ENCS5341
Machine Learning and Data Science

Linear Algebra and Probability Review

Slides are based on Stanford CS229 course

Linear Algebra

Basic Notation

- By $x \in \mathbb{R}^n$, we denote a vector with n entries.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

The Identity Matrix

- The identity matrix, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A$$

- Ex:

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad , \qquad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Diagonal matrices

- A diagonal matrix is a matrix where all non-diagonal elements are 0. This is typically denoted $D = diag(d_1, d_2, \dots, d_n)$, with

$$D_{ij} = \begin{cases} d_i, & i = j \\ 0, & i \neq j \end{cases}$$

- For example the identity matrix $I = diag\ (1,1, \dots, 1)$

Vector-Vector Product

- inner product or dot product

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

- outer product

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}.$$

Matrix-Vector Product

- If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

Matrix-Vector Product

- If we write A by columns, then we have:

$$y = Ax = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a^1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a^2 \\ \\ \end{bmatrix} x_2 + \ldots + \begin{bmatrix} \\ a^n \\ \\ \end{bmatrix} x_n \ .$$

y is a linear combination of the columns of A.

Matrix-Vector Product

- It is also possible to multiply on the left by a row vector.
  - If we write A by columns, then we can express $x^T A$ as,

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a^1 & x^T a^2 & \cdots & x^T a^n \end{bmatrix}$$

Matrix-Vector Product

- It is also possible to multiply on the left by a row vector.
  - expressing A in terms of rows we have

$$y^T = x^T A = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix} \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}$$

$$= x_1 \begin{bmatrix} — & a_1^T & — \end{bmatrix} + x_2 \begin{bmatrix} — & a_2^T & — \end{bmatrix} + \ldots + x_m \begin{bmatrix} — & a_m^T & — \end{bmatrix}$$

y^T is a linear combination of the rows of A.

Matrix-Matrix Multiplication (different views)

1. As a set of vector-vector products (dot product)

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b^1 & a_1^T b^2 & \cdots & a_1^T b^p \\ a_2^T b^1 & a_2^T b^2 & \cdots & a_2^T b^p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b^1 & a_m^T b^2 & \cdots & a_m^T b^p \end{bmatrix}$$

2. As a sum of outer products

$$C = AB = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^p \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^T & — \\ — & b_2^T & — \\ & \vdots & \\ — & b_p^T & — \end{bmatrix} = \sum_{i=1}^{p} a^i b_i^T$$

Matrix-Matrix Multiplication (different views)

3. As a set of matrix-vector products.

$$C = AB = A \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab^1 & Ab^2 & \cdots & Ab^n \\ | & | & & | \end{bmatrix}$$

4. As a set of vector-matrix products

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} B = \begin{bmatrix} — & a_1^T B & — \\ — & a_2^T B & — \\ & \vdots & \\ — & a_m^T B & — \end{bmatrix}$$

Matrix-Matrix Multiplication (properties)

- Associative: (AB)C = A(BC).
- Distributive: A(B + C) = AB + AC.
- In general, not commutative; that is, it can be the case that AB ≠ BA. (For example, if $A \in \mathbb{R}^{m \times n}$ and B ∈ $\mathbb{R}^{n \times q}$, the matrix product BA does not even exist if m and q are not equal!)

The Transpose

- The transpose of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the n × m matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

- The following properties of transposes are easily verified:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$

Norms

- A norm of a vector ||x|| is informally a measure of the “length” of the vector.

- More formally, a norm is any function f : $\mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Examples of Norms

The commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

The ℓ_1 norm,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

The ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

The Inverse of a Square Matrix

- The ***inverse*** of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

- We say that A is ***invertible*** or ***non-singular*** if A^{-1} exists and ***non-invertible*** or ***singular*** otherwise.

- In order for a square matrix A to have an inverse A^{-1}, then A must be full rank.

- Properties (Assuming $A, B \in \mathbb{R}^{n \times n}$ are non-singular):
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T}.

Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector*** if

$$Ax = \lambda x, \quad x \neq 0.$$

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x, but scaled by a factor λ.

