Lecture 2: Probability and linear algebra basics

Statistical Learning (BST 263)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Linear algebra basics

Probability basics

Random vectors

Outline

Linear algebra basics

Probability basics

Random vectors

Linear algebra in this course

- A little bit of linear algebra is essential for understanding many machine learning methods.
  - E.g., linear regression, logistic regression, LDA, QDA, PCA, GAMs, kernel ridge, SVMs, K-means.

- Linear algebra is not a prerequisite for this course, so I made the following slides to give you the basic concepts needed.
- You will need to study this material carefully if you are not already familiar with it.

Matrices and transposes

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

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

  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

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

  In other words, $(A^{\mathtt{T}})_{ij} = A_{ji}$.
- Note: $x \in \mathbb{R}^n$ 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.)

- $(AB)C = A(BC)$
  - Consequently, we can write ABC without specifying the order in which the multiplications are performed.
- $A(B + C) = AB + AC$
- $(B + C)A = BA + CA$
- Except in special circumstances, AB is not equal to BA.
- $(AB)^{\mathtt{T}} = B^{\mathtt{T}} A^{\mathtt{T}}$
- $(A + B)^{\mathtt{T}} = A^{\mathtt{T}} + B^{\mathtt{T}}$

Identity, inverse, and trace

- The $n \times n$ *identity matrix*, denoted $I_{n\times n}$ or I for short, is

$$I = I_{n\times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

- $IA = A = AI$
- 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*.
- $(A^{-1})^{\mathtt{T}} = (A^{\mathtt{T}})^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- The *trace* of a square matrix $A \in \mathbb{R}^{n\times n}$, denoted $\mathrm{tr}(A)$, is defined as $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$.
- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ if AB is a square matrix.

Symmetric and definite matrices

- A is *symmetric* if $A = A^{\mathtt{T}}$.
- A is *symmetric positive semi-definite* (SPSD) if and only if $A = B^{\mathtt{T}}B$ for some $B \in \mathbb{R}^{m \times n}$ and some m.
- 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 (which is why I wrote "if and only if"). I believe the definitions above are the easiest to understand and use.

Outline

Linear algebra basics

Probability basics

Random vectors

Discrete random variables

- Informally, a random variable (r.v.) is a quantity that probabilistically takes any one of a range of values.
- Notation: Uppercase for r.v.s, lowercase for values taken.

- A random variable 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 technically correct also.

- Properties (discrete case):

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

Continuous random variables

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

- $p(x)$ is called the *density* of X.
- Careful! $p(x)$ is not the probability that X equals x.
- Note that $\int_{\mathbb{R}} p(x)dx = 1$, but $p(x)$ can be > 1.

- The same definitions apply to random vectors $X \in \mathbb{R}^n$, with $\mathbb{R}^n$ in place of $\mathbb{R}$.

- The *cumulative distribution function* (c.d.f.) of $X \in \mathbb{R}$ is

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

Joint distributions of multiple random variables/vectors

- $p(x, y)$ 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.
  - $\mathbb{P}(X \in A, Y \in B) = \int_{A \times B} p(x, y) dx\, dy$ if X and Y are continuous.
  - $\mathbb{P}(X = x, Y \in B) = \int_B p(x, y) dy$ if X is discrete and Y is continuous.
- The density of X can be recovered from the joint density by *marginalizing* over Y:
  - $p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$ if Y is discrete,
  - $p(x) = \int_{\mathcal{Y}} p(x, y) dy$ if Y is continuous.
- Note: 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

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

- 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

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

  for all $x_1, \ldots, x_n$.

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

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

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

Expectations (a.k.a. expected values)

- Suppose $h(x)$ is a real-valued function of x.

- The *expectation* of $h(X)$, denoted $\mathrm{E}(h(X))$, is
  - $\mathrm{E}(h(X)) = \sum_{x \in \mathcal{X}} h(x)p(x)$ if X is discrete,
  - $\mathrm{E}(h(X)) = \int_{\mathcal{X}} h(x)p(x)dx$ if X is continuous.

- The *conditional expectation* of $h(X)$ given $Y = y$ is
  - $\mathrm{E}(h(X) \mid Y = y) = \sum_{x \in \mathcal{X}} h(x)p(x|y)$ if X is discrete,
  - $\mathrm{E}(h(X) \mid Y = y) = \int_{\mathcal{X}} h(x)p(x|y)dx$ if X is continuous.

- $\mathrm{E}(h(X)|Y)$ is defined as $g(Y)$ where $g(y) = \mathrm{E}(h(X)|Y = y)$.
- *Law of iterated expectations*: $\mathrm{E}(\mathrm{E}(h(X)|Y)) = \mathrm{E}(h(X))$.

Outline

Linear algebra basics

Probability basics

Random vectors

Random vectors

- If $Z_1, \ldots, Z_n \in \mathbb{R}$ are random variables, then

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} = (Z_1, \ldots, Z_n)^{\mathtt{T}}$$

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

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

$$\mathrm{E}(Z) = \begin{bmatrix} \mathrm{E}(Z_1) \\ \vdots \\ \mathrm{E}(Z_n) \end{bmatrix}.$$

Random vectors

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

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

  where

$$\begin{aligned}\mathrm{Cov}(Z_i, Z_j) &= \mathrm{E}\Big((Z_i - \mathrm{E}(Z_i))(Z_j - \mathrm{E}(Z_j))\Big) \\ &= \mathrm{E}(Z_i Z_j) - \mathrm{E}(Z_i)\mathrm{E}(Z_j).\end{aligned}$$

- Equivalently,

$$\begin{aligned}\mathrm{Cov}(Z) &= \mathrm{E}\Big((Z - \mathrm{E}(Z))(Z - \mathrm{E}(Z))^{\mathtt{T}}\Big) \\ &= \mathrm{E}(ZZ^{\mathtt{T}}) - \mathrm{E}(Z)\mathrm{E}(Z)^{\mathtt{T}}.\end{aligned}$$

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

Random vectors

- $\mathrm{Cov}(Z)$ is always SPSD.

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

$$\mathrm{E}(AZ + b) = A\,\mathrm{E}(Z) + b$$

  and

$$\mathrm{Cov}(AZ + b) = A\,\mathrm{Cov}(Z)A^{\mathtt{T}}$$

  for any fixed (i.e., nonrandom) $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

- If $Y, Z \in \mathbb{R}^n$ are independent random vectors, then $\mathrm{Cov}(Y + Z) = \mathrm{Cov}(Y) + \mathrm{Cov}(Z)$.

Multivariate normal distribution

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

- *Standard multivariate normal*: If $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ independently and $Z = (Z_1, \ldots, Z_n)^{\mathtt{T}}$, then $Z \sim \mathcal{N}(0, I)$.

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

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

Multivariate normal distribution

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

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

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

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