CSE422: Artificial Intelligence
Linear Algebra & Probability Review

Rafiad Sadat Shahir

1 Linear Algebra

Linear algebra provides a way to compactly represent and operate on sets of linear equations.

1.1 Basic Terminologies

Tensors: In machine learning, the term tensor informally refers to multidimensional arrays.
Matrix: A matrix is a rank-2 tensor. By $A \in \mathbb{R}^{n \times m}$, we denote a matrix with m rows and n columns, where the entries of A are real numbers.
Vector: A vector is a rank 1 tensor. A matrix is a rank-1 tensor. By $x \in \mathbb{R}^n$, we denote a vector with n entries. By convention, a vector of n-dimension is often thought of as a matrix with n rows and 1 column, known as a column vector. If we want to explicitly represent a row vector by a matrix with 1 row and n columns, we typically write x^T (transpose of x).
Scalar: A scalar is a rank-0 tensor.

1.2 Products

Vector-Vector Products: Given two vectors $x, y \in \mathbb{R}^n$, the quantity $x^T y$ (inner product or dot product of the vectors) is a real number given by

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

Matrix-Vector Products: Given a matrix $A \in \mathbb{R}^{n \times m}$ and a vector $x \in \mathbb{R}^m$, their product is a vector $y = Ax \in \mathbb{R}^n$. y can be expressed as

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_n^T x \end{bmatrix}$$

Matrix-Matrix Products: The product of two matrices $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times m}$ is the matrix $C = AB \in \mathbb{R}^{n \times m}$. C can be expressed as

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \dots & b_m \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \dots & a_1^T b_m \\ a_2^T b_1 & a_2^T b_2 & \dots & a_2^T b_m \\ \vdots & \vdots & \ddots & \vdots \\ a_n^T b_1 & a_n^T b_2 & \dots & a_n^T b_m \end{bmatrix}$$

2 Probability

Probability theory is the study of uncertainty. We will be relying on concepts from probability theory to derive machine learning algorithms.

2.1 Axioms of Probability

Sample Space: The set of all the outcomes of a random experiment is referred to as the sample space Ω.
Set of Events: A set of events, denoted by $\mathcal{F}$, is a set whose elements are subsets of Ω.
The following three properties of a probability function $P : \mathcal{F} \to \mathbb{R}$ are called axioms of probability:

1. $P(A) \geq 0$, for all $A \in \mathcal{F}$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \dots A_n$ are disjoint events then

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

2.2 Random Variables

A random variable X is a function $X : \Omega \to \mathbb{R}$. Random variables are broadly classified into two types: discrete and continuous.

2.2.1 Probability Mass Function

A way to represent the probability measure associated with a discrete random variable is the probability mass function (PMF). In particular, a PMF is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x).$$

2.2.2 Cumulative Distribution Function

A cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ that specifies a probability measure as

$$F_X(x) \triangleq P(X \leq x).$$

2.2.3 Probability Density Function

For some continuous random variables, the cumulative distribution function is differentiable. In these cases, we define the probability density function (PDF) as the derivative of the CDF, i.e.

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

2.2.4 Some Common Random Variables

The following are some common random variables:

- $X \sim Bernoulli(\phi)$, where $0 \leq \phi \leq 1$

$$p(x) = \phi^x (1 - \phi)^{1-x}$$

- $X \sim Normal(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

2.2.5 Expectation

Let X be a discrete random variable with a PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$. We define the expectation or expected value of g(X) as

$$E[g(X)] \triangleq \sum_{x \in V(X)} g(x) p_X(x)$$

If X is a continuous random variable with a PDF $f_X(x)$, the expected value of g(X) is defined as

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x)$$

2.3 Two Discrete Random Variables

2.3.1 Joint and Marginal PMFs

If X and Y are discrete random variables, the joint PMF $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ can be expressed as

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

In addition, the marginal PMF p_{XY} can be expressed as

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

2.3.2 Conditional Distributions

The conditional PMF $p_{Y|X}(y|x)$ of Y given X can be expressed as

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

assuming that $p_X(x) \neq 0$

2.3.3 Bayes Rule

A useful formula that often arises when trying to derive an expression for the conditional probability of one variable given another is the Bayes rule. In the case of discrete random variables X and Y,

$$P_{Y|X}(y|x) = \frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum\limits_{y_i \in Y} P_{X|Y}(x|y_i)P_Y(y_i)}$$

Generally, $P_{Y|X}(y|x)$, $P_{X|Y}(x|y)$, $P_Y(y)$, and $P_X(x)$ are called posterior probability, likelihood, prior probability, and marginal probability, respectively.

2.3.4 Independence

Two discrete random variables X and Y are independent if

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

for all $x \in V(X), y \in V(Y)$. In addition, $p_{Y|X}(y|x) = p_Y(y)$ for all $y \in V(Y)$ whenever $p_X(x) \neq 0$.

2.3.5 Conditional Independence

Two discrete random variables X and Y are conditionally independent given Z if

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$$

for all $x \in V(X), y \in V(Y), z \in V(Z)$.