Chapter 2

Preliminaries

Contents

This chapter provides some basic definitions and theorems.

2.1 Linear Algebra

Definition 2.1.1. (Inner Product) [118] Let V be a vector space over the field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. An inner product on a vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}$, satisfying the following axioms.

For $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in V$ and $\alpha \in \mathbb{K}$,

- $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \langle \mathbf{u}_2, \mathbf{u}_1 \rangle$
- $\langle \mathbf{u}_1 + \mathbf{u}_2, \mathbf{u}_3 \rangle = \langle \mathbf{u}_1, \mathbf{u}_3 \rangle + \langle \mathbf{u}_2, \mathbf{u}_3 \rangle$
- $\langle \alpha \mathbf{u}_1, \mathbf{u}_2 \rangle = \alpha \langle \mathbf{u}_1, \mathbf{u}_2 \rangle$
- $\langle \mathbf{u}_1, \mathbf{u}_1 \rangle \geq \mathbf{0}$ and $\langle \mathbf{u}_1, \mathbf{u}_1 \rangle = \mathbf{0} \Leftrightarrow \mathbf{u}_1 = \mathbf{0}$.

Definition 2.1.2. (Inner Product Space) [118]
Let V be a vector space over the field $\mathbb{K}$. The vector space V with an inner product $\langle \cdot, \cdot \rangle$ is called an inner product space.

Definition 2.1.3. (Norm)[71]
Let V be a vector space over the field $\mathbb{K}$. Then the norm of a vector $\mathbf{u}$ in V is a real number denoted by $\|\mathbf{u}\|$ and defined as

$$\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}}.$$

A norm satisfying the following properties:

1. $\|\mathbf{u}\| = \mathbf{0}$ if and only if $\mathbf{u} = \mathbf{0}$
2. $\|\lambda \mathbf{u}\| = |\lambda| \|\mathbf{u}\|, \quad \forall \lambda \in \mathbb{K}, \forall \mathbf{u} \in V$
3. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in V$

The norm $\|\cdot\|$ induces a metric. i.e. distance on V is defined as:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|. \tag{2.1.1}$$

Definition 2.1.4. (Dot Product) [118]
Let $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$ and $g(\cdot) = \sum_{j=1}^{n} \beta_j k\left(\cdot, x'_j\right)$ be $\mathbb{K}$- valued functions. Then the dot product between f and g is defined as $\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \, \beta_j k\left(x_i, x'_j\right)$.
The dot product $\langle f, g \rangle$ is symmetric and positive definite.

Definition 2.1.5. (Quadratic Form) [92]
For an $n \times n$ real symmetric matrix $\mathbf{A}$ and an n-dimensional vector $\mathbf{x}$, the form

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mathbf{x}_i\mathbf{x}_j, \quad a_{ij} = a_{ji}$$

is called a real quadratic form or quadratic form, denoted by $Q(x)$.
For an Hermitian matrix $\mathbf{A}$ and a complex n-dimensional vector $\mathbf{x}$, the form

$$\mathbf{x}^*\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\bar{\mathbf{x}}_i\mathbf{x}_j, \quad a_{ij} = \bar{a_{ji}}$$

is called a complex quadratic form.

Definition 2.1.6. (Definite and Semidefinite Quadratic Form) [92]
A Definite Quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a real symmetric matrix (or $\mathbf{x}^*\mathbf{A}\mathbf{x}$ where $\mathbf{A}$is a Hermitian matrix), is said to be

- positive definite if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} > 0), \quad \text{for} \quad \mathbf{x} \neq \mathbf{0} \tag{2.1.2}$$
$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} = 0), \quad \text{for} \quad \mathbf{x} = \mathbf{0} \tag{2.1.3}$$

- positive semidefinite if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0, \quad (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} \geq 0), \quad \text{for} \quad \mathbf{x} \neq \mathbf{0} \tag{2.1.4}$$
$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 0, \quad (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} = 0), \quad \text{for} \quad \mathbf{x} = \mathbf{0} \tag{2.1.5}$$

