Linear algebra

Some linear algebra is important for understanding many machine learning methods, such as linear or logistic regression.

Matrices and transposes

A is a $m \times n$ real matrix, written $A \in \mathbb{R}^{m \times n}$ if

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

where $a_{ij} \in \mathbb{R}$. The (i, j)th entry of A is $A_{ij} = a_{ij}$.

The transpose of $A \in \mathbb{R}^{m \times n}$ is defined as

$$A^T = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & a_{2n} & \cdots & A_{mn} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

That is, $(A^T)_{ij} = A_{ji}$.
Note that $x \in \mathbb{R}^n$ is considered to be a column vector in $\mathbb{R}^{n \times 1}$.

Sums and products of matrices

The sum of matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$ is the matrix $A + B \in \mathbb{R}^{m \times n}$ such that

$$(A + B)_{ij} = A_{ij} + B_{ij}.$$

The product of matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times \ell}$ is the matrix $AB \in \mathbb{R}^{m \times \ell}$ such that

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Basic matrix properties

In the following properties, it is assumed that the matrix dimensions are compatible. (For example, if we write $A + B$ then it is assumed that A and B are the same size.)

- $(AB)C = A(BC)$
- $A(B + C) = AB + AC$
- $(B + C)A = BA + CA$
- Except in certain situations, AB is not equal to BA.
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$.

Identity

The $n \times n$ identity matrix denoted $I_{n \times n}$ or I is

$$I = I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

$$IA = A = AI$$

Inverse

If it exists, the inverse of A denoted A^{-1} is a matrix such that $A^{-1}A = I$ and $AA^{-1} = I$.

If A^{-1} exists, we say that A is invertible.

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

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

Trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted trA is defined as

$$tr(A) = \sum_{i=1}^{n} A_{ii}.$$

$tr(AB) = tr(BA)$ if AB is a square matrix.

Symmetric and definite matrices

A is symmetric if $A = A^T$.

A is symmetric positive semi-definite (SPSD) if and only if $A = B^T B$ for some $B \in \mathbb{R}^{m \times n}$.

A is symmetric positive definite (SPD) if and only if A is SPSD and A^{-1} exists.

There are many equivalent definitions of SPSD and SPD, however, these are the ones that I will provide for this course.

Discrete random variables

Informally, a random variable (r.v.) is a quantity that probabilistically takes any one of a range of values.

Usual: uppercase for a r.v. and lowercase for the observed value.

A r.v. X is discrete if it takes values in a countable set $\mathcal{X} = \{x_1, x_2, \ldots\}$.

Examples: Bernoulli, Binomial, Poisson, Geometric.

The density of a discrete r.v. is the function $p(x) = \mathbb{P}(X = x)$= probability that X equals x.

Sometimes, $p(x)$ is called the probability mass function in the discrete case, but "density" is also correct.

Properties:

$$0 \leq p(x) \leq 1, \quad \sum_{x \in \mathcal{X}} p(x) = 1 \quad \mathbb{P}(X \in A) = \sum_{x \in A} p(x).$$

Continuous random variables

A random variable $X \in \mathbb{R}$ is continuous if there is a function $p(x) \geq 0$ such that $P(X \in A) = \int_A p(x)\, dx$ for all $A \subseteq \mathbb{R}$. Examples: Normal, Uniform, Beta, Gamma, Exponential.

We call $p(x)$ the probability density function of X. But, it's not the probability that X equals x!

While $\int_A p(x)\, dx = 1$, it can occur that $p(x) > 1$.

Note that the same definitions apply to random vectors $X \in \mathbb{R}^n$.

The cumulative distribution function (cdf) of $X \in \mathbb{R}$ is

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{\infty} p(x')\, dx'$$

Joint distributions and random variables

Let $p(x, y)$ denotes the joint density of $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.

- $\mathbb{P}(X = x, Y = y) = p(x, y)$ if X and Y are discrete r.v.
- $\mathbb{P}(X \in A, Y \in B) = \int_{A \times B} p(x, y)\, dx\, dy$ if X and Y are continuous.
- The density of X can be recovered from the joint density by marginality (summing/integrating) over Y:
  - $p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$ if Y discrete.
  - $p(x) = \int_{\mathcal{Y}} p(x, y)\, dy$ if Y continuous.

It is common to use "p" to denote all densities and follow the convention that X is taking the value x, Y is taking the value y, etc.

Conditional densities and independence

If $p(y) > 0$ then the conditional density of X given $Y = y$ is

$$p(x \mid y) = \frac{p(x, y)}{p(y)}.$$

X and Y are independent if $p(x, y) = p(x)p(y)$ for all x, y.

