Stochastic Processes

Topic 1

Probability and Linear Algebra – Review

nctuee 09f

Summary

This lecture reviews several fundamental concepts in Linear Algebra and Probability that we will see very often in this course. Specifically, I will discuss:

- Eigenvectors and eigenvalues of a matrix
- Hermitian matrices
- Singular value decomposition (SVD)
- Random variable
- Conditional probability
- Expectation and conditional expectation

Notation

We will use the following notation rules, unless otherwise noted, to represent symbols during this course.

- Boldface upper case letter to represent **MATRIX**
- Boldface lower case letter to represent **vector**
- Superscript $(\cdot)^T$ and $(\cdot)^H$ to denote transpose and hermitian (conjugate transpose), respectively

1 Linear Algebra

(1) **Eigenvector and Eigenvalue**

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. An ***eigenvector*** of $\mathbf{A}$ is a non-zero vector $\mathbf{v} \in \mathbb{C}^{n\times 1}$ such that

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}.$$

The constant $\lambda \in \mathbb{C}$ is called the ***eigenvalue*** associated with $\mathbf{v}$.

(2) **Finding Eigenvalues**

Use the fact that $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$ if and only if

$$\det(\mathbf{A} - \lambda \cdot \mathbf{I}) = 0$$

to find eigenvalues. Having obtained all the eigenvalues, solve the linear equation $(\mathbf{A} - \lambda \cdot \mathbf{I}) \cdot \mathbf{v} = 0$ to determine associated eigenvectors $\mathbf{v}'s$.

(3) **Matrix Decomposition**

Suppose that $\mathbf{A} \in \mathbb{C}^{n\times n}$ admits n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. Then, we can decompose the matrix $\mathbf{A}$ into

$$\mathbf{A} = \mathbf{E} \cdot \mathbf{\Lambda} \cdot \mathbf{E}^{-1},$$

where $\mathbf{E} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]_{n\times n}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. With this, we say the matrix $\mathbf{A}$ is diagonalizable.

Remark

— Note that NOT all square matrices have the above decompositions. There exist certain conditions for matrices to be diagonalizable. And having said that $\mathbf{A}$ have n linearly independent eigenvectors satisfies the condition.

(4) **Hermitian Matrices**

— A matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is called Hermitian if $\mathbf{A} = \mathbf{A}^H$. That is, $A_{ij} = A_{ji}^*$.

— Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Then, $\mathbf{A}$ has n orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ that form a basis for $\mathbb{C}^n$. (Orthonormal means: $\mathbf{v}_i^H \mathbf{v}_j = 0$ for $i \neq j$, and $||\mathbf{v}_i||^2 = 1$.)

(5) **Decomposition for Hermitian Matrices**

Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Then, we can decompose the matrix $\mathbf{A}$ into

$$\mathbf{A} = \mathbf{V} \cdot \mathbf{\Lambda} \cdot \mathbf{V}^H,$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]_{n \times n}$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. The matrix $\mathbf{V}$ consisting of the eigenvectors of $\mathbf{A}$ is unitary, i.e., $\mathbf{V}^H \mathbf{V} = \mathbf{I}$.

Remark: For any $\mathbf{x} \in \mathbb{C}^{n \times 1}$ and $||\mathbf{x}|| = 1$, one can show

$$\lambda_{\min} \leq \mathbf{x}^H \mathbf{A} \mathbf{x} \leq \lambda_{\max}.$$

(6) ***Singular Value Decomposition*** **(SVD)**

Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a rectangular matrix with rank r (implying that $r \leq \min(m,n)$). Then, the matrix $\mathbf{A}$ can be decomposed into

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}^H,$$

where $\mathbf{U}$ and $\mathbf{V}$ are $m \times m$ and $n \times n$ unitary matrices, respectively, and the matrix

$$\mathbf{D} = \left[\begin{array}{c|c} \mathbf{\Sigma}_{r\times r} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

is a simple structured $m \times n$ matrix with $\mathbf{\Sigma}_{r\times r} = \mathrm{diag}(\sigma_1, \cdots, \sigma_r)$. The diagonal terms of $\mathbf{\Sigma}_{r\times r}$ are called the singular values of $\mathbf{A}$, and are the square roots of the positive eigenvalues of $\mathbf{A}^H\mathbf{A}$ or $\mathbf{A}\mathbf{A}^H$.

