{\partial L(x, \lambda)}{\partial x} = 0 \quad \text{or} \quad \frac{\partial L(x, \lambda)}{\partial x_1} = \cdots = \frac{\partial L(x, \lambda)}{\partial x_n} = 0.
    \]

    and \[
    \frac{\partial L(x, \lambda)}{\partial \lambda} = 0 \quad \text{or} \quad \frac{\partial L(x, \lambda)}{\partial \lambda_1} = \cdots = \frac{\partial L(x, \lambda)}{\partial \lambda_m} = 0.
    \]
Optimization Basics

- Lagrange Multiplier Example: SVM Optimization
  - Minimize the margin cost:
    \[ \Phi(w) = \frac{1}{2} w^T w \]
    subject to the following constraints
    \[ y_i (w^T x_i + b) \geq 1 \quad \text{for} \quad i = 1, 2, \ldots, N \]
  - Construct the unconstrained cost with Lagrangian multipliers
    \[ J(w, b, \lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i [y_i (w^T x + b) - 1] \]
Optimization Basics

- Lagrange Multiplier Example: SVM Optimization (cont.)
  - Apply the optimality condition
    \[
    \frac{\partial J(w, b, \lambda)}{\partial w} = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{N} \lambda_i y_i x_i
    \]
    \[
    \frac{\partial J(w, b, \lambda)}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \lambda_i y_i = 0
    \]
  - Insert to the original cost function to get a dual cost function
    \[
    Q(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j
    \]
    - Apply the quadratic programming technique to find solutions
      \[
      \lambda_i^* : \text{indicator on whether } x_i \text{ is a SV}
      \]
      Kernel SVM
      \[
      \Phi(x_i)^T \Phi(x_j) = K(x_i, x_j)
      \]
Where does a machine learning algorithm come from?

“Model”
A mechanism that can encode knowledge for problem solving, which is often a parametric model. Learning is going to find “appropriate” parameters.

“Cost/Utility function”
The performance criterion: the function defined based on a learning model for a specific learning goal to judge how well the parameters of this model are set based on a training data set.

“Learning algorithm”
The algorithm comes from an optimisation process that minimises/maximises the error/cost/loss/likelihood function with respect to parameters, which derives a learning rule for finding “appropriate” parameters with a given training data set.