- negative definite if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} < 0, \quad (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} < 0), \quad \text{for} \quad \mathbf{x} \neq \mathbf{0} \tag{2.1.6}$$
$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 0, \quad (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{x} = 0), \quad \text{for} \quad \mathbf{x} = \mathbf{0} \tag{2.1.7}$$

- negative semidefinite if

$$\mathbf{x}^T\mathbf{A}\mathbf{y} \leq 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{y} \leq 0), \quad \text{for} \quad \mathbf{x} \neq \mathbf{0} \tag{2.1.8}$$
$$\mathbf{x}^T\mathbf{A}\mathbf{y} = 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{y} = 0), \quad \text{for} \quad \mathbf{x} = \mathbf{0} \tag{2.1.9}$$

- indefinite if for some $\mathbf{x} \in V$, the quadratic form

$$\mathbf{x}^T\mathbf{A}\mathbf{y} \geq 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{y} \geq 0)$$

and for some $\mathbf{x} \in V$, the quadratic form

$$\mathbf{x}^T\mathbf{A}\mathbf{y} \leq 0, (\text{or} \quad \mathbf{x}^*\mathbf{A}\mathbf{y} \leq 0)$$

Definition 2.1.7. (Non-Degenerate Quadratic Form) [40] A quadratic form is said to be non degenerate quadratic form if $Q(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \mathbf{0}$.

Definition 2.1.8. (Moore Penrose Inverse) [55]
Let A be an $m \times n$ matrix, then the Moore Penrose Inverse of A is the unique $n \times m$ matrix B with the following properties:

- $ABA=A$
- $BAB=B$
- $(AB)^T = AB$
- $(BA)^T = BA$.

Definition 2.1.9. (Gram Matrix) [118]
Let χ be a nonempty subset of $\mathbb{R}^n$. $k : \chi \times \chi \to \mathbb{K}$ where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ be given function and $x_1, x_2, ...x_n \in \chi$. Then the $n \times n$ matrix K with elements $K_{ij} = k(x_i, x_j)$ is called Gram matrix of k with respect to $\{x_1, x_2, ...x_n\}$.

2.2 Optimization

Definition 2.2.1. (Convex Set and Convex Function) [107]
The set T = $\{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_m\}$ in a vector space over $\mathbb{R}^n$ is said to be a convex if the line segment joining any two points of T lies entirely in T. i.e. for $\mathbf{x}_1, \mathbf{x}_2 \in$ T, $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in T$, where $0 \leq \lambda \leq 1$.
If f $\in \mathbb{R}^n$ such that $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2)$, where $0 \leq \lambda \leq 1$, then f is said to be a convex function.

Definition 2.2.2. (Convex Hull) [132]
Convex hull C of the set of points T = $\{x_1, x_2, ...x_m\}$ is the intersection of all convex sets containing T and is given by $C \equiv a_1x_1+a_2x_2+...+a_mx_m$, where $a_i \geq 0$, i=1,2,...,m and $\sum_{i=1}^{m} a_i = 1$

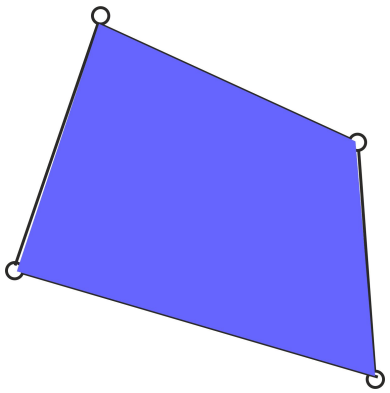

Figure 2.1: Convex Hull

Definition 2.2.3. (Reduced/Soft Convex Hull) [47]
Reduced convex hull of the sample points $\{x_1, x_2, ...x_m\}$ is convex combination $a_1x_1+a_2x_2+...+a_mx_m$ such that $\sum_{i=1}^{m} a_i = 1$ and $0 \leq a_i \leq \mu$, where $\frac{1}{m} \leq \mu \leq 1$.
If $\mu = 1$ then it is usual convex hull.