Probability Theory

Definitions, Axioms, and Corollaries

- Performing an **experiment** $\rightarrow$ **outcome**
- **Sample Space** (S): set of all possible outcomes of an experiment
- **Event** (E): a subset of S ($E \subseteq S$)
- **Probability** (**Bayesian** definition)

  A number between 0 and 1 to which we ascribe meaning
  i.e. our belief that an event E occurs
- **Frequentist** definition of probability

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

Definitions, Axioms, and Corollaries

Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

Axiom 3: If E and F are mutually exclusive ($E \cap F = \emptyset$), then $P(E) + P(F) = P(E \cup F)$

Corollary 1: $P(E^C) = 1 - P(E) \quad (= P(S) - P(E))$

Corollary 2: $E \subseteq F$, then $P(E) \leq P(F)$

Corollary 3: $P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Let's apply conditional probability to obtain **Bayes' Rule**!

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

Conditioned Bayes' Rule: given events A, B, C,

$$P(A \mid B, C) = \frac{P(B \mid A, C)P(A \mid C)}{P(B \mid C)}$$

Law of Total Probability

Let $B_1, \ldots, B_n$ be n disjoint events whose union is the entire sample space. Then, for any event A,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

$$= \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

We can then write Bayes' Rule as:

$$P(B_k \mid A) = \frac{P(B_k) P(A \mid B_k)}{P(A)}$$

$$= \boxed{\frac{P(B_k) P(A \mid B_k)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}}$$

Law of Total Probability

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins. Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**? [1]

Solution:

$$
\begin{aligned}
P(A \mid G) &= \frac{P(A)P(G \mid A)}{P(A)P(G \mid A) + P(B)P(G \mid B)} \\
&= \frac{0.5 \times 1}{0.5 \times 1 + 0.5 \times 0.6} \\
&= \boxed{0.625}
\end{aligned}
$$

Chain Rule

For any n events $A_1, ..., A_n$, the joint probability can be expressed as a product of conditionals:

$$P(A_1 \cap A_2 \cap ... \cap A_n)$$
$$= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2 \cap A_1)...P(A_n \mid A_{n-1} \cap A_{n-2} \cap ... \cap A_1)$$

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Implication: If two events are independent, observing one event does not change the probability that the other event occurs.

In general: events $A_1, \ldots, A_n$ are **mutually independent** if

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

for any subset $S \subseteq \{1, \ldots, n\}$.

Random Variables

- A **random variable** X is a variable that probabilistically takes on different values. It maps outcomes to real values
- X takes on values in $Val(X) \subseteq \mathbb{R}$ or Support $\text{Sup}(X)$
- $X = k$ is the **event** that random variable X takes on value k

Discrete RVs:

- $Val(X)$ is a set
- $P(X = k)$ can be nonzero

Continuous RVs:

- $Val(X)$ is a range
- $P(X = k) = 0$ for all k. $P(a \leq X \leq b)$ can be nonzero.

Probability Mass Function (PMF)

Given a **discrete** RV X, a PMF maps values of X to probabilities.

$$p_X(x) := p(x) := P(X = x)$$

For a valid PMF, $\sum_{x \in Val(x)} p_X(x) = 1$.

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a probability (i.e. $\mathbb{R} \to [0, 1]$)

$$F_X(a) := F(a) := P(X \leq a)$$

A CDF must fulfill the following:

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- If $a \leq b$, then $F_X(a) \leq F_X(b)$ (i.e. CDF must be nondecreasing)

Also note: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

Probability Density Function (PDF)

PDF of a continuous RV is simply the derivative of the CDF.

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

Thus,

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

A valid PDF must be such that

- for all real numbers x, $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X:

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x) p_X(x)$$

- If X is a continuous RV with PDF f_X:

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Intuitively, expectation is a weighted average of the values of $g(x)$, weighted by the probability of x.

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Linearity of Expectation

Given n real-valued functions $f_1(X), ..., f_n(X)$,

$$\mathbb{E}[\sum_{i=1}^{n} f_i(X)] = \sum_{i=1}^{n} \mathbb{E}[f_i(X)]$$

Variance

The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$\begin{aligned} Var(X) &:= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

Interpretation: $Var(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbb{R}$, real-valued function $f(X)$

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, ...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, ...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ for all $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

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

  Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$
- **Marginal PMF** of X, given joint PMF of X, Y:

$$p_X(x) = \sum_y p_{XY}(x, y)$$

Joint and Marginal Distributions

- **Joint PDF** for continuous X, Y:

$$f_{XY}(x,y) = \frac{\delta^2 F_{XY}(x,y)}{\delta x \delta y}$$

  Note that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$

- **Marginal PDF** of X, given joint PDF of X, Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$