

4 Lecture 4: Jan 25

Last time

- Linear algebra: vector and vector space, rank of a matrix
- Column space (JM Appendix A)

Today

- Nullspace
- Probability review

Null space

Definition: The null space of a matrix, denoted by $\mathcal{N}(\mathbf{A})$, is $\mathcal{N}(\mathbf{A}) = \{\mathbf{y} : \mathbf{A}\mathbf{y} = \mathbf{0}\}$.

Result A.3

If \mathbf{A} has full-column rank, then $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.

proof:

Theorem A.1

Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\dim(\mathcal{C}(\mathbf{A})) = r$ and $\dim(\mathcal{N}(\mathbf{A})) = n - r$, where $r = \text{rank}(\mathbf{A})$.

See JM Appendix Theorem A.1 for the proof.

proof: Denote $\dim(\mathcal{N}(\mathbf{A}))$ by k , to be determined, and construct a set of basis vectors for $\mathcal{N}(\mathbf{A}) : \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \dots, \mathbf{u}^{(k)}\}$, so that $\mathbf{A}\mathbf{u}^{(i)} = \mathbf{0}$, for $i = 1, 2, \dots, k$. Now, construct a basis for \mathbb{R}^n by adding the vectors $\{\mathbf{u}^{(k+1)}, \dots, \mathbf{u}^{(n)}\}$, which are not in $\mathcal{N}(\mathbf{A})$. Clearly, $\mathbf{A}\mathbf{u}^{(i)} \in \mathcal{C}(\mathbf{A})$ for $i = k+1, \dots, n$, and so the span of these vectors form a subspace of $\mathcal{C}(\mathbf{A})$. These vectors $\{\mathbf{A}\mathbf{u}^{(i)}, i = k+1, \dots, n\}$ are also linearly independent from the following argument: suppose $\sum_{i=k+1}^n c_i \mathbf{A}\mathbf{u}^{(i)} = \mathbf{0}$; then $\sum_{i=k+1}^n c_i \mathbf{A}\mathbf{u}^{(i)} = \mathbf{A} [\sum_{i=k+1}^n c_i \mathbf{u}^{(i)}] = \mathbf{0}$, and hence $\sum_{i=k+1}^n c_i \mathbf{u}^{(i)}$ is a vector in $\mathcal{N}(\mathbf{A})$. Therefore, there exist b_i such that $\sum_{i=k+1}^n c_i \mathbf{u}^{(i)} = \sum_{i=1}^k b_i \mathbf{u}^{(i)}$, or $\sum_{i=1}^k b_i \mathbf{u}^{(i)} - \sum_{i=k+1}^n c_i \mathbf{u}^{(i)} = \mathbf{0}$. Since $\{\mathbf{u}^{(i)}\}$ form a basis for \mathbb{R}^n , c_i must all be zero. Therefore $\mathbf{A}\mathbf{u}^{(i)}, i = k+1, \dots, n$ are linearly independent. At this point, since $\text{span}\{\mathbf{A}\mathbf{u}^{(k+1)}, \dots, \mathbf{A}\mathbf{u}^{(n)}\} \subseteq \mathcal{C}(\mathbf{A})$, $\dim(\mathcal{C}(\mathbf{A}))$ is at least $n - k$. Suppose there is a vector \mathbf{y} that is in $\mathcal{C}(\mathbf{A})$, but not in the span; then there exists $\mathbf{u}^{(n+1)}$ so that $\mathbf{y} = \mathbf{A}\mathbf{u}^{(n+1)}$ and $\mathbf{u}^{(n+1)}$ is linearly independent of $\{\mathbf{u}^{(k+1)}, \dots, \mathbf{u}^{(n)}\}$ (and clearly not in $\mathcal{N}(\mathbf{A})$), making $n+1$ linearly independent vectors in \mathbb{R}^n . Since that is not possible, the span is equal to $\mathcal{C}(\mathbf{A})$ and $\dim(\mathcal{C}(\mathbf{A})) = n - k = r = \text{rank}(\mathbf{A})$, so that $k = \dim(\mathcal{N}(\mathbf{A})) = n - r$.