Proof

$\longrightarrow$ First, you should know

— Nonzero eigenvalues of $\mathbf{A}^H\mathbf{A}$ and $\mathbf{A}A^H$ are identical.
— $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^H\mathbf{A}) = \mathrm{rank}(\mathbf{A}\mathbf{A}^H)$.
— Eigenvalues of $\mathbf{A}^H\mathbf{A}$ are non-negative.

$\longrightarrow$ Consider the case $m > n$. Similar proof applies to the other case.

$\longrightarrow$ It is clear that $\mathbf{A}^H\mathbf{A}$ is Hermitian, and can be decomposed into

$$\begin{aligned} \mathbf{A}^H\mathbf{A} &= \mathbf{V} \cdot \left[\begin{array}{c|c} \mathbf{\Sigma}^2_{r\times r} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]_{n\times n} \cdot \mathbf{V}^H \\ \left[\begin{array}{c} \mathbf{V}_1^H \\ \hline \mathbf{V}_2^H \end{array}\right] \mathbf{A}^H\mathbf{A}[\mathbf{V}_1 \mid \mathbf{V}_2] &= \left[\begin{array}{c|c} \mathbf{\Sigma}^2_{r\times r} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]_{n\times n}. \end{aligned}$$

Therefore, we have $\mathbf{V}_1^H\mathbf{A}^H\mathbf{A}\mathbf{V}_1 = \mathbf{\Sigma}^2_{r\times r}$ and $\mathbf{V}_2^H\mathbf{A}^H\mathbf{A}\mathbf{V}_2 = 0$.

$$\begin{cases} \mathbf{V}_1^H\mathbf{A}^H\mathbf{A}\mathbf{V}_1 = \mathbf{\Sigma}^2_{r\times r} \\ \mathbf{V}_2^H\mathbf{A}^H\mathbf{A}\mathbf{V}_2 = 0 \quad \Rightarrow \quad \mathbf{A}\mathbf{V}_2 = 0. \end{cases}$$

$\longrightarrow$ Note that

$$\mathbf{V}_1^H \mathbf{A}^H \mathbf{A} \mathbf{V}_1 = \mathbf{\Sigma}_{r\times r}^2$$

has a symmetric structure, allowing us to perform further manipulations and create a new $m \times r$ matrix

$$\mathbf{U}_1 = \mathbf{A}\mathbf{V}_1 \mathbf{\Sigma}^{-1}$$

such that $\mathbf{U}_1^H \mathbf{U}_1 = \mathbf{I}$. The above tells us that $\mathbf{U}_1$ has r orthonormal columns.

$\longrightarrow$ Expand $\mathbf{U}_1$ into an $m \times m$ unitary matrix

$$\mathbf{U} = [\mathbf{U}_1 \mid \mathbf{U}_2]$$

such that $\mathbf{U}^H \mathbf{U} = \mathbf{I}$, with

$$\begin{cases} \mathbf{U}_1^H \mathbf{U}_1 = \mathbf{I} \\ \mathbf{U}_1^H \mathbf{U}_2 = \mathbf{0}. \end{cases}$$

Then, we can prove

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}^H.$$

■

2 Probability [1]

(1) Elements of a Probabilistic Model

— The ***sample space*** Ω, which is the set of all possible ***outcomes*** of an ***experiment***.

⇒ An experiment is a process involved in every probabilistic model and will produce exactly one ***outcome***; e.g. tossing a die.

⇒ A subset of the sample space is called an ***event***.

— The **probability law**, which assigns an event A of possible outcomes a nonnegative number $P(A)$ (called the probability of A) that encodes our knowledge of belief about the collective likelihood of the elements of A.

(2) Probability Axioms

— (**Nonnegativity**) $P[A] > 0$ for every event A.

— (**Additivity**) If A and B are two disjoint events, then the probability of their union satisfies

$$P[A \cup B] = P[A] + P[B].$$

— (**Normalization**) The probability of the entire sample space Ω is equal to 1, $P[\Omega] = 1$.

(3) **Random Variable**

— A random variable is a real-valued ***function*** of the experimental ***outcome***.

— Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome.

— A function of random variable defines another random variable.

Examples of random variables:

(a) Flip a coin. Define a function $X(\text{head}) = 1$ and $X(\text{tail}) = 0$. Then, X is a random variable.

(b) In an experiment involving a sequence of 5 flips of a coin, the number of heads in the sequence is a random variable.

