part I: Linear Algebra

Some notations : $A_{m\times n}$ is matrix with m rows and n columns , $A_{m\times n} \in \mathbb{R}^{m\times n}$

A_{ij} is the (i,j)th element of A , A^T is the transpose of A

- $$A = \begin{bmatrix} a_{11}, & a_{12}, & \cdots, & a_{1N} \\ \vdots & & \ddots & \vdots \\ a_{m1}, & a_{m2}, & \cdots, & a_{mn} \end{bmatrix} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n],$$

  with $\mathbf{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{mi} \end{bmatrix}$ $\qquad a_i^t = [a_{i1}, a_{i2}, \cdots a_{in}]$, the ith row of A

  the ith column of A.

- $C = AB$, where $A_{m\times k}$, $B_{k\times n}$, we have $C_{ij} = a_i^t \cdot b_j$, $C \in \mathbb{R}^{m\times n}$

- $C = \sum_{i=1}^{k} a_i \cdot b_i^t$, outer-product formulation

- $(A\cdot B)^T = (B^T \cdot A^T)$, where $A_{m\times k}$, $B_{k\times n}$ are not necessarily square

- $A\cdot A^{-1} = A^{-1}\cdot A = I$, where A must be square matrix and non-singular

- $(A^T)^{-1} = (A^{-1})^T$

short proof : $\because (A\cdot A^{-1})^T = I^T = I$

$\therefore (A^{-1})^T \cdot A^T = I \Rightarrow (A^{-1})^T = (A^T)^{-1}$ //

Some Special Matrices :

① Identity matrix, Let $A = I_N = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots & \\ & & & 1 \end{bmatrix}$, whose diagonal elements are all "1"s and off diagonal elements are all zeros.

② Symmetric matrix : Let A be a square matrix, we call it a symmetric matrix if $A = A^T$.

③ Idempotent matrix : Let A be a square matrix, that satisfies $A = A \cdot A$, then A is called a idempotent matrix.

④ orthonormal matrix : Let A be a square matrix, we call it orthonormal matrix if $A^T \cdot A = I$, which also implies $A^{-1} = A^T$

⑤ positive Semi-definite matrix : A is said to be a positive semi definite matrix if it is satisfied that :

(a) $A = A^T$ (b) $y^T \cdot A \cdot y \geq 0$, for any $y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$

⑥ positive definite matrix : A is said to be a positive definite matrix if (a) $A = A^T$ (b) $y^T A y > 0$, for any $y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$
$y \neq 0$

Trace & Determinant (apply to square matrices)

- Definition of trace: Let A be an $n \times n$ matrix. The trace of A, denoted by $tr(A)$, is defined to be the sum of the diagonal elements of A, i.e., $tr(A) = \sum_{i=1}^{n} a_{ii}$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} \cdots & a_{3n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \longrightarrow \text{Summation} \longrightarrow tr(A)$$

- properties:

① For $B_{m \times n}$, $C_{n \times m}$, we have $tr(B \cdot C) = tr(C \cdot B)$

② For $B_{m \times n}$, $C_{n \times q}$, $D_{q \times m}$, we have

$$\underbrace{tr(BCD) = tr(DBC) = tr(CDB)}_{\text{cyclic property}}$$

③ For $A_{n \times n}$, $B_{n \times n}$, we have $tr(\alpha \cdot A + \beta \cdot B) = \alpha\, tr(A) + \beta\, tr(B)$

where α, β are constant scalars.

trace is a linear operator.

properties of determinant and rank

- $|AB| = |A|\,|B|$, where A and B are $n \times n$ matrices (NOT very easy to prove!)
- $|A^{-1}| = \frac{1}{|A|}$, A is non-singular $\Rightarrow$ A^{-1} exists

  Proof: $A^{-1}A = I \Rightarrow |A^{-1}A| = |A^{-1}|\,|A| = |I| = 1 \Rightarrow |A^{-1}| = \frac{1}{|A|}$
- $|A| = |A^T|$
- $|\alpha A| = \alpha^n |A|$

Rank of a matrix A of size $m \times n$.

