

Section 01: Probability and Linear Algebra Review

Definitions

Norms

We haven't covered norms yet but they are incredibly useful, and they show up quite often! We will cover them later in more detail, but for now it is sufficient to get yourself familiar with definitions for some of the most widely used norms! For any vector v that is n -dimensional, i.e. $v \in \mathbb{R}^n$, we have the following

- (a) **One-norm** (ℓ_1): $\|v\|_1 = \sum_{i=1}^n |v_i|$
- (b) **Two-norm** (ℓ_2): $\|v\|_2 = \sqrt{v^T v} = \sqrt{\sum_{i=1}^n v_i^2}$
- (c) **∞ -norm**: $\|v\|_\infty = \max_i |v_i|$

Symmetric Matrices and the Quadratic Form

Let us define a matrix $A \in \mathbb{R}^{n \times n}$.

- (a) We have that the matrix A is symmetric iff $A = A^T$
- (b) The quadratic form is defined to be $x^T A x$ for any vector $x \in \mathbb{R}^n$. The matrix A is said to be positive semi-definite if $x^T A x \geq 0$

1. Probability Review: PDF, CDF and Expectation

The **Cumulative Density Function** (CDF) $F_X : \mathbb{R} \rightarrow [0, 1]$ of a random variable X is defined as $\mathbb{P}(X \leq x)$. The **Probability Density Function** (PDF), $f_X : \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$, of the same random variable is defined as $f_X(x) = \frac{d}{dx} F_X(x)$.

Note that the CDF can be computed from the PDF, and vice versa; e.g. $F_X = \int_{-\infty}^x f(x) dx$.

We can use these functions to directly compute the expectation of random variables, since the expectation is defined in terms of the PDF: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$.

These functions can also be used to compute the distribution of any one-to-one transformation $g(\cdot)$ of the random variable: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx$.

Note: this section focuses on the continuous case, but equivalent formulations hold in the discrete case by replacing integration with summation.

- (a) You've just started a new exercise regimen. You start on the 2nd floor of CSE1, and then make a random choice:
 - With probability p_1 you run up 2 flights of stairs.
 - With probability p_2 you run up 1 flight of stairs.
 - With probability p_3 you walk down 1 flight of stairs.

Where $p_1 + p_2 + p_3 = 1$.

You will do two iterations of your exercise scheme (with each draw being independent). Let X be the floor you're on at the end of your exercise routine. Recall you start on floor 2.

- (i) Let Y be the difference between your ending floor and your starting floor in one iteration. What is $\mathbb{E}[Y]$ (in terms of p_1, p_2, p_3)?

(ii) What is $\mathbb{E}[X]$ (use your answer from the previous part)

(iii) You change your scheme: instead of doing two independent iterations, you decide the second iteration of your regimen will just use the same random choice as your first (in particular they are no longer independent!). Does $\mathbb{E}[X]$ change? (Optional)

Fact 1. Let $X_{(j)}$ denote the j th order statistic in a sample of i.i.d. random variables; that is, the j th element when the items are sorted in increasing order $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$.

The PDF of $X_{(j)}$ is given by:

$$f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!} [F(x)]^{j-1} [1-F(x)]^{n-j} f(x). \quad (1)$$

(b) When a sample of $2N + 1$ i.i.d. random variables is observed, the $(N + 1)^{\text{st}}$ smallest is called the sample median. If a sample of size 3 from a uniform distribution over $[0, 1]$ is observed, find the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$. Hint: use Fact 1.

2. Linear Algebra Review

Let $X \in \mathbb{R}^{m \times n}$. X may not have full rank. We explore properties about the four fundamental subspaces of X .

2.1. Summation form v.s. Matrix form

(a) Let $w \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Let $x_i^\top$ denote each row in X and y_i in Y . Show $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

(b) Let $L(w) = \|Xw - Y\|_2^2$. What is $\nabla_w L(w)$? (Hint: You can use either summation or matrix form from first sub-problem).

2.2. Subspaces of X

What is the row space, column space, nullspace, and rank of $X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$?

2.3. Connections between subspaces of X

Check the following facts.

- (a) The row space of X is the column space of $X^\top$, and vice versa.
- (b) The nullspace of X and the row space of X are orthogonal complements. This can be written in shorthand as $\text{Null}(X) = \text{Range}(X^\top)^\perp$. This is further equivalent to saying $\text{Range}(X^\top) = \text{Null}(X)^\perp$.
- (c) The nullspace of $X^\top$ is orthogonal to the column space of X . This can be written in shorthand as $\text{Null}(X^\top) = \text{Range}(X)^\perp$.

2.4. Linear algebra facts for linear regression

We saw in lecture on Linear Regression that the closed form expression for linear regression without an offset involves the term $(X^\top X)^{-1}$.

- (a) Is it true that the matrix $X^\top X$ is always symmetric and positive semidefinite?
- (b) State and prove the connection between the nullspace of X and the nullspace of $X^\top X$. That is, your statement should look like one of the following: $\text{Null}(X) \subseteq \text{Null}(X^\top X)$, or $\text{Null}(X) \supseteq \text{Null}(X^\top X)$ or $\text{Null}(X) = \text{Null}(X^\top X)$.
- (c) Is it true that $X^\top X$ is always invertible?
- (d) Based on the above fact about the connection between the nullspaces of X and $X^\top X$ and the expression for linear regression without an offset (that we referred to two problems above), justify the use of “tall skinny” data matrix X as opposed to a “short wide” matrix X .
- (e) The column space and row space of $X^\top X$ are the same, and are equal to the row space of X . (Hint: Use the relationship between nullspace and row space.)