

Linear Vector Spaces:

Definition: A linear vector space, X , is a set of elements (vectors) defined over a scalar field, F , that satisfies the following conditions:

- 1) If $x \in X$ and $y \in X$ then $x+y \in X$.
- 2) $x+y = y+x$
- 3) $(x+y)+z = x+(y+z)$
- 4) There is a unique vector $0 \in X$, such that $x+0=x$ for all $x \in X$.
- 5) For each vector $x \in X$ there is a unique vector in X , to be called $(-x)$, such that $x+(-x)=0$.
- 6) multiplication, for all scalars $a \in F$, and all vectors $x \in X$, $7) \text{ For any } x \in X, 1x=x \text{ (for scalar 1).}$
- 8) For any two scalars $a \in F$ and $b \in F$ and any $x \in X$, $a(bx) = (ab)x$.
- 9) $(a+b)x = ax + bx$.
- 10) $a(x+y) = ax + ay$.

Linear Independence: Consider n vectors $\{x_1, x_2, \dots, x_n\}$. If there exists n scalars a_1, a_2, \dots, a_n , at least one of which is nonzero, such that $a_1x_1 + a_2x_2 + \dots + a_nx_n = 0$, then the $\{x_i\}$ are linearly dependent.

Spanning a Space:

Let X be a linear vector space and let $\{u_1, u_2, \dots, u_n\}$ be a subset of vectors in X . This subset spans X if and only if for every vector $x \in X$ there exist scalars x_1, x_2, \dots, x_n such that $x = x_1u_1 + x_2u_2 + \dots + x_nu_n$.

Inner Product: (x, y) for any scalar function of x and y .

1. $(x, y) = (y, x)$
2. $(x, ay_1 + by_2) = a(x, y_1) + b(x, y_2)$
3. $(x, x) \geq 0$, where equality holds iff x is the zero vector.

Norm: A scalar function $\|x\|$ is called a norm if it satisfies:

1. $\|x\| \geq 0$
2. $\|x\| = 0$ if and only if $x = 0$.
3. $\|ax\| = |a|\|x\|$
4. $\|x + y\| \leq \|x\| + \|y\|$

Angle: The angle θ bet. 2 vectors x and y is defined by $\cos \theta = \frac{(x, y)}{\|x\| \|y\|}$

Orthogonality: 2 vectors $x, y \in X$ are said to be orthogonal if $(x, y) = 0$.

Gram Schmidt Orthogonalization:

Assume that we have n independent vectors y_1, y_2, \dots, y_n . From these vectors we will obtain n orthogonal vectors v_1, v_2, \dots, v_n .

$$v_1 = y_1, \quad v_k = y_k - \sum_{i=1}^{k-1} \frac{(y_k, v_i)}{(v_i, v_i)} v_i,$$

where $\frac{(v_i, y_k)}{(v_i, v_i)} v_i$ is the projection of y_k on v_i

Vector Expansions:

$$x = \sum_{i=1}^n x_i v_i = x_1 v_1 + x_2 v_2 + \dots + x_n v_n,$$

for orthogonal vectors, $x_j = \frac{(v_j, x)}{(v_j, v_j)}$

Reciprocal Basis Vectors:

$$(r_i, v_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}, \quad x_j = (r_j, x)$$

To compute the reciprocal basis vectors: set $B = [v_1 \ v_2 \ \dots \ v_n]$,

$R = [r_1 \ r_2 \ \dots \ r_n]$, $R^T = B^{-1}$ In matrix form: $x^v = B^{-1} x^s$

Transformations:

A transformation consists of three parts:

domain: $X = \{x_i\}$, range: $Y = \{y_i\}$, and a rule relating each $x_i \in X$ to an element $y_i \in Y$.

Linear Transformations: transformation A is linear if:

1. for all $x_1, x_2 \in X$, $A(x_1 + x_2) = A(x_1) + A(x_2)$
2. for all $x \in X, a \in R$, $A(ax) = aA(x)$

Matrix Representations:

Let $\{v_1, v_2, \dots, v_n\}$ be a basis for vector space X , and let $\{u_1, u_2, \dots, u_n\}$ be a basis for vector space Y . Let A be a linear transformation with domain X and range Y : $A(x) = y$

The coefficients of the matrix representation are obtained from

$$A(v_j) = \sum_{i=1}^m a_{ij} u_i$$

Change of Basis: $B_t = [t_1 \ t_2 \ \dots \ t_n]$, $B_w = [w_1 \ w_2 \ \dots \ w_n]$
 $A' = [B_w^{-1} A B_t]$