Def: The dimension of the row space and the column space of the matrix A is called the rank of A., denoted by $\mathrm{rank}(A)$

properties:
- $\mathrm{rank}(A) \leq \min(m, n)$
- If $\mathrm{rank}(A) = \min(m, n) \Rightarrow$ A is of full rank!
- If $\mathrm{rank}(A) < \min(m, n) \Rightarrow$ A is rank deficient!

Vector Norm.

① L2 norm of $\mathbf{x} = [x_1, x_2, x_3, \dots x_n]^T$ is defined as

$$\|\mathbf{x}\|_2 \triangleq \sqrt{x_1^2 + x_2^2 + \dots x_n^2}$$

and L2 norm is also often referred as Euclidean norm.

② L1 norm of $\mathbf{x} = [x_1, x_2, \dots x_n]^T$ is defined as

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \dots + |x_n|$$
$$= \sum_{i=1}^{n} |x_i|$$

Vector norms will be used when we talk about regularized least-squares.

Some other interesting norms that you may see in the literature:

③ L_0 Norm of $\mathbf{x} = [x_1, x_2, \dots x_n]^T$ is defined as

$$\|\mathbf{x}\|_0 = \#(i \mid x_i \neq 0),$$

i.e., equal to the number of non-zero entries of $\mathbf{x}$.

④ L_∞ norm of $\mathbf{x} = [x_1, x_2, \dots x_n]^T$ is defined as

$$\|\mathbf{x}\|_\infty = \max_i \{|x_i|\},$$

i.e., equal to the maximum entry's magnitude of the vector $\mathbf{x}$.

Subspace : def : a set $S \subseteq \mathbb{R}^m$ is called a subspace if for any $\alpha, \beta \in \mathbb{R}$,

$$x, y \in S \Rightarrow \alpha x + \beta y \in S$$

well-known subspaces :

① span : given a collection of vectors $\{a_1, a_2, \cdots a_n\} \subseteq \mathbb{R}^m$

$$\text{span}\{a_1, a_2, \cdots, a_n\} \triangleq \{x \in \mathbb{R}^m \mid x = \sum_{i=1}^{n} \alpha_i a_i, \alpha_i \in \mathbb{R}\}$$

② orthogonal Complement subspace :

given a subset $S \subseteq \mathbb{R}^m$

$$S_\perp = \{x \in \mathbb{R}^m \mid x^T y = 0, \text{ for all } y \in S\}$$

$S_\perp$ is an orthogonal Complement subspace of S.

③ range space :

give $A \in \mathbb{R}^{m \times n}$

$$R(A) \triangleq \{x \in \mathbb{R}^m \mid x = Ay, y \in \mathbb{R}^n\}$$

$$\equiv \text{span}\{a_1, a_2, \cdots, a_n\}$$

proof $\because x = Ay = \sum_{i=1}^{n} a_i \cdot y_i$, where y_i is the ith element of y and $y_i \in \mathbb{R}^n$

This corresponds to the definition of span.

④ Null space : given $A \in \mathbb{R}^{m \times n}$, $N(A) \triangleq \{x \in \mathbb{R}^n \mid Ax = 0\}$

$R(A)_\perp = N(A^T)$!

proof : $\mathcal{N}(A^T) \triangleq \{x \in \mathbb{R}^m \mid A^T x = 0\}$

$\mathcal{R}(A) \triangleq \{\tilde{x} \in \mathbb{R}^m \mid \tilde{x} = Ay, y \in \mathbb{R}^n\}$

$\because \tilde{x}^T x = y^T A^T x = 0$ for all $x \in \mathcal{N}(A^T)$

$\therefore \mathcal{N}_{\perp}(A^T) = \mathcal{R}(A)$

Derivatives:

Suppose ① $f(x) : \mathbb{R}^1 \to \mathbb{R}^1$ is a scalar-valued function of a scalar argument. x

② $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^1$ is a scalar-valued function of an n-vector argument $\mathbf{x} = [x_1, x_2, \dots x_n]^T$. Sometimes, we write out the n scalar arguments, $x_1, x_2, \dots, x_n$:

$$f(\mathbf{x}) = f(x_1, x_2, \dots x_n)$$

③ $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ is a vector-valued function of an n-vector argument $\mathbf{x}$. We can write $f(\mathbf{x})$ as

$$f(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}$$

where $f_i(\mathbf{x})$ is a scalar-valued function of $\mathbf{x} = [x_1, x_2, \dots x_n]^T$.