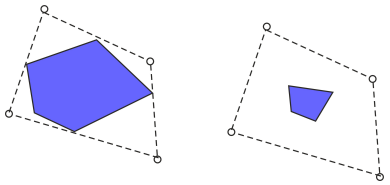

Figure 2.2: Reduced Convex Hull

Definition 2.2.4. (Hessian Matrix) [107] Square matrix of second order partial derivatives of a scaler valued function f is called Hessian Matrix of f which is defined as:

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Precursory 2.2.1. Karush-Kuhn-Tucker(KKT) conditions [41]
Consider optimization problem:

$$\min_{\{x\in\mathbb{R}^n\}} f(x) \tag{2.2.1}$$

subject to constraints,

$$\begin{aligned} g_i(x) &\leq 0, i = 1, 2, ..., m \\ h_j(x) &= 0, j = 1, 2, ..., r. \end{aligned}$$

Define the generalized Lagrangian,

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{r} \beta_j h_j(x)$$

where α and β are Lagrange's multiplier.
The Karush-Kuhn-Tucker (KKT) conditions are as follows:

$$\begin{aligned} \frac{\partial}{\partial x_i} L(x^*, \alpha^*, \beta^*) &= 0, \quad i = 1, 2, ..., n \quad (\texttt{stationarity}) \\ \frac{\partial}{\partial \beta_i} L(x^*, \alpha^*, \beta^*) &= 0, \quad i = 1, 2, ..., r \quad (\texttt{stationarity}) \\ \alpha_i^* g_i(x_i^*) &= 0, \quad i = 1, 2, ..., m \quad (\texttt{Complementary slackeness}) \\ g_i(x^*) &\leq 0, \quad i = 1, 2, ..., m \quad (\texttt{primal feasibility}) \\ \alpha^* &\geq 0, \quad i = 1, 2, ..., m. \quad (\texttt{dual feasibility}) \end{aligned}$$

where, x^*, α^* and β^* are the values of x, α, β satisfying the KKT conditions. For convex optimization problem KKT conditions are necessary and sufficient for global minimum.

Precursory 2.2.2. Grid Search Method [107] This method involves setting up grid in the design space. Let the lower and upper bound of the design variable x_i be l_i and u_i respectively. Divide the range (l_i, u_i) into p_{i-1} equal parts, so that $x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p_i)}$ are grid points along x_i-axis, $i = 1, 2, ...n$. Therefore, $p_1 p_2 ... p_n$ grid points are there in the design space. In this method objective function is evaluated at each of these grid points.

When number of design variables is small, this method can be conveniently used to find an approximate optimum value.

2.3 Functional Analysis

Definition 2.3.1. (Cauchy Sequence) [77]
A sequence (x_n) of a metric space X is said to be Cauchy if for every $\epsilon > 0$, there is some N_0, such that $d(x_n, x_m) < \epsilon$ for all $n, m \geq N_0$.

Definition 2.3.2. (Completeness, Hilbert Space) [118]
Let $\mathbf{X}$ be an inner product space. If every cauchy sequence in $\mathbf{X}$ is convergent then $\mathbf{X}$ is called a complete inner product space. A complete inner product space is called a Hilbert space.

Definition 2.3.3. (Kernel Function) [118]
Let χ be a nonempty subset of $\mathbb{R}^n$. A function $k : \chi \times \chi \to \mathbb{R}$, such that $k(x, x')$ returns a real number giving the similarity between two patterns x and x' is called a kernel function.

