Mathematical Analysis in Machine Learning reading notes

I haven't thought of a title yet

School of Computer Science
WuHan University

Franz Hu

What is the first

This is my first time to write notes all in English.

Contents

1 Preliminaries

1.1 Set and Collections

Definition 1.1.1 *Let $\mathcal{D}$ be a collection of sets. The union of $\mathcal{C}$, denoted by $\bigcup \mathcal{C}$, is the set defined by*

$$\bigcup \mathcal{C} = \{x | x \in S \text{ for some } S \in \mathcal{C}\}.$$

If $\mathcal{C}$ is a non-empty collection, its intersection is the set $\bigcap \mathcal{C}$ given by

$$\bigcap \mathcal{C} = \{x | x \in S \text{ for every } S \in \mathcal{C}\}.$$

Here, the generalized union and generalized intersection of sets simply involve performing the union or intersection operation on each element within the set.

Note that if $\mathcal{C}$ and $\mathcal{D}$ are two collections such that $\mathcal{C} \subseteq \mathcal{D}$, then

$$\bigcup \mathcal{C} \subseteq \bigcup \mathcal{D} \text{ and } \bigcap \mathcal{D} \subseteq \bigcap \mathcal{C}$$

Within the framework of collections of subsets of a given set $\mathcal{S}$, we extend the previous definition by taking $\bigcap \varnothing = S$ for the empty collection of subsets of $\mathcal{S}$.

This is consistent with the fact that $\varnothing \subseteq \mathcal{C}$ implies $\bigcap \mathcal{C} \subseteq \mathcal{S}$.

Definition 1.1.2 *The symmetric difference of sets denoted by $\oplus$ is defined by $U \oplus V = (U - V) \cup (V - U)$ for all sets U, V.*

The symmetric difference operation is easily shown to satisfy symmetry and associativity, and also $U \oplus U = \varnothing$. Next, we will prove the associativity.

Proof 1.1.1

$$\begin{aligned}
LEFT &= ((U - V) \cup (V - U)) \oplus T \\
&= (((U - V) \cup (V - U)) - T) \cup (T - ((U - V) \cup (V - U))) \\
&= ((U\bar{V} \cup V\bar{U}) \cap \bar{T}) \cup (T \cap (\bar{U}\bar{V} \cup UV)) \\
&= \sum_{i=1,2,4,7} m_i
\end{aligned}$$

$$RIGHT = LEFT$$

However, the above method is rather cumbersome. We can instead adopt the characteristic function approach for the proof. Here, we first supplement the relevant concepts of characteristic functions.

Definition 1.1.3 *For any set A, the function*

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A \end{cases}$$

is called the characteristic function of the set A.

The characteristic function has the following properties:

1. The necessary and sufficient condition for $A = X$ is $\chi_A(x) \equiv 1$, and the necessary and sufficient condition for $A = \varnothing$ is $\chi_A(x) \equiv 0$;
2. The necessary and sufficient condition for $A \subset B$ is
$$\chi_A(x) \leqslant \chi_B(x), (\forall x \in X);$$
3. $\chi_{\bar{P}}(x) = 1 - \chi_P(x)$.
4. $\chi_{A\cup B}(x) = \chi_A(x) + \chi_B(x) - \chi_{A\cap B}(x)$.
5. $\chi_{A\cap B}(x) = \chi_A(x) \cdot \chi_B(x)$;
6.
$$\chi_{\bigcup_{\alpha\in\Lambda} A_\alpha}(x) = \max_{\alpha\in\Lambda} \chi_{A_\alpha}(x)$$
$$\chi_{\bigcap_{\alpha\in\Lambda} A_\alpha}(x) = \min_{\alpha\in\Lambda} \chi_{A_\alpha}(x)$$
7. Let $\{A_k\}$ be an arbitrary sequence of sets, then
$$\chi_{\overline{\lim}_{k\to\infty} A_k}(x) = \overline{\lim_{k\to\infty}} \chi_{A_k}(x)$$
$$\chi_{\underline{\lim}_{k\to\infty} A_k}(x) = \underline{\lim_{k\to\infty}} \chi_{A_k}(x);$$
8. The necessary and sufficient condition for $\lim_{k\to\infty} A_k$ to exist is that $\lim_{k\to\infty} \chi_{A_k}(x)$ exists $(\forall x \in X)$, and when the limit exists, we have
$$\chi_{\lim_{k\to\infty} A_k}(x) = \lim_{k\to\infty} \chi_{A_k}(x) \quad (x \in X).$$

$$\overline{\lim_{k\to\infty}} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

The limit superior of a sequence of sets reflects the collection of elements that ”repeatedly appear” in the sequence of sets.

$$\underline{\lim_{k\to\infty}} A_k = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

The limit inferior of a sequence of sets reflects the collection of elements that exhibit ”stable membership” within the sequence of sets.

$$\overline{\lim_{k\to\infty}} \chi_{A_k}(x) = \lim_{n\to\infty} \big(\sup_{k\geqslant n} \chi_{A_k(x)}\big)$$
$$\underline{\lim_{k\to\infty}} \chi_{A_k}(x) = \lim_{n\to\infty} \big(\inf_{k\geqslant n} \chi_{A_k(x)}\big)$$

Next, we will employ the method of characteristic functions to complete the proof of the associativity of the symmetric difference operation.