(b) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

Why Introducing the Notion of Random Variable?

For ***mathematical convenience***.

$\longrightarrow$ We can describe complicated events using simple math expressions by means of random variables

$\longrightarrow$ This is particularly useful when outcomes of the considered experiment do not involve with any numerical values, *e.g.* coin flip (head, tail)

Examples:

Flip a coin 3 times. Define the random variable $X_i = 1$ if the ith flip is a head, and $X_i = 0$ if tail.

— $F = \{$Two heads in 3 flips$\}$

— $G = \{$1st flip is a head, 2nd and 3rd flips have different results$\}$

$\longrightarrow$ Every event has its particular physical meaning, and can be described precisely and elegantly by properly defined random variables.

(4) **Conditional Probability**

The conditional probability of an event A, given an event B with $P[B] > 0$, is defined by

$$P[A|B] \triangleq \frac{P[A \cap B]}{P[B]}.$$

Clarification

For independent random variables X and Y, which of the following statements for an appropriate function $g(\cdot)$ is correct?

(i) $P\left[g(X,Y) \in A | Y = y_0\right] = P\left[g(X, y_0) \in A\right]$ (correct?)

(ii) $P\left[g(X,Y) \in A \cap Y = y_0\right] = P\left[g(X, y_0) \in A\right]$ (correct?)

Consider an example first.
Toss a dice twice, and let the outcome for the first toss and the second toss be X_1 and X_2, respectively. What is the probability $P\left[X_1 + X_2 \leq 8 \cap X_1 = 5\right]$? And, what is the probability $P\left[X_1 + X_2 \leq 8 | X_1 = 5\right]$?

(5) **Total Probability Theorem**

Let $A_1, \cdots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in one and only one of the events $A_1, \cdots, A_n$) and assume that $P(A_i) > 0$, for all $i = 1, \cdots, n$. Then, for any event B, we have

$$P[B] = P[A_1]P[B|A_1] + \cdots + P[A_n]P[B|A_n].$$

(6) **Bayes' Rule**

Let $A_1, \cdots, A_n$ be disjoint events that form a partition of the sample space and assume that $P(A_i) > 0$, for all $i = 1, \cdots, n$. Then, for any event B such that $P[B] > 0$, we have

$$\begin{aligned} P[A_i|B] &= \frac{P[A_i]P[B|A_i]}{P[B]} \\ &= \frac{P[A_i]P[B|A_i]}{\sum_{j=1}^{n} P[A_j]P[B|A_j]}. \end{aligned}$$

Remarks:

Bayes' rule is often used to ***infer*** the most likely unobserved cause (***statistical inference***) of a particular observed effect, by finding and comparing the conditional probabilities $P[A_i|B]$ of all possible causes A_i's given that we have observed the effect B

— The conditional probability $P[A_i|B]$ is referred to as the ***posterior*** probability, as compared to the ***prior*** probability $P[A_i]$ of the event A_i

Example: (Total Probability and Bayes' Rule)
Consider a person's chest X-ray, and let the sample space be all the possible outcomes of the X-ray images. The X-ray images that appear to have at least a shaded region is a subset of the sample space, and thus is an event, denoted by B.

Suppose we observe a shade in the person's X-ray; that is the event B is observed (the effect).

Objective:

We want to infer which of the following three mutually exclusive and collectively exhaustive potential ***causes*** is the most likely one leading the the effect B:

1. Cause 1 (event A_1): there is a malignant tumor
2. Cause 2 (event A_2): there is a nonmalignant tumor
3. Cause 3 (event A_3): this corresponds to reasons other than a tumor

Assumptions:

We assume we know the prior probabilities $P[A_i]$ and the cause-effect transition probabilities $P[B|A_i]$ for all i.

Approaches:

Given that we have observed a shade (event B occurs), find the posterior probabilities $P[A_i|B]$ for all i using Bayes' rule:

$$P[A_i|B] = \frac{P[A_i]P[B|A_i]}{P[A_1]P[B|A_1] + P[A_2]P[B|A_2] + P[A_3]P[B|A_3]}, \quad i = 1, 2, 3.$$

Choose the cause that has the ***largest posterior probability*** to be the most likely cause.