Interpretation: “dimension of column space + dimension of null space = # columns”

Mis-Interpretation: Columns space and null space are orthogonal complement to each other. They are of different orders in general! Next result gives the correct statement.

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- [Review of Probability Theory](#) by Arian Maleki and Tom Do

Probability theory review

A few basic elements to define a probability on a set:

- **Sample space** S is the set that contains all possible outcomes of a particular experiment.
- An **event** is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).
- Event operations

1. Union: The union of A and B , written $A \cup B$, is the set of elements that belong to either A or B or both:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

2. Intersection: The intersection of A and B , written $A \cap B$, is the set of elements that belong to both A and B :

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

3. Complementation: The complement of A , written as A^c , is the set of all elements that are not in A :

$$A^c = \{x : x \notin A\}.$$

- **Sigma algebra (or Borel field):** A collection of subsets of S is called a sigma algebra (or Borel field), denoted by \mathcal{B} , if it satisfies the following three properties:

1. $\emptyset \in \mathcal{B}$ (the empty set is an element of \mathcal{B})
2. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ (\mathcal{B} is closed under complementation).
3. If $A_1, A_2, \dots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ (\mathcal{B} is closed under countable unions).

- **Axioms of probability:** Given a sample space S and an associated sigma algebra \mathcal{B} , a *probability function* is a function $\Pr()$ with domain \mathcal{B} that satisfies

1. $\Pr(A) \geq 0$ for all $A \in \mathcal{B}$
2. $\Pr(S) = 1$.
3. If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$.

Properties:

If $\Pr()$ is a *probability function* and A and B are any sets in \mathcal{B} , then

- $\Pr(\emptyset) = 0$, where \emptyset is the empty set

Proof:

- $\Pr(A) \leq 1$

Proof:

- $\Pr(A^c) = 1 - \Pr(A)$

Proof:

- $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$

Proof:

- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Proof:

- $\Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

- If $A \subset B$, then $\Pr(A) \leq \Pr(B)$.

Proof:

Conditional probability

Definition: If A and B are events in S , and $\Pr(B) > 0$, then the conditional probability of A given B , written $\Pr(A|B)$, is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Note that what happens in the conditional probability calculation is that B becomes the sample space: $\Pr(B|B) = 1$, in other words, $\Pr(A|B)$ is the probability measure of the event A after observing the occurrence of event B .

Definition: Two events A and B are statistically independent if $\Pr(A \cap B) = \Pr(A) \Pr(B)$. When A and B are independent events, then $\Pr(A|B) = \Pr(A)$ and the following pairs are also independent

- A and B^c

proof:

- A^c and B

- A^c and B^c

Random variables

Definition: A random variable is a function from a sample space S into the real numbers.

Experiment	Random variable
Toss two dice	$X = \text{sum of the numbers}$
Toss a coin 25 times	$X = \text{number of heads in 25 tosses}$
Apply different amounts of fertilizer to corn plants	$X = \text{yield/acre}$

Suppose we have a sample space

$$S = \{s_1, \dots, s_n\}$$

with a probability function \Pr and we define a random variable X with range $\mathcal{X} = \{x_1, \dots, x_m\}$. We can define a probability function \Pr_X on \mathcal{X} in the following way. We will observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Thus,

$$\Pr_X(X = x_i) = \Pr(\{s_j \in S : X(s_j) = x_i\}).$$

We will simply write $\Pr(X = x_i)$ rather than $\Pr_X(X = x_i)$.

A note on notation: Random variables are often denoted with uppercase letters and the realized values of the variables (or its range) are denoted by corresponding lowercase letters.

Distribution functions

Definition: The cumulative distribution function or cdf of a random variable (r.v.) X , denoted by $F_X(x)$ is defined by

$$F_X(x) = \Pr(X \leq x), \text{ for all } x.$$

The function $F(x)$ is a cdf if and only if the following three conditions hold:

1. $\lim_{x \rightarrow \infty} F(x) = 1$.
2. $F(x)$ is a nondecreasing function of x .
3. $F(x)$ is right-continuous; that is, for every number x_0 , $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

Definition: A random variable X is continuous if $F(x)$ is a continuous function of x . A random variable X is discrete if $F(x)$ is a step function of x .