Theorem 2.3.1. *(Mercer's Theorem) [91]*
Let χ be a closed subset of $\mathbb{R}^n, n \in \mathbb{N}$. Let $k : \chi \times \chi \to \mathbb{R}$ be a symmetric function, i.e. $k(x, x') = k(x', x)$ where $x \in \mathbb{R}^n$ then k to be a valid kernel called Mercer's kernel, it is necessary and sufficient that for any finite set of points $\{x_1, x_2, ..., x_m\}$ and real numbers $\{a_1, a_2, ..., a_m\}$, $\sum_{i,j}^{m} a_i a_j k(x_i, x_j) \geq 0$, i.e. the corresponding kernel matrix K is symmetric positive semi definite (definition 2.1.6).

Definition 2.3.4. (Reproducing Kernel) [118]
Let k be the real valued positive definite kernel and χ be a non empty subset of $\mathbb{R}^n$. Define the non linear function $\phi : \chi \to \mathbf{R}^{\chi}$ such that it maps x to $k(\cdot, x)$:

$$\phi : x \to k(\cdot, x)$$

and $\mathbf{R}^{\chi}$ be the space of functions from χ to $\mathbb{R}$, viz.

$$\mathbf{R}^{\chi} := \{\phi : \chi \to \mathbb{R}\} \in \mathbb{R}^{\chi}. \tag{2.3.1}$$

Construct a vector space containing the images of input patterns under the mapping ϕ, as

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \tag{2.3.2}$$

where $m \in \mathbb{N}, \alpha_i \in \mathbb{R}$ and $x_i, x_2, ..., x_m \in \chi$ are arbitrary.
The vector space is:

$$\text{Span}\left(\{\phi(x) : x \in \chi\}\right) = \left\{ f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) : m \in \mathbb{N}, x_i \in \chi, \alpha_i \in \mathbb{R} \right\}$$

Let two functions $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$ and $g(\cdot) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$, and define the inner product,

$$\langle f, g \rangle_{Hk} = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \tag{2.3.3}$$

It is a valid inner product, as

- It is symmetric, since k is symmetric:

$$\langle g, f \rangle_{Hk} = \sum_{j=1}^{m'} \sum_{i=1}^{m} \beta_j \alpha_i k(x'_j, x_i) = \langle f, g \rangle_{Hk}$$

- It is bilinear:

$$\begin{aligned} \langle f, g \rangle_{Hk} &= \sum_{i=1}^{m} \beta_j \sum_{j=1}^{m'} \alpha_i k(x_i, x'_j) \\ &= \sum_{j=1}^{m'} \beta_j f(x'_j). \end{aligned}$$

  Therefore,

$$\begin{aligned} \langle f_1 + f_2, g \rangle &= \sum_{j=1}^{m'} \beta_j (f_1(x'_j) + f_1(x'_j)) \\ &= \sum_{j=1}^{m'} \beta_j f_1(x'_j) + \sum_{j=1}^{m'} \beta_j f_2(x'_j) \\ &= \sum_{j=1}^{m'} \beta_j f_1(x'_j) + \sum_{j=1}^{m'} \beta_j f_2(x'_j) \\ &= \langle f_1, g \rangle_{Hk} + \langle f_2, g \rangle_{Hk} \end{aligned}$$

Similarly, we can show that

$$\langle f, (g_1 + g_2) = \langle f, g_1 \rangle_{Hk} + \langle f, g_2 \rangle_{Hk} \rangle$$

Then for all functions (2.3.2), we have $\langle k(., x), f \rangle = f(x)$ where k is positive definite kernel and is called the reproducing kernel.

Definition 2.3.5. (Reproducing Kernel Hilbert Space(RKHS)) [118]
Let χ be a nonempty set. $\mathbf{R}^{\chi}$ be a Hilbert space of functions $f : \chi \to \mathbb{R}$ define as (2.3.1), endowed with the dot product $\langle \cdot, \cdot \rangle$ (definition 2.1.4) and the norm $\|f\| = \sqrt{\langle f, f \rangle}$ and let $k : \chi \times \chi \to \mathbb{R}$ be a reproducing kernel, i.e. $\langle f, k(x, .) \rangle = f(x)$ for all $f \in \mathcal{H}$ and k span $\mathbf{R}^{\chi}$, then $\mathbf{R}^{\chi}$ is called Reproducing Kernel Hilbert Space.