Proof 1.1.2

$$\begin{aligned}\chi_{A\oplus B}(x) &= \chi_{(A-B)\cup(B-A)}(x)\\ &= \chi_{A-B}(x) + \chi_{B-A}(x)\\ &= \chi_A(x)\cdot\chi_{\bar{B}}(x) + \chi_{\bar{A}}(x)\cdot\chi_B(x)\\ &= \chi_A(x) + \chi_B(x) - 2\chi_{A\cap B}(x)\end{aligned}$$

$$\begin{aligned}\textit{LEFT} &= \chi_{(U\oplus V)\oplus T}(x)\\ &= \chi_U(x)+\chi_V(x)+\chi_T(x)-2\chi_{U\cap V}(x)-2\chi_{V\cap T}(x)-2\chi_{U\cap T}(x)+4\chi_{U\cap V\cap T}(x)\\ &= \textit{RIGHT}\end{aligned}$$

Definition 1.1.4 *Below are several concepts pertaining to sets.*

1. *An ordered pair is a collection of sets $\{\{x,y\},\{x\}\}$. It can be readily verified that x and y are determined uniquely.*
2. *Let $\{\{x,y\},\{x\}\}$ be an ordered pair. Then x is the first component of p and y is the second component of p.*
3. *Let X, Y be two sets. Their product is the set $X \times Y$ that consists of all pairs of the form (x,y) where $x \in X$ and $y \in Y$.*
4. *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of sets such that $\bigcup\mathcal{C} = \bigcup\mathcal{D}$. $\mathcal{D}$ is a **refinement** of $\mathcal{C}$ if, for every $D \in \mathcal{D}$, there exists $C \in \mathcal{C}$ such that $D \subseteq C$.This is denoted by $\mathcal{C} \sqsubseteq \mathcal{D}$.*
5. *A collection of sets $\mathcal{C}$ is **hereditary** if $U \in \mathcal{C}$ and $W \subseteq U$ implies $W \in \mathcal{C}$.*
6. *The set of subsets of S that contain k elements is denoted by $\mathcal{P}_k(S)$.Clearly, for every set S, we have $\mathcal{P}_0(S) = \{\varnothing\}$. The set of all finite subsets of a set S is denoted by $\mathcal{P}_{fin}(S) = \bigcup_{k\in\mathcal{N}}\mathcal{P}_k(S)$.*
7. *Let $\mathcal{C}$ be a collection of sets and let U be a set. The **trace** of the collection $\mathcal{C}$ on the set U is the collection $\mathcal{C}_U = \{U \cap C | C \in \mathcal{C}\}$.*

Definition 1.1.5 *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of sets.*

1. $\mathcal{C} \vee \mathcal{D} = \{C \cup D | C \in \mathcal{C}$ *and* $D \in \mathcal{D}\}$,
2. $\mathcal{C} \wedge \mathcal{D} = \{C \cap D | C \in \mathcal{C}$ *and* $D \in \mathcal{D}\}$,
3. $\mathcal{C} - \mathcal{D} = \{C - D | C \in \mathcal{C}$ *and* $D \in \mathcal{D}\}$.

Attention, unlike $\cup$ and cap, the operations $\vee$ and $\wedge$ between collections of sets are not idempotent. Indeed, we have, for example,

$$\mathcal{D} \vee \mathcal{D} = \{\{y\},\{x,y\},\{u,y,z\},\{u,x,y,z\}\} \neq \mathcal{D}.$$

Definition 1.1.6 *A partition of a non-empty subsets of S that are pairwise disjoint and whose union equals S.*

The members of π are referred to as the blocks of the partition π. The collection of partitions of a set S is denoted by $PART(S)$. A partition is finite if it has a finite number of blocks. The set of finite partitions of S is denoted by $PART_{fin}(S)$.

If $\pi \in PART(S)$ then a subset T of S is π-saturated if it is a union of blocks of π.