$X_1, \ldots, X_n$ are independent if

$$p(x_1, \ldots, x_n) = p(x_1) \times p(x_n)$$

for all $x_1, \ldots, x_n$.

$X_1, \ldots, X_n$ are conditionally independent given Y if

$$p(x_1, \ldots, x_n \mid y) = p(x_1 \mid y) \times p(x_n \mid y)$$

for all $x_1, \ldots, x_n, y$.

Expectations

Suppose $h(x)$ is a real-valued function of x.
The expectation of $h(X)$, denoted $E(h(X))$ is

- $E(h(X)) = \sum_{x \in \mathcal{X}} h(x)p(x)$ if X is discrete.
- $E(h(X)) = \int_{\mathcal{X}} h(x)p(x)dx$ if X is continuous.

The conditional expectation of $h(X)$ given $Y = y$ is

- $E(h(X) \mid Y = y) = \sum_{x \in \mathcal{X}} h(x)p(x \mid y)$ if X is discrete.
- $E(h(X) \mid Y = y) = \int_{\mathcal{X}} h(x)p(x \mid y)dx$ if X is continuous.

Let $g(Y) = E[h(X) \mid Y]$, where $g(y) = E[h(X) \mid Y = y]$.

The law of iterated expectations is

$$E\left[E(h(X) \mid Y) = E(h(X)\right]$$

Random vectors

Let $Z_1, \ldots, Z_n \in \mathbb{R}$ be r.v.. Then

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} Z_1 & Z_2 & \cdots & Z_n \end{pmatrix}^T$$

is a random vector in $\mathbb{R}^n$.

The expectation of a random vector $Z \in \mathbb{R}^n$ is

$$E(Z) = \begin{pmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{pmatrix}$$

Covariance matrix

The covariance matrix of a random vector $Z \in \mathbb{R}^n$ is the matrix $Cov(Z) \in \mathbb{R}^{n \times n}$ with (i, j)th entry

$$Cov(Z)_{ij} = Cov(Z_i, Z_j).$$

where

$$\begin{aligned} Cov(Z_i, Z_j) &= E\left[(Z_i - E(Z_i))(Z_j - E(Z_j))\right] && (1) \\ &= E(Z_i Z_j) - E(Z_i)E(Z_j) && (2) \end{aligned}$$

It is equivalent that

$$\begin{aligned} Cov(Z) &= E\left[(Z - E(Z))(Z - E(Z))^T\right] && (3) \\ &= E(ZZ^T) - E(Z)E(Z)^T && (4) \end{aligned}$$

Recall that $Z \in \mathbb{R}^n$ is considered to be a column vector in $\mathbb{R}^{n \times 1}$ so ZZ^T is a matrix in $\mathbb{R}^{n \times n}$.

Covariance matrix

$Cov(Z)$ is always SPSD.

If $Z \in \mathbb{R}^n$ is a random vector, then

$$E[AZ + b] = AE[Z] + b$$

and

$$Cov(AZ + b) = ACov(Z)A^T.$$

for any fixed (non-random) $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.
If $Y, Z \in \mathbb{R}^n$ are independent, random vectors, then

$$Cov(Y + Z) = Cov(Y) + Cov(Z).$$

Multivariate normal distribution

If $\mu \in \mathbb{R}^n$ and $C \in \mathbb{R}^{n \times n}$ is SPSD, then $Z \sim N(\mu, C)$ denotes that Z is multivariate normal with $E(Z) = \mu$ and $Cov(Z) = C$.

Standard Multivariate normal: If $Z_1, \ldots, Z_n \sim N(0, 1)$ independently and $Z = (Z_1, \ldots, Z_n)^T$ then $Z \sim N(0, I)$.

Affine transformation property: If $Z \sim N(\mu, C)$ then $AZ + b \sim N(A\mu + b, ACA^T)$ for any fixed matrix $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \mu \in \mathbb{R}^n$ and SPSD $C \in \mathbb{R}^{n \times n}$.

Any multivariate normal distribution can be obtained via an affine transformation $(AZ + b)$ of $Z \sim N(0, I_{n \times n})$ for an appropriate choice of $n, A, andb$.

Multivariate normal distribution

Sum property: If $Y \sim N(\mu_1, C_1)$ and $Z \sim N(\mu_2, C_2)$, independently, then $Y + Z \sim N(\mu_1 + \mu_2, C_1 + C_2)$.

Density: If $Z = (Z_1, \ldots, Z_n)^T \sim N(\mu, C)$ and C^{-1} exists, the Z has density:

$$p(x) = \frac{1}{2(\pi)^{n/2}|det(C)|^{1/2}} \exp\{\frac{-1}{2}(z-\mu)^T C^{-1}(z-\mu)\}$$

for all $z \in \mathbb{R}^n$