The above three cases are more often seen. However, a complete list of all cases are given in the next page. As a short summary, we may have

① $f(x) : \mathbb{R}^1 \to \mathbb{R}^1$ ④ $f(x) : \mathbb{R}^1 \to \mathbb{R}^m$

② $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^1$ ⑤ $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$

③ $f(X) : \mathbb{R}^{m \times n} \to \mathbb{R}^1$ ⑥ $f(x) : \mathbb{R}^{m \times n} \to \mathbb{R}^m$

$f(X)$ 不同类别，X 不同类别

	$f(X): R^1 \to R^1$ $R^n \nearrow$ $R^{m\times n} \nearrow$	$f(X): R^1 \to R^n$ $R^n \to R^n$ $R^{m\times n} \to R^n$	$f(X): R^1 \to R^{m\times n}$ $: R^n \to R^{m\times n}$ $R^{m\times n} \to R^{m\times n}$
$X: R^1$	$\frac{\partial f(X)}{\partial X}$, scalar	$\frac{\partial f(X)}{\partial X}$, column vector $n\times 1$	$\frac{\partial f(X)}{\partial X}$, $m\times n$ matrix
$X: R^n$	$\frac{\partial f(X)}{\partial X}$, column vector $n\times 1$	$\frac{\partial f(X)}{\partial X}$, $n\times n$ matrix	$\frac{\partial f(X)}{\partial X}$, hard
$X: R^{m\times n}$	$\frac{\partial f(X)}{\partial X}$, $m\times n$ matrix	$\frac{\partial f(X)}{\partial X}$, difficult	$\frac{\partial f(X)}{\partial X}$, hard

Part II : Statistics

1. Expectation and Covariance matrix

Let us first assume an n-vector $\mathbf{x} = [x_1, x_2, \cdots x_n]^T$ is a random vector, that follows the probability density function (pdf) $p(\mathbf{x})$.

1.1 The expected value of $\mathbf{x}$ is given by

$$\boldsymbol{\mu} \triangleq E(\mathbf{x}) = \int_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x} \cdot p(\mathbf{x})\, d\mathbf{x}$$

The n-vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots \mu_n]^T$, whose elements are given by

$$\mu_i \triangleq \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n \text{ times}} x_i \, p(\mathbf{x})\, d\mathbf{x} \qquad , i = 1, 2, \cdots, n$$

$$= \int_{-\infty}^{\infty} x_i \, p(x_i)\, dx_i$$

where $p(x_i)$ is the marginal distribution of $\overbrace{p(\mathbf{x})}^{\text{the joint distribution}}$.

1.2 The covariance matrix K associated with a real random vector $\mathbf{x}$ is given by

$$K \triangleq \mathrm{cov}(\mathbf{x}) = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right],$$

Where the (i,j)th component of K is

$$K_{ij} \triangleq E[(x_i - \mu_i)(x_j - \mu_j)].$$

It is easy to verify that K is a symmetric matrix with

$$K_{ij} = K_{ji} \qquad \text{(pls verify this point by yourself!)}$$

Next, Let us introduce some short-hand notations, as follows:

$$K_{ij} \triangleq \begin{cases} \sigma_{ij}, & \text{if } i \neq j \\ \sigma_i^2, & \text{if } i = j \end{cases}$$

then the matrix K can be written as

$$K = \begin{bmatrix} \sigma_1^2, \sigma_{12}, \sigma_{13}, \cdots & \sigma_{1n} \\ \sigma_{21}, \sigma_2^2, \sigma_{23}, \cdots & \sigma_{2n} \\ \vdots \qquad \ddots & \vdots \\ \sigma_{n1}, \sigma_{n2}, \cdots\cdots & \sigma_n^2 \end{bmatrix}$$

where the diagonal terms all are variances. $E[(x_i-\mu_i)(x_i-\mu_i)]$, which you have learned in the probability theory where uni-variate random variable is introduced.

Note that: don't confuse the covariance matrix K with the correlation matrix R, which is given by $R \triangleq E[x x^T]$.