1.2 Relations and Functions

Definition 1.2.1 *Let X, Y be two sets. A relation on X, Y is a subset ρ of the set product $X \times Y$. If $X = Y = S$, we refer to ρ as a relation on S. The relation ρ on S is:*

- *reflexive if $(x,x) \in \rho$ for every $x \in S$;*
- *irreflexive if $(x,x) \notin \rho$ for every $x \in S$;*

- *symmetric if* $(x, y) \in \rho$ *implies* $(y, x) \in \rho$ *for all* $x, y \in S$*;*
- *antisymmetric if* $(x, y) \in \rho$ *and* $(y, x) \in \rho$ *imply* $x = y$ *for all* $x, y \in S$*;*
- *transitive if* $(x, y) \in \rho$ *and* $(y, z) \in \rho$ *imply* $(x, z) \in \rho$ *for all* $x, y, z \in S$*.*

A **partial order** on S is a relation ρ that belongs to REFL$(S) \cap$ ANTISYMM$(S) \cap$ TRAN(S), that is, a relation that is reflexive, symmetric and transitive.

In current mathematical practice, we often write $x \rho y$ instead on $(x, y) \in \rho$, where ρ is a relation of S and $x, y \in S$. This alternative way to denote the fact that (x, y) belongs to ρ is known as the infix notation.

Definition 1.2.2 *Let* X*,* Y *be two sets. A function(or a mapping) from* X *to* Y *is a relation* f *on* X*,* Y *such that* $(x, y), (x, y') \in f$ *implies* $y = y'$*.*

Let X*,* Y *be two sets and let* $f : X \to Y$*.*

The domain of f *is the set*

$$Dom(f) = \{x \in X | y = f(x) \text{ for some } y \in Y\}.$$

The range of f *is the set*

$$Ran(f) = \{y \in Y | y = f(x) \text{ for some } x \in X\}.$$

Definition 1.2.3 *Let* X *be a set,* $Y = \{0, 1\}$ *and let* L *be a subset of* S*. The characteristic function is discussed above. The indicator function of* L *is the function* $I_L : S \to \mathcal{R} \cup \infty$ *defined by*

$$I_L(x) = \begin{cases} 1 & \text{if } x \in L. \\ \infty & \text{otherwise} \end{cases}$$

for $x \in S$*.*

Definition 1.2.4 *A function* $f : X \to Y$ *is:*

1. ***injective or one-to-one*** *if* $f(x_1) = f(x_2)$ *implies* $x_1 = x_2$ *for* $x_1, x_2 \in Dom(f)$*;*
2. ***surjective or onto*** *if* $Ran(f) = Y$*;*
3. ***total*** *if* $\boldsymbol{Dom}(f) = X$*.*

If f *both injective and surjective, then it is a* ***bijective*** *function.*

Theorem 1.2.1 *A function* $f : X \to Y$ *is injective if and only if there exists a function* $g : Y \to X$ *such that* $g(f(x)) = x$ *for every* $x \in \boldsymbol{Dom}(f)$*.*

A function $f : X \to Y$ *is surjective if and only if there exists a function* $h : X \to Y$ *such that* $f(h(y)) = y$ *for every* $y \in Y$*.*

Here we provide the proof for the latter theorem.

Proof 1.2.1 *Suppose that* f *is a surjective function. The collection* $\{f^{-1}(y) | y \in Y\}$ *indexed by* Y *consists of non-empty sets. By the Axiom of Choice there exists a choice function for this collection, that is a function* $h : Y \to \bigcup_{y \in Y} f^{-1}(y)$ *such that* $h(y) \in f^{-1}(y)$*, or* $f(h(y)) = y$ *for* $y \in Y$*.*

Conversely, suppose that there exists a function $h : X \to Y$ *such that* $f(h(y)) = y$ *for every* $y \in Y$*. Then,* $f(x) = y$ *for* $y = h(y)$*, which shows that* f *is surjective.*

The Axiom of Choice: Let $\mathcal{C} = \{C_i | i \in I\}$ be a collection of non-empty sets indexed by a set I. There exists a function $\phi : I \to \bigcup \mathcal{C}$(known as a choice function) such that $\phi(i) \in C_i$ for each $i \in I$.

Theorem 1.2.2 *There is a bijection $\Psi : \mathcal{P}(S) \to (S \to \{0,1\})$ between the set of subsets of S and the set of characteristic functions defined on S.*

Definition 1.2.5 *A set S is indexed be a set I if there exists a surjection $f : I \to S$. In this case we refer to I as an index set.*

If S is indexed by the function $f : I \to S$ we write the element $f(i)$ just as s_i, if there is no risk of confusion.

Definition 1.2.6 *A sequence of length n on a set X is a function $x : \{0, 1, \cdots, n-1\} \to X$.*