Recap of Probability Theory

Segwang Kim

February 18, 2023

Why do we learn probability theory for deep learning? The mathematical formulation of probability is pretty far from what we do in deep learning. Nonetheless, probability is an essential tool to define machine learning problems or formulate loss functions of neural networks. Once and for all, let's learn probability theory as rigorous as possible in the view of practitioners. I hope this recap of probability theory will prevent a feeling of doubt about mathematical formulations which you may have someday.

1 Probability Space

Definition) σ-field. A family $\mathcal{F}$ of subsets of Ω (=sample space), is a σ-field if

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \;\Rightarrow\; A^c \in \mathcal{F}$
3. $A_1,\ A_2,\ \cdots \in \mathcal{F} \;\Rightarrow\; \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

We say that $(\Omega, \mathcal{F})$ is an event space. Each element of $\mathcal{F}$ is called an event.

Intuitively, σ-field is a playground for *probability* which satisfies countable additivity.

Exercise) event space of coin toss. For one toss of a fair coin, we observe its outcome which is either Head ($= H$) or Tail ($= T$). Define the event space $(\Omega, \mathcal{F})$ for the coin toss.

Definition) Borel-field. For the set $\mathcal{T} = \{A \in \mathbb{R} \mid A :\ \text{open}\}$ of all open sets in $\mathbb{R}$, the smallest σ-field that contains $\mathcal{T}$ is called Borel-field, denoted by $\mathcal{B}(\mathbb{R})$[1]. In other words, Borel-field is the σ-field generated by $\mathcal{T}$.

Definition) probability measure. For a σ-field $\mathcal{F}$ of a sample space Ω, a set function $\mu : \mathcal{F} \to \mathbb{R}$ is called measure if

[1] The necessity for Borel-field, instead of the set of all subsets in $\mathbb{R}$, is explained in Section 8.1.

1. $\forall A \in \mathcal{F},\ \mu(A) \geq 0$
2. $\forall A_i \in \mathcal{F}\ (i = 1,\ \dots),\ A_i \cap A_j = \emptyset\ (i \neq j)\ \Rightarrow\ \mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$: countable additivity

Example) typical measures. The followings are noteworthy measures:

1. If a measure μ satisfies $\mu(\Omega) = 1$, μ is called probability measure or simply probability.
2. If a measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies $\lambda([a, b]) = b - a$, λ is called Lebesgue measure (Not precise).

Bear in mind that **countable additivity** is key for all of useful theorems about probability, such as central limit theorem (CLT), strong law of large numbers (SLLN), and so