A Trip from Machine Learning to Measure and Probability

Shichang Zhang

02/18/2022

Preface

These notes go over some core concepts and terminologies in measure-theoretical probability theory. The target readers are those who are interested in doing research in artificial intelligence/machine learning/data mining, having background in basic probability, but haven't seen measure-theoretical probability theory yet. The goal is to help readers to understand modern machine learning research papers written by statisticians and hopefully also make readers appreciate the necessity and beauty of measure theory. These notes emphasize on intuition so proofs are omitted, rigorous theorems are only included as needed, and *Remarks* may be skipped at the first reading. The scope of these notes is far below doing actually research in probability theory, and readers are encouraged to check the sources these notes rely on for a rigorous treatment of this topic: lecture notes by Sourav Chatterjee, lecture notes by John Hunter, and lecture notes by Amir Dembo (the most rigorous and detailed).

The notes are under development so any feedback regarding errors, typos, or general comments are welcome. (shichang@cs.ucla.edu).

1 Measure

The first key concept is the *measure*. Given a set Ω, a measure is a special set function μ that maps a subset $A \subseteq \Omega$ to a number in $[0, \infty]$. Intuitively, $\mu(A)$ represents the "size" of A, which can be mass of a set, length, area, and volume of a geometric shape, or more abstractly, probability of an event. We would like measures to match our common intuition of these notions. However, not all subsets $A \in 2^\Omega$ can be rigorously assigned a notion of "size" that align with intuition (an example is the Vitali set). We thus only talk about measures of *measurable* sets, for which we can give a well-defined definition of "size". This is where a weird mathematical object called the σ-algebra comes into play. A σ-algebra is simply a collection of subsets of Ω specifying which subsets are measurable.

Definition 1.1 (σ-algebra)**.** Given a set Ω, $\mathcal{F} \in 2^{\Omega}$ is a σ-algebra if

(a) $\emptyset \in \mathcal{F}$
(b) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
(c) $A_i \in \mathcal{F} \implies \cup_i A_i \in \mathcal{F}$, where $i \in \mathbb{N}$ is countable

Remark 1.2. If we keep property (b), equivalent properties of (a) and (c) are

(a') $\Omega \in \mathcal{F}$
(c') $A_i \in \mathcal{F} \implies \cap_i A_i \in \mathcal{F}$, where $i \in \mathbb{N}$ is countable (by DeMorgan's law)

Notice that though we haven't formally defined what a measure is, by definition of the σ-algebra, a subset A is measurable as long as it is in the σ-algebra. This is pretty much like how a topology defines openness. Similarly as a topological space, we now define a *measurable space*.

Definition 1.3 (Measurable Space, Measure, and Measure Space)**.** Given a set Ω and a σ-algebra $\mathcal{F}$ on Ω. The pair $(\Omega, \mathcal{F})$ is called a *measurable space*, and a *measure* $\mu : \mathcal{F} \to [0, \infty]$ is a set function satisfying the following properties.

(a) Non-negativity, i.e. $\mu(A) > 0$ for all $A \in \mathcal{F}$
(b) Null empty set, i.e. $\mu(\emptyset) = 0$
(c) Countable additivity, i.e. $\mu(\cup_i A_i) = \sum_i \mu(A_i)$ for **disjoint** $A_i \in \mathcal{F}$, where $i \in \mathbb{N}$ is countable

The triplet $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*.

To construct a measure space, first equip any set with a σ-algebra to produce a measurable space, and then equip the measurable space with any μ satisfying the three properties above to produce a measure space. Notice that these properties match our expected behavior of "size". Generate the σ-algebra is a key step for such construction. However, Definition 1.1 only tells us how to justify whether a collection of subsets is a σ-algebra, but not how to generate one, nor even how to concisely describe one.

Simple σ-algebras can be easily described by enumerating subsets in it. For example, the simplest σ-algebra is $\{\emptyset, \Omega\}$. Another example for $\Omega_0 = \{1, 2, 3\}$ is $\mathcal{F}_0 = \{\emptyset, \{1\}, \{2, 3\}, \Omega_0\}$. More practical σ-algebras, on the other hand, are hard to describe by enumeration, especially when there are uncountably many sets. We thus develop the following definition to help us describe and generate a σ-algebra, where we demand it to include the sets we care about while only containing the minimum required sets for it to be a valid σ-algebra.

