Linear Algebra and Probability Review[1]

Xin Bing

Department of Statistical Sciences
University of Toronto

[1] Slides adapted from Ian Goodfellow's *Deep Learning* textbook lectures

About this tutorial

- Not a comprehensive survey of all of linear algebra and probability.
- Focused on the subset most relevant to machine learning.

Scalars

- A scalar is a single number
- Integers, real numbers, rational numbers, etc.
- Typically denoted in italic font:

$$a, n, x$$

Vectors

- A vector is an array of d numbers
- x_i be integer, real, binary, etc.
- Notation to denote type and size:

$$\boldsymbol{x} \in \mathbb{R}^d$$

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

Matrices

- A matrix is an array of numbers with two indices
- $A_{i,j}$ be integer, real, binary, etc.
- Notation to denote type and size:

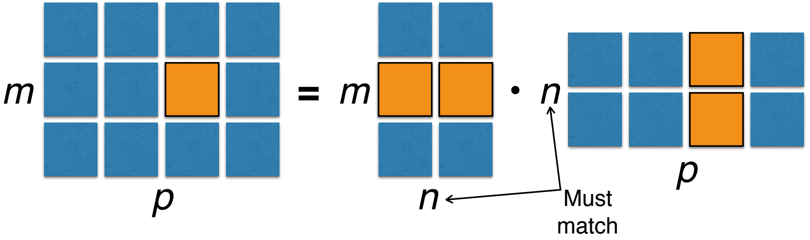
$$A \in \mathbb{R}^{m \times n}$$

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

Matrix (Dot) Product

Matrix product AB is the matrix such that

$$(AB)_{i,j} = \sum_k A_{i,k} B_{k,j}.$$



(Goodfellow 2016)

This also defines matrix-vector products $A\boldsymbol{x}$ and $\boldsymbol{x}^\top A$.

Identity Matrix

The identity matrix for $\mathbb{R}^d$ is the matrix I_d such that

$$\forall \mathbf{x} \in \mathbb{R}^d, \; I_d \mathbf{x} = \mathbf{x}$$

For example, I_3:

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix Transpose

The transpose of a matrix A is the matrix $A^\top$ such that $(A^\top)_{i,j} = A_{j,i}$.

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \implies A^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

The transpose of a matrix can be thought of as a mirror image across the main diagonal. The transpose switches the order of the matrix product.

$$(AB)^\top = B^\top A^\top$$

Systems of equations

The matrix equation

$$A\mathbf{x} = \boldsymbol{b}$$

expands to

$$\begin{aligned} A_{1,1}x_1 + A_{1,2}x_2 + \cdots A_{1,n}x_n &= b_1 \\ A_{2,1}x_1 + A_{2,2}x_2 + \cdots A_{2,n}x_n &= b_2 \\ &\vdots \\ A_{m,1}x_1 + A_{m,2}x_2 + \cdots A_{m,n}x_n &= b_m \end{aligned}$$

Solving Systems of Equations

A linear system of equations can have:

- No solution
- Many solutions
- Exactly one solution, i.e. multiplying by the matrix is an invertible function

Matrix Inversion

The matrix inverse of A is the matrix A^{-1} such that

$$A^{-1}A = I_d$$

Solving a linear system using an inverse:

$$\begin{aligned} A\mathbf{x} &= \boldsymbol{b} \\ A^{-1}A\mathbf{x} &= \boldsymbol{b} \\ I_d\mathbf{x} &= A^{-1}\boldsymbol{b} \end{aligned}$$

Can be numerically unstable to implement it this way in the computer, but useful for analysis.

Invertibility

Be careful, the matrix inverse does not always exist. For example, a matrix cannot be inverted if...

- More rows than columns
- More columns than rows
- Rows or columns can be written as linear combinations of other rows or columns ("linearly dependent")

Norms

- A norm is a function that measures how “large” a vector is
- Similar to a distance between zero and the point represented by the vector

$$f(\mathbf{x}) = 0 \implies \mathbf{x} = 0$$
$$f(\mathbf{x} + \boldsymbol{y}) \leq f(\mathbf{x}) + f(\boldsymbol{y}) \text{ (the triangle inequality)}$$
$$\forall a \in \mathbb{R},\ f(a\mathbf{x}) = |a| f(\mathbf{x})$$

Norms

- L^p norm

$$||\mathbf{x}||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

- Most popular norm: L2 norm, $p = 2$, i.e., the Euclidean norm.
- L1 norm:

$$||\mathbf{x}||_1 = \sum_i |x_i|$$

- Max norm, infinite norm:

$$||\mathbf{x}||_\infty = \max_i |x_i|$$

Special Matrices and Vectors

- Unit vector:

$$||\mathbf{x}||_2 = 1$$

- Symmetric matrix:

$$A = A^\top$$

- Orthogonal matrix

$$A^\top A = AA^\top = I_d$$
$$A^\top = A^{-1}$$

Trace

- The trace of an $n \times n$ matrix is the sum of the diagonal

$$\mathrm{Tr}(A) = \sum_i A_{i,i}$$

- It satisfies some nice commutative properties

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$$

In particular, for any vectors $v_1, v_2 \in \mathbb{R}^d$,

$$v_1^\top v_2 = \mathrm{Tr}\left(v_1^\top v_2\right) = \mathrm{Tr}\left(v_1 v_2^\top\right).$$

How to learn linear algebra

- Lots of practice problems.
- Start writing out things explicitly with summations and individual indexes.
- Eventually you will be able to mostly use matrix and vector product notation quickly and easily.

What is random and what is not random?

- In Probability & Statistics, we use capitalized letters for generic random variables (e.g. X and Y).
- The parameters such as $\beta_1, \ldots, \beta_p$ or the function $f : \mathcal{X} \to \mathcal{Y}$ are treated as deterministic (non-random). Of course, being Bayesian is an exception.
- The data points (x_i, y_i) for $1 \leq i \leq n$ are actual values, observed in practice. They can be thought as the realizations of random variables (X_i, Y_i) for $1 \leq i \leq n$.
- When we talk about estimators (e.g. the OLS estimator) which, by definition, are functions of (X_i, Y_i), hence are random.
- Nevertheless, we will NOT distinguish between (x_i, y_i) and (X_i, Y_i) throughout the term, but you should have in mind that the training data (x_i, y_i) are random realizations.

Review of probability facts

Let X and Y be two random variables.

- $$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- More generally, for any function f,

  $$\text{Var}(f(X)) = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] = \mathbb{E}[(f(X))^2] - (\mathbb{E}[f(X)])^2.$$

- X is said to be uncorrelated with Y if

$$\mathrm{Cov}(X, Y) = 0.$$

In particular, the fact that X is independent of Y implies that $\mathrm{Cov}(X, Y) = 0$.

- For any constants a, b,

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y).$$

In particular, if X is uncorrelated with Y, then

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y).$$

- For any function f and g, if X is independent of Y, then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)],$$

  and

$$\mathbb{E}[f(X) \mid Y] = \mathbb{E}[f(X)].$$

- For any function h,

$$\begin{aligned}\mathbb{E}[h(X, Y)] &= \mathbb{E}_X\left[\mathbb{E}_{Y|X}[h(X, Y) \mid X]\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[h(X, Y) \mid Y]\right]\end{aligned}$$

  where $\mathbb{E}_X$ is the expectation w.r.t. the randomness of X whereas $\mathbb{E}_{Y|X}$ is w.r.t. the randomness of $Y \mid X$.