The following two statements are equivalent:

1. The random variables X and Y are identically distributed.
2. $F_X(x) = F_Y(x)$ for every x .

Density and mass functions

Definition: The probability mass function (pmf) of a discrete random variable X is given by

$$f_X(x) = \Pr(X = x) \text{ for all } x.$$

Example (Geometric probabilities) For the geometric distribution, we have the pmf

$$f_X(x) = \Pr(X = x) = \begin{cases} p(1-p)^{x-1} & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

Definition: The probability density function or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \quad \text{for all } x.$$

A note on notation: The expression “ X has a distribution given by $F_X(x)$ ” is abbreviated symbolically by “ $X \sim F_X(x)$ ”, where we read the symbol “ \sim ” as “is distributed as”.

Example (Logistic distribution) For the logistic distribution, we have

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

and, hence,

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if

1. $f_X(x) \geq 0$ for all x
2. $\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (pdf).

Expectations

The expected value, or expectation, of a random variable is merely its average value, where we speak of “average” value as one that is weighted according to the probability distribution.

Definition: The expected value or mean of a random variable $g(X)$, denoted by $\mathbf{E}(g(X))$, is

$$\mathbf{E}(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x)f_X(x) = \sum_{x \in \mathcal{X}} g(x)\Pr(X = x) & \text{if } X \text{ is discrete,} \end{cases}$$

Exponential mean

Suppose $X \sim \text{Exp}(\lambda)$ distribution, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda}e^{-x/\lambda}, \quad 0 \leq x < \infty, \quad \lambda > 0$$

Then $\mathbf{E}(X)$ is:

Binomial mean

If X has binomial distribution, i.e. $X \sim \text{binomial}(n, p)$, its pmf is given by

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n,$$

where n is a positive integer, $0 \leq p \leq 1$, and for every fixed pair n and p the pmf sums to 1. The expected value of a binomial random variable is then given by

$$\mathbf{E}(X) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x}$$

Now, use the identity $x \binom{n}{x} = n \binom{n-1}{x-1}$ to derive the Expected value.

properties:

Let X be a random variable and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

1. $\mathbf{E}(a \cdot g_1(X) + b \cdot g_2(X) + c) = a\mathbf{E}(g_1(X)) + b\mathbf{E}(g_2(X)) + c.$
2. If $g_1(x) \geq 0$ for all x , then $\mathbf{E}(g_1(X)) \geq 0.$
3. If $g_1(x) \geq g_2(x)$ for all x , then $\mathbf{E}(g_1(X)) \geq \mathbf{E}(g_2(X)).$
4. If $a \leq g_1(x) \leq b$ for all x , then $a \leq \mathbf{E}(g_1(X)) \leq b.$

Moments

The various moments of a distribution are an important class of expectations.

Definition: For each integer n , the n^{th} moment of X (or $F_X(x)$), μ'_n , is

$$\mu'_n = \mathbf{E}(X^n).$$

The n^{th} central moment of X , μ_n , is

$$\mu_n = \mathbf{E}((X - \mu)^n),$$

where $\mu = \mu'_1 = \mathbf{E}(X).$

Variance

Definition: The variance of a random variable X is its second central moment, $\mathbf{Var}(X) = \mathbf{E}((X - EX)^2)$. The positive square root of $\mathbf{Var}(X)$ is the standard deviation of X .

Exponential variance

Let X have the exponential(λ) distribution, $X \sim \text{Exp}(\lambda)$. Then the variance of X is

properties

1. $\mathbf{Var}(aX + b) = a^2 \mathbf{Var}(X)$.

proof:

2. $\mathbf{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$.

proof:

Moment generating function

Definition: Let X be a random variable with cdf F_X . The moment generating function or mgf of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = \mathbf{E}(e^{tX}),$$

provided that the expectation exists for t in some neighborhood of 0. That is, there exists an $h > 0$ such that for all t in $-h < t < h$, $\mathbf{E}(e^{tX})$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

Property: If X has mgf $M_X(t)$, then

$$\mathbf{E}(X^n) = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \left. \frac{d^n}{dt^n} M_X(t) \right|_{t=0}.$$