Definition 1.4 (Generated σ-algebra)**.** Given an index set Γ (not necessarily countable) and its corresponding collection of subsets $\{A_\gamma \subseteq \Omega | \gamma \in \Gamma\}$ (short noted as $\{A_\gamma\}$), we denote the smallest σ-algebra containing $\{A_\gamma\}$ by $\sigma(\{A_\gamma\})$. That is

$$\sigma(\{A_\gamma\}) = \cap\{\mathcal{F} | \mathcal{F} \in 2^{\Omega} \text{ is a } \sigma\text{-algebra and } \{A_\gamma\} \in \mathcal{F}\}$$

We call $\sigma(\{A_\gamma\})$ the σ-algebra generated by $\{A_\gamma\}$ and $\{A_\gamma\}$ the *generator*.

Definition 1.4 gives us a convenient way to describe, generate, and compare σ-algebras through their generators. The $\mathcal{F}_0$ mentioned above can be described as $\sigma(\{1\})$. For comparison, $\sigma(\{A_\gamma\}) = \sigma(\{A_\beta\})$ if and only if $\{A_\gamma\} \subseteq \sigma(\{A_\beta\})$ and $\{A_\beta\} \subseteq \sigma(\{A_\gamma\})$.

A special type of generated σ-algebra that comes up all the time is the Borel σ-algebra.

Definition 1.5 (Borel σ-algebra)**.** Suppose Ω is a topological space, i.e. equipped with a notion of open sets, the Borel σ-algebra $\mathcal{B}_\Omega$ is defined as $\mathcal{B}_\Omega = \sigma(\{O \subseteq \Omega | O \text{ is open}\})$. An element of $\mathcal{B}_\Omega$ is called a Borel set meaning that the set can be constructed by applying the σ-algebra properties in Definition 1.1 on open sets.

Remark 1.6. When $\Omega = \mathbb{R}$, $\mathcal{B}_\Omega$ is usually denoted as $\mathcal{B}$. Other than the definition using all open sets as the generator, there are many equivalent expressions for $\mathcal{B}$ in terms of smaller interval generators, i.e.

$$\begin{aligned}\mathcal{B} &= \sigma(\{(a,b)|a < b \in \mathbb{R}\}) = \sigma(\{[a,b]|a < b \in \mathbb{R}\}) = \sigma(\{[a,b)|a < b \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty,b]|b \in \mathbb{R}\}) = \sigma(\{(-\infty,b]|b \in \mathbb{Q}\})\end{aligned}$$

So far we have discussed how to construct a measurable space $(\Omega, \mathcal{F})$ by generating the σ-algebra with a generator. The next step is to define the measure μ to make it a measure space. Just like the problem we faced above, it is often impossible to explicitly enumerate the value of μ on all $A \in \mathcal{F}$. Instead, we want to specify the values of a set function μ_0 on a generator $\mathcal{A}$ and then extend μ_0 on $\mathcal{A}$ to an actual measure μ on $\mathcal{F}$. (Notice that we don't say μ_0 is a measure because a measure is only defined on a measurable space, and $\mathcal{A}$ is not expected to be a σ-algebra already, which means $(\Omega, \mathcal{A})$ is not a measurable space). Such extension is nontrivial. Ideally, we would want $\mathcal{A}$ to be smaller than $\sigma(\mathcal{A})$, and we would also want the guaranteed existence and uniqueness of μ.

Given these requirements, we choose the *algebra* defined below as our generator. It defer from σ-algebra by property (c) in Definition 1.1, where algebra is only closed under unions rather than countable unions.

Definition 1.7 (Algebra)**.** $\mathcal{A} \in 2^\Omega$ is an algebra if

(a) $\emptyset \in \mathcal{A}$
(b) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$
(c) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$

We pick algebras as generators because they are better behaved and much smaller collections than σ-algebras (think about why smaller before checkout this example). More importantly, the following theorem guarantees the existence and uniqueness of an extension from a set function on an algebra.