2.4 Machine Learning

Precursory 2.4.1. One-to-One Algorithm

- Let there be K classes in a multiclassification problem.
- Divide the K classes in to $\frac{K(K-1)}{2}$ binary classes.
- Train each of these binary classes and set parameter values.
- Test the unknown data sample to all these binary classifiers.
- The class that gets the highest number of positive prediction, is predicted as the class of unseen data sample.

Precursory 2.4.2. Confusion Matrix
Confusion matrix gives error of classifier model and also types of the error.

Table 2.1: Confusion Matrix for Binary Class

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

where,

- TP : number of correctly classified samples from positive class.
- TN : number of correctly classified samples from negative class.
- FP : number of wrong classified samples from positive class.
- FN : number of wrong classified samples from negative class.

Precursory 2.4.3. (Accuracy)
Accuracy is an important evaluation for evaluating any classifier. The accuracy of a classifier is also called predicted positive condition rate. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precursory 2.4.4. (Precision) [30]
Precision or true positive rate measures that, among all positive predicted samples how many samples are actually positive. i.e. proportion of positive that are correctly identified and define as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precursory 2.4.5. (Recall)[30]
Recall measures that, among all the samples which are actually positive, what fraction are detected as positive and it is define as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

For imbalanced dataset if classification accuracy is measured according to definition (2.4.3), then the classifier can predict the value of the majority class for all predictions and achieve high classification accuracy which is not correct. This drawback can be overcome by measuring the classification accuracy using F-Score.

Precursory 2.4.6. (F-Score) [105]
F-score is the balance between Precision and Recall, it is the harmonic mean of Recall and Precision and is defined as:

$$\text{F-Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precsion} + \text{Recall}} \tag{2.4.1}$$

Precursory 2.4.7. (G-Score) [105]
G-score is another measure of accuracy which also do not account the size of positive and negative classes and provide a fair comparison. It is geometric mean of Precision and Recall and is defined as:

$$\text{G-Score} = \sqrt{\text{Precision} \cdot \text{Recall}} \tag{2.4.2}$$

Precursory 2.4.8. (K-fold Cross Validation)
K-fold validation is a technique to validate model during training phase. Following are the steps for the K-fold cross validation technique:

1. Partition the training data set into k equal part, where each partition is called a fold.
2. for $i = 1$ to k
    - Take i^{th} fold as validation set (testing set) and remaining $k - 1$ folds in the Cross validation set (training set).
    - Train classifier using $k - 1$ training set and calculate the accuracy of the model obtained by temporary training-testing set.
    - end
    - Estimate the accuracy of the classifier by averaging the accuracies obtained from all k folds of cross validation.

In the k-fold cross validation technique, all the samples of the original training data set are used for both training as well as for validation. Also, each sample is used for validation just once.

Theorem 2.4.1. *(Universal Approximation Theorem) [27, 26]*
Let $\phi(\cdot)$ be any non constant, bounded and monotonically increasing continuous function. Let I_n, the n-dimensional unit hypercube $[0,1]^n$. Then $\forall f \in C(I_n)$, and $\forall \varepsilon > 0$, $\exists p \in N$,sets of real constants α_j, $\theta_j \in R, w_{ij}$ where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$ such that $G(x) = \sum_{j=1}^{p} \alpha_j \phi(\sum_{i=1}^{n} w_{ij}^T x_i - \theta_j)$ $x \in I_n, w \in \mathbb{R}^{n \text{x} p}$ as an approximation of function $f(\cdot)$ independent of ϕ, that is for $x \in I_m$, $|G(x) - f(x)| < \epsilon$ for all $x \in I_m$