It is easy to verify that $K = R - \mu\mu^T$.

2. uncorrelated random vectors

orthogonal random vectors

independent random vectors

Definition: Consider two real n-random vectors $\mathbf{x}$ and $\mathbf{y}$ with respective mean vectors μ_x, μ_y, and pdfs $p(\mathbf{x}), p(\mathbf{y})$.

① If the expected value of their outer product satisfies

$$E[xy^T] = \mu_x \mu_y^T,$$

then $\mathbf{x}$ and $\mathbf{y}$ are said to be uncorrelated.

② If

$$E[xy^T] = O_{n\times n} \quad (\text{a zero-matrix}),$$

then $\mathbf{x}$ and $\mathbf{y}$ are said to be orthogonal.

③ If the joint pdf of $\mathbf{x}$, and $\mathbf{y}$, defined as $p(x, y)$ satisfies

$$p(x, y) = p(x) \cdot p(y),$$

then $\mathbf{x}$ and $\mathbf{y}$ are said to be independent.

Remarks:

① Independence $\Rightarrow$ uncorrelatedness

② uncorrelatedness $\not\Rightarrow$ independence

③ For multi-variate Gaussian RVs, independence $\Longleftrightarrow$ uncorrelatedness.

<u>Exercise</u>: Show that the covariance matrix K is positive semi-definite (PSD).

3 Gaussian distributed RV

3.1 univariate case

Let $x \in \mathbb{R}^1$ be a Gaussian distributed random variable, with the pdf $p(x; \mu, \sigma^2)$ given by

$$p(x) = \underline{N(x; \mu, \sigma^2)} \triangleq \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

The expected mean is given by

$$\mu = E(x) = \int_{R^1} x\, p(x)\, dx = \int_{R^1} x \cdot N(x; \mu, \sigma^2)\, dx$$

$$\sigma^2 = E[(x-\mu)^2] = \int_{R^1} (x-\mu)^2 p(x)\, dx = \int_{R^1} (x-\mu)^2 \cdot N(x; \mu, \sigma^2)\, dx$$

3.2 multi-variate case

Let $\mathbf{x} \in \mathbb{R}^n$ be a Gaussian distributed n-vector random variable, with the pdf $p(\mathbf{x})$ given by

$$p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \triangleq \frac{1}{(\sqrt{2\pi})^{n/2} \cdot |\Sigma|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right]$$

The expected mean and the covariance matrix are given by

$$\boldsymbol{\mu} = E(\mathbf{x}) = \int_{R^n} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} = \int_{R^n} \mathbf{x} \cdot N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)\, d\mathbf{x}$$

$$\Sigma = E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\} = \int_{R^n} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \cdot p(\mathbf{x})\, d\mathbf{x}$$

$$= \int_{R^n} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)\, d\mathbf{x}$$

Multi-variate Gaussian Random Variable :

Theorem 1: If $Y_1, Y_2, \cdots Y_N$ are jointly Gaussian and mutually independent

i.e. $p(Y_1, Y_2) = p(Y_1) \cdot p(Y_2)$, then they are mutually uncorrelated.

Theorem 2: If $Y_1, Y_2, \cdots Y_N$ are jointly Gaussian and mutually uncorrelated,

i.e. $E(Y_1 Y_2) = E(Y_1) \cdot E(Y_2)$, then they are mutually independent.

or $Cov(Y_1, Y_2) = 0$

Theorem 3: If $Y_1, Y_2, \cdots Y_N$ are jointly Gaussian and mutually uncorrelated,

then
$$Cov(\underline{Y}) = \Sigma = \begin{bmatrix} Var(Y_1) & & & \\ & Var(Y_2) & & \\ & & \ddots & \\ & & & Var(Y_N) \end{bmatrix}$$

Remark: when the elements of a multivariate Gaussian random vector $X = [X_1, X_2, \cdots X_n$ are mutually uncorrelated, then the covariance matrix Σ is a diagonal matrix, i.e.,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

where

$$\begin{cases} E\{(X_i - \mu_i)(X_j - \mu_j)\} = 0, & i \neq j, \\ E\{(X_i - \mu_i)^2\} = \sigma_i^2, & i = j. \end{cases}$$