Theorem 1.8 (Carathéodory's Extension Theorem)**.** *If μ_0 is a countably additive set function on an algebra $\mathcal{A}$. Then there exists a measure μ on $(\Omega, \sigma(\mathcal{A}))$ such that $\mu = \mu_0$ on $\mathcal{A}$. Furthermore, if $\mu_0(\Omega) < \infty$, then μ is unique.*

Remark 1.9. In Theorem 1.8, the uniqueness of μ requires μ_0 to be *finite* ($\mu_0(\Omega) < \infty$). This finite condition can actually be replaced by a weaker condition called *σ-finite*, which is more commonly found in the statement of this theorem. μ_0 being σ-finite only requires Ω to be covered by a countable union of sets where each of them has a finite measure. That is μ_0 is σ-finite if

$$\exists \text{ a countable collection } \{A_i\} \text{ s.t. } \Omega = \cup_i A_i \text{ and } \mu_0(A_i) < \infty \text{ for all } i$$

σ-finite is considered as "the next best thing than finite" because you can approximate $\mu_0(\Omega)$ using a sequence of finite numbers although itself might be infinite.

We now know how to specify values of a measure on a σ-algebra. We are almost there for a concrete example of a measure.

Proposition 1.10. *Given Ω, a measure μ can be constructed following either of the two approaches*

(a) Start from a generated σ-algebra $\mathcal{F}$. Since we know $(\Omega, \mathcal{F})$ is a measurable space, we can pick any set function μ that satisfies the three properties in Definition 1.3 as a measure.

(b) Start from a set function μ with some ideal properties. We construct a σ-algebra that is compatible with these ideal properties of μ, which at the same time contains sets required by the three properties in Definition 1.3.

The approach (a) in Proposition 1.10 seems more straight forward upon our discussion so far. It follows the order of $\mathcal{F}$ and then μ. Whatever set is in $\mathcal{F}$, it is by definition measurable. In this case, μ **doesn't** decide whether a set is measurable or not so forcing some special "non-measurable" sets in $\mathcal{F}$ makes define a meaningful μ impossible (the "non-measurable" here is not a standard description. Roughly speaking, by "non-measurable", I mean we cannot assign a notion of "size" to it that matches our intuition of length, volume, and etc. To be more strict, "non-measurable" means not Lebesgue measurable, which we will define below). For example, the Vitali set $\mathbf{V} \subseteq [0, 1]$ is "non-measurable". However, we can define a σ-algebra $\mathcal{F} = \{\emptyset, \mathbf{V}, \mathbf{V}^c, \Omega\}$, and $\mathbf{V}$ will be measurable by definition. Given such $\mathcal{F}$, μ cannot be something meaningful like the "length" of $\mathbf{V}$, but only some trivial function like the zero measure, i.e. $\mu : \mathcal{F} \to 0$.

The approach (b) is a more desirable approach, we can follow it to define a set function μ with properties we want. Then we collect "μ-measurable" (a non-standard description, only in contrast to the "non-measurable" above, but will be made concrete later) sets as a generator and generate a σ-algebra compatible with μ. In this sense, the measure μ **does** decide whether a set is "μ-measurable" or not. Next, we go through a concrete example of constructing the Lebesgue measure following this approach.

Proposition 1.11. *Let $\mathcal{A}$ be the set of all subsets of $\mathbb{R}$ that are finite disjoint unions of half-open intervals of the form $(a, b] \cap \mathbb{R}$, where $-\infty \leq a \leq b \leq \infty$ (we include $\cap\mathbb{R}$ to ensure the interval is (a, ∞) when $b = \infty$). Let $\lambda : \mathcal{A} \to \mathbb{R}$ be a* *set function satisfying*

$$\lambda(\cup_{i=1}^{n}(a_i, b_i] \cap \mathbb{R}) = \sum_{i}^{n}(b_i - a_i)$$

Notice that the property of λ is exactly what we want for length of intervals, which is also the property of the Lebesgue measure as a natural notion of "size". The construction of the Lebesgue measure follows these steps below.

