Linear Algebra (https://probml.github.io/pml-book/book1.html)

Vector

- A vector is a list of numbers
- $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$
- $x^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$

Inner product

- given two vectors: $x, y \in \mathbb{R}^n$
- notations: $\langle x, y \rangle$ or $x^\top y$
- $x^\top y = y^\top x = \sum_{i=1}^n x_i y_i$ (a scalar)

Euclidean norm (2-norm)

- the length of the vector
- given a vectors: $x \in \mathbb{R}^n$
- $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$
- $\|x\|^2 = x^\top x$
- unit vector: $\frac{x}{\|x\|}$

Orthogonal and orthonormal

- given three vectors: $x, y, z \in \mathbb{R}^n$
- if $\langle x, y \rangle = 0$, then x and y is **orthogonal** (perpendicular) to each other
- the set of vectors $\{x, y, z\}$ is said to be orthonormal if x, y, z are mutually orthogonal, and $\|x\| = \|y\| = \|z\| = 1$

Linear independence

- given a set of m vectors $\{v_1, v_2, \ldots, v_m\}$
- the set is **linearly dependent** if a vector in the set can be represented as a linear combination of the remaining vectors
  - $v_m = \sum_{i=1}^{m-1} \alpha_i v_i,\ \alpha_i \in \mathbb{R}$
- otherwise, the set is **linearly independent**

Span

- the **span** of $\{v_1, \dots, v_m\}$ is the set of all vectors that can be expressed as a linear combination of $\{v_1, \dots, v_m\}$
  - span$(\{v_1, \dots, v_m\})$ = $\left\{u : u = \sum_{i=1}^m \alpha_i v_i\right\}$
- if $\{v_1, \dots, v_m\}$ is linearly independent, where $v_i \in \mathbb{R}^m$, then any vector $u \in \mathbb{R}^m$ can be written as a linear combination of v_1 through v_m
  - in other words, span$(\{v_1, \dots, v_m\}) = \mathbb{R}^m$
  - e.g. $v_1 = \begin{bmatrix}1\\0\end{bmatrix}$ $v_2 = \begin{bmatrix}0\\1\end{bmatrix}$, span$\left(\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\}\right) = \mathbb{R}^2$
  - we call $\{v_1, \dots, v_m\}$ a **basis** of $\mathbb{R}^m$

Matrix

- A matrix $A \in \mathbb{R}^{m \times n}$ is a 2d array of numbers with m rows and n columns
  - if $m = n$, we call A a square matrix
- Matrix multiplication
  - given two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$
  - $C = AB \in \mathbb{R}^{m \times p}$, where $C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$
  - not commutative: $AB \neq BA$
- Matrix-vector multiplication
  - can be viewed as a linear combination of the columns of a matrix

Transpose $A^\top$

- The transpose of a matrix results from flipping the rows and columns
  - $(A^\top)_{ij} = A_{ji}$
- Some properties:
  - $(AB)^\top = B^\top A^\top$
  - $(A + B)^\top = A^\top + B^\top$
- If $A^\top = A$, we call A a **symmetric** matrix

Trace $\operatorname{tr}(A)$

- The trace of a square matrix is the sum of its diagonal
  - $\operatorname{tr}(A) = \sum_{i=1}^n A_{ii}$
- If A, B are square matrix, then $\operatorname{tr}(AB) = \operatorname{tr}(BA)$

Inverse A^{-1}

- The inverse of a square matrix A is the unique matrix such that
  - $A^{-1}A = I = AA^{-1}$

- Let A be a $n \times n$ matrix. A is invertible iff
  - the column vectors of A is linearly independent (spans $\mathbb{R}^n$)
  - $\det(A) \neq 0$
- Assume that both A, B are invertible. Some properties:
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^\top = (A^\top)^{-1}$

Orthogonal matrix

- A square matrix is orthogonal if all its columns are **orthonormal**
- U is a orthogonal matrix iff
  - $U^TU = I = UU^\top$
  - $U^{-1} = U^\top$
- Note: if U is not square but has orthonormal vectors, we still have $U^TU = I$ but not $UU^\top = I$

Calculus

Chain rule

- $\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$ (single-variable)
- e.g., $y = (x^2+1)^3$, $\frac{dy}{dx} =?$
  - let $u = x^2 + 1$, then $y = u^3$
  - $\frac{dy}{dx} = 3(x^2+1)^2(2x)$

Critical points

- $f' > 0$ => f is increasing
- $f' < 0$ => f is decreasing
- $f'' > 0$ => f' is increasing => f is concave up
- $f'' < 0$ => f' is decreasing => f is concave down
- We call $x = a$ a critical point of a continous function $f(x)$ if either $f'(a) = 0$ or $f'(a)$ is undefined
- different kinds of critical points: local minimum, local maximum, inflection point
  - use second derivative test to check the concavity

Taylor series

- $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$ (single-variable)

- second order approximation:
  - $f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$ (single-variable)
  - $f(w) \approx f(u) + \nabla f(u)^\top (w-u) + \frac{1}{2}(w-u)^\top H_f(u)(w-u)$ (multivariate)
    - $w, u \in \mathbb{R}^d$
    - $\nabla {f_w}^\top = [\frac{\partial f}{\partial w_0}, \cdots, \frac{\partial f}{\partial w_d}]$
    - H_f is the Hessian matrix, where $(H_f)_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j}$

Probability

Conditional probability

- $P(A|B)$: the conditional probability of event A happening given that event B has occurred
  - $P(A|B) = \frac{P(A,B)}{P(B)}$
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (**Bayes rule**)
- $\{A_i\}_{i=1}^n$ is a partition of the sample space. We have $P(B) = \sum_{i=1}^n P(B, A_i) = \sum_{i=1}^n P(B|A_i)P(A_i)$

Independece and conditional independence

- random variable X is independent with Y iff
  - $P(X|Y) = P(X)$
  - $P(X,Y) = P(X)P(Y)$
- random variable X is conditionally independent with Y given Z iff
  - $P(X|Y,Z) = P(X|Z)$
  - $P(X,Y|Z) = P(X|Z)P(Y|Z)$

Expectation

- $\mathcal{X}$: the **sample space**. The set of all possible outcomes of an experiment.
  - e.g., the face of a dice, $\mathcal{X} = \{1,2,3,4,5,6\}$
- $p(x)$: probability mass function (discrete) or probability density function (continuous)
- $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x)$ (discrete)
- $\mathbb{E}[X] = \int_{\mathcal{X}} x\, p(x)\, dx$ (continuous)
- $\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$
- If X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$
  - the converse is *not* true

Variance and covariance

- $\mathrm{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right]$, where $\mu = \mathbb{E}[X]$
    - $= \mathbb{E}\left[X^2\right] - \mu^2$
- $\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$
- $\mathrm{Cov}[X,Y] = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$
- If X and Y are independent, then $\mathrm{Cov}[X,Y]=0$
    - the converse is *not* true

Covariance matrix

- Consider d random variables $X_1,\dots,X_d$
- $\mathrm{Cov}[X_i,X_j] = \mathbb{E}\left[(X_i-\mu_{X_i})(X_j-\mu_{X_j})\right]$
- We can write a $d\times d$ matrix Σ, where $\Sigma_{ij} = \mathrm{Cov}[X_i,X_j]$
    - when $i=j$ (diagonal), $\mathbb{E}\left[(X_i-\mu_{X_i})(X_i-\mu_{X_i})\right] = \mathrm{Var}[X_i]$