Eigenvalues & Eigenvectors: $Az = \lambda z$, $|[A - \lambda I]| = 0$

Diagonalization: $B = [z_1 \ z_2 \ \dots \ z_n]$,

where $\{z_1, z_2, \dots, z_n\}$ are the eigenvectors of a square matrix A ,
 $[B^{-1} A B] = \text{diag}([\lambda_1 \ \lambda_2 \ \dots \ \lambda_n])$

Perceptron Architecture:

$$a = \text{hardlim}(Wp + b), \quad W = [w_1^T \ w_2^T \ \dots \ w_s^T]^T, \\ a_i = \text{hardlim}(n_i) = \text{hardlim}(w_i^T p + b_i)$$

Decision Boundary: $w_i^T p + b_i = 0$

The decision boundary is always orthogonal to the weight vector. Single-layer perceptrons can only classify linearly separable vectors.

Perceptron Learning Rule

$$W^{new} = W^{old} + ep^T, \quad b^{new} = b^{old} + e, \\ \text{where } e = t - a$$

Hebb's Postulate: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."

Linear Associator: $a = \text{purelin}(Wp)$

The Hebb Rule: Supervised Form: $w_{ij}^{new} = w_{ij}^{old} + t_{qi} p_{qi}$

$$W = t_1 p_1^T + t_2 p_2^T + \dots + t_Q p_Q^T \\ W = [t_1 \ t_2 \ \dots \ t_Q] \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_Q^T \end{bmatrix} = TP^T$$

Pseudoinverse Rule: $W = TP^+$

When the number, R , of rows of P is greater than the number of columns, Q , of P and the columns of P are independent, then the pseudoinverse can be computed by $P^+ = (P^T P)^{-1} P^T$

Variations of Hebbian Learning:

Filtered Learning (Ch.14): $W^{new} = (1 - \gamma)W^{old} + \alpha t_q p_q^T$

Delta Rule (Ch.10): $W^{new} = W^{old} + \alpha(t_q - a_q)p_q^T$

Unsupervised Hebb (Ch.13): $W^{new} = W^{old} + \alpha a_q p_q^T$

Taylor: $F(x) = F(x^*) + \nabla F(x^T)|_{x=x^*} (x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 F(x^T)|_{x=x^*} (x - x^*) + \dots$

Grad $\nabla F(x) = \left[\frac{\partial}{\partial x_1} F(x) \quad \frac{\partial}{\partial x_2} F(x) \quad \dots \quad \frac{\partial}{\partial x_n} F(x) \right]^T$

Hessian: $\nabla^2 F(x) =$

$$\begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(x) & \frac{\partial^2}{\partial x_1 \partial x_2} F(x) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} F(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(x) & \frac{\partial^2}{\partial x_2^2} F(x) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} F(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} F(x) & \frac{\partial^2}{\partial x_n \partial x_2} F(x) & \dots & \frac{\partial^2}{\partial x_n^2} F(x) \end{bmatrix}$$

Directional Derivatives:

1st Dir.Der.: $\frac{p^T \nabla F(x)}{\|p\|}$, **2nd Dir.Der.:** $\frac{p^T \nabla^2 F(x) p}{\|p\|^2}$

Minima:

Strong Minimum: if a scalar $\delta > 0$ exists, such that $F(x) < F(x + \Delta x)$ for all Δx such that $\delta > \|\Delta x\| > 0$.

Global Minimum: if $F(x) < F(x + \Delta x)$ for all $\Delta x \neq 0$

Weak Minimum: if it is not a strong minimum, and a scalar $\delta > 0$ exists, such that $F(x) \leq F(x + \Delta x)$ for all Δx such that $\delta > \|\Delta x\| > 0$.

Necessary Conditions for Optimality:

1st-Order Condition: $\nabla F(x)|_{x=x^*} = 0$ (Stationary Points)

2nd-Order Condition: $\nabla^2 F(x)|_{x=x^*} \geq 0$ (Positive Semi-definite Hessian Matrix).

Quadratic fn.: $F(x) = \frac{1}{2} x^T A x + d^T x + c$

$$\nabla F(x) = Ax + d, \quad \nabla^2 F(x) = A, \quad \lambda_{\min} \leq \frac{p^T A p}{\|p\|^2} \leq \lambda_{\max}$$