1. The $\mathcal{A}$ defined in Proposition 1.11 is an algebra
2. The σ-algebra $\mathcal{A}$ generates is the Borel σ-algebra, i.e. $\sigma(\mathcal{A}) = \mathcal{B}$
3. The λ defined in Proposition 1.11 is countably additive and σ-finite.
4. By Theorem 1.8, there is a measure as the unique extension of λ on $\mathcal{B}$, which we call the Lebesgue measure.

Following the steps above we get the Lebesgue measure properly defined on $(\mathbb{R}, \mathcal{B})$, with $\mathcal{B}$ stands for the Borel σ-algebra. In other words, all sets in $\mathcal{B}$ are compatible with the Lebesgue measure. However, $\mathcal{B}$ doesn't give us all the "λ-measurable" (Lebesgue-measurable) sets, which form a strictly larger σ-algebra than $\mathcal{B}$. For practical purpose, $\mathcal{B}$ is often good enough, but as promised we will properly define Lebesgue-measurable and construct the Lebesgue measure on all such sets, since Lebesgue-measurable is quite a frequently used term that is easily confused with Borel (defined in Definition 1.5) or Borel-measurable (haven't defined yet).

Next we introduce how we can generate a σ-algebra containing all "μ-measurable" sets for a given μ, and then we go through the special case for μ being the Lebesgue measure.

We first define a new function called the *outer measure* to help us construct the proper σ-algebra, which relaxes the countable additivity property of a measure stated in Definition 1.3.

Definition 1.12 (Outer Measure)**.** Given a set Ω, a set function $\mu^* : 2^\Omega \to [0, \infty]$ is an outer measure if it satisfies the following properties.

(a) Non-negativity, i.e. $\mu^*(A) > 0$ for all $A \in 2^\Omega$

(b) Null empty set, i.e. $\mu^*(\emptyset) = 0$

(c) Monotone, i.e. $A \subseteq B \implies \mu^*(A) \leq \mu^*(B)$

(d) Countable sub-additivity, i.e. $\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$ for **disjoint** $A_i \in \Omega$, where $i \in \mathbb{N}$ is countable

Note that an outer measure is not necessarily a measure because of the relaxation, but a measure is an outer measure because monotone and countable sub-additivity follow from countable additivity. Unlike measures, the definition of an outer measure doesn't rely on a σ-algebra. Rather, we use μ^* to pick out the μ^**-measurable* sets defined below, which form a σ-algebra.

Definition 1.13. Given μ^* as an outer measure on a set Ω, a subset $A \subseteq \Omega$ is called μ^*-measurable if for all $S \subseteq \Omega$,

$$\mu^*(A) = \mu^*(A \cap S) + \mu^*(A \cap S^C)$$

The Definition 1.13, is sometimes referred as the Carathéodory criterion or the "clean division" criterion. It basically says that a μ^*-measurable set should be cleanly divided to two pieces by any subset $S \subseteq \Omega$ and its complement. Nothing from this division should be extra or left out when measured by μ^*. Although this definition is not very intuitive, it is a very clean way of defining measurable sets (Curious readers can read more about how Lebesgue defined measurable sets in the first place using Lebesgue inner measure and Lebesgue outer measure, see these notes. Lebesgue's approach was more intuitive but much more complicated. After that Carathéodory provided this equivalent and cleaner criterion).

The important theorem about the outer measure that serves our purpose is the following, which gives us the σ-algebra we need, and the measure that follows.

Theorem 1.14. *Given a set Ω and an outer measure μ^* on Ω. Let $\mathcal{F}$ be the collection of all μ^*-measurable sets in 2^Ω. Then $\mathcal{F}$ is a σ-algebra and μ^* is a measure on $\mathcal{F}$.*

Theorem 1.14 can be used to construct all kinds of measures from the corresponding outer measures. Following it, we are no longer constructing the σ-algebra from a generator, but pick out all measurable sets from the power set 2^Ω. A special case is the Lebesgue measure from the Lebesgue outer measure defined below.