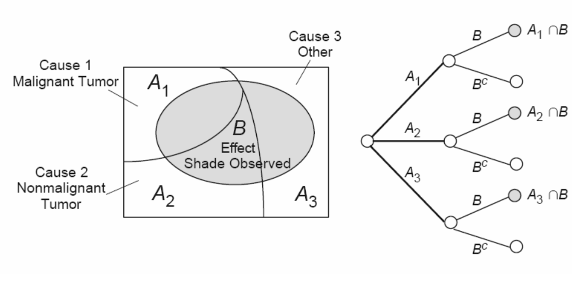

Figure 1: Illustration of the above example.

(7) **Expectation**

We define the expected value of a discrete random variable X, with prob. mass function (PMF) $p_X(x)$ by

$$E[X] \triangleq \sum_x x p_X(x),$$

where $p_X(x) = P[X = x]$.

(8) **Expectation for Functions of Random Variables**

Let X be a random variable with PMF $p_X(x)$, and let $g(X)$ be a real-valued function of X. Then, the expected value of the random variable $g(X)$ is given by

$$E[g(X)] = \sum_x g(x) p_X(x).$$

The above can be extended to continuous case.

Example:

Let X be a random variable with $P[X = -1] = 0.2$, $P[X = 0] = 0.5$, and $P[X = 1] = 0.3$. Find $E[X^2]$.

(9) **Joint PMF**

The joint PMF of two discrete random variables X and Y is defined by

$$p_{X,Y}(x,y) = P[X = x, Y = y]$$

for all pairs of numerical values (x, y) that X and Y can take. (For notational convenience, we use $P[X = x, Y = y]$ to mean $P[X = x \cap Y = y]$).

(10) **Marginal PMF from Joint PMF**

The marginal PMF $p_X(x)$ and $p_Y(y)$ can be calculated using

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

(11) **Conditional PMF**

The conditional PMF $p_{X|A}(x)$ of a random variable X, conditioned on a particular event A with $P[A] > 0$, is defined by

$$p_{X|A}(x) = P[X = x|A] = \frac{P[\{X = x\} \cap A]}{P[A]}.$$

(12) **Marginal PMF from Conditional PMF**

The marginal PMF $p_X(x)$ can be calculated using

$$\begin{aligned} p_X(x) &= P[X = x] = \sum_y p_{X,Y}(x,y) \\ &= \sum_y p_Y(y) p_{X|Y}(x|y) \\ &= E[\, \underbrace{p_{X|Y}(x|Y)}_{\text{a function of } Y} \,] \end{aligned}$$

(13) **Conditional Expectation**

The conditional expectation of X given a value y of Y is defined by

$$\begin{aligned} E[X|Y=y] &\triangleq \sum_x x p_{X|Y}(x|y) \\ &= \sum_x x P[X=x|Y=y] \end{aligned}$$

Remarks about Conditional Expectation

(a) $E[X|Y=y]$ is a number whose value depends on y.

(b) $E[X|Y]$ is a function of the random variable Y, hence is a ***random variable***.

(14) **Cumulative Distribution Function (CDF)**

The CDF, or sometimes called probability distribution function, of a random variable X is denoted by F_X and provides the probability $P[X \leq x]$. In particular, for continuous random variable X, we have

$$F_X(x) \triangleq P[X \leq x] = \int_{-\infty}^{x} f_X(\alpha) d\alpha,$$

where $f_X(\alpha)$ is the probability density function of X.

Remarks

(a) The probability density function (pdf) can be calculated by

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

(b) We know $P[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1)$.

(15) **Conditional Density Function**

Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. For any fixed y with $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$, is defined by

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

(16) **Total Probability in Density Version**

The probability density function $f_Y(y)$ of a continuous random variable Y can be evaluated by

$$f_Y(y) = \sum_i P_N[i] f_{Y|N}(y|i).$$

(17) **Conditional Probability on a Continuous Random Variable**

We are often interested in knowing the conditional probability $P[N = n|Y = y]$ conditioned on a continuous random variable Y at $Y = y$. This is given by

$$P[N = n|Y = y] = \frac{P_N[n] f_{Y|N}(y|n)}{\sum_i P_N[i] f_{Y|N}(y|i)}.$$

Example:

A binary signal $S \in \{-1, +1\}$ is transmitted, and we are given that $P(S = 1) = P(S = -1) = 1/2$. The received signal at the receiver is

$$Y = S + N,$$

where N is normal noise, with zero mean and variance σ^2, independent of S.

What is the probability that $S = -1$, given that we have observed $Y = y$?

References

[1] Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd edition, 2008.