Definition 1.15 (Lebesgue Outer Measure)**.** The Lebesgue outer measure of a subset $S \subseteq \mathbb{R}$, denoted by $\lambda^*(S)$, is defined as the infimum taken over all countable collections of intervals I_i whose union covers S. That is, with $I_i = (a_i, b_i)$ denotes the interval and $l(I_i) = (b_i - a_i)$ denotes the length,

$$\lambda^*(S) = \inf\{\sum_{i=1}^{\infty} l(I_i) | S \in \cup_{i=1}^{\infty} l(I_i)\}$$

Remark 1.16. The interval I_i being open is an arbitrary choice. They can be picked to be closed, half-open, etc.

The more general construction of the Lebesgue measure on all Lebesgue measurable sets follows these steps below.

1. The λ^* defined in Definition 1.15 is an outer measure
2. By Theorem 1.14, λ^* generates an σ-algebra $\mathcal{F}$ with all λ^*-measurable (Lebesgue measurable) sets

3. Restriction of λ^* on $\mathcal{F}$ gives the Lebesgue measure λ

This construction gives the same Lebesgue measure as the other construction we discussed above but on a strictly larger σ-algebra $\mathcal{F}$ containing $\mathcal{B}$. (For practical purpose, $\mathcal{B}$ is often good enough, but if you are really curious, here is an example of a Lebesgue-measurable but not Borel set). In fact, $\mathcal{F}$ is a *completion* of $\mathcal{B}$ which contains all Borel sets and measure-zero sets as we define below.

Definition 1.17 (Complete Measure Space)**.** We call a measure space $(\Omega, \mathcal{F}, \mu)$ *complete* if any subset S of any $A \in \mathcal{F}$ with $\mu(S) = 0$ is also in $\mathcal{F}$. The $\mathcal{F}$ is also referred as a complete σ-algebra for the measure μ when mentioned solely. For every measure space, there exists a complete measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$, called the *completion* of $(\Omega, \mathcal{F}, \mu)$, such that $\mathcal{F} \subseteq \bar{\mathcal{F}}$ and $\bar{\mu} = \mu$ on $\mathcal{F}$.

Remark 1.18. We can define the collection of μ-null sets as $\mathcal{N} = \{N | N \subseteq A \text{ for } A \in \mathcal{F} \text{ and } \mu(A) = 0\}$. Then the completion in Definition 1.17 is given by $\bar{\mathcal{F}} = \{A \cup N | A \in \mathcal{F}, N \in \mathcal{N}\} = \sigma(\mathcal{F}, \mathcal{N})$ and $\bar{\mu}(A \cup N) = \mu(A)$. This means we can throw the μ-null sets into a σ-algebra to complete it. The completion of the Borel σ-algebra with λ-null sets is called the Lebesgue σ-algebra, which contains all Lebesgue measurable sets and is strictly larger than the Borel σ-algebra.

The Lebesgue measure is perhaps the most important measure. It can be extended to $\mathbb{R}^n$ following similar steps like above, and it has the following properties align with our intuition of length, area, and volume in $\mathbb{R}^n$

Proposition 1.19. *The Lebesgue measure λ on $\mathbb{R}^n$ satisfies*

(a) Non-negativity, i.e. $\lambda(A) \geq 0$ for any set A.
(b) Matching length of intervals, i.e. $\lambda([a, b]) = b - a$
(c) Translation invariance, i.e. $\lambda(A) = \lambda(A + x)$ for any $x \in \mathbb{R}^n$, where $A + x = \{a + x | a \in A\}$.
*(d) Countable additivity, i.e. $\lambda(\cup_i A_i) = \sum_i \lambda(A_i)$ for **disjoint** $A_i \subseteq \mathbb{R}^n$, where $i \in \mathbb{N}$ is countable*

These four properties holds on all Lebesgue measurable sets and have been proven to be impossible to hold simultaneously for all sets of $\mathbb{R}^n$.

Remark 1.20. Being a notion of "size", the Lebesgue measure is different from cardinality or density. Although finite and countable sets all have Lebesgue measure zero, there are sets with uncountably many points but have measure zero, e.g. the Cantor set or other examples. There are even more interesting sets that are nowhere dense (contain no intervals) but with positive measures, e.g. the fat Cantor set.

2 Measurable Functions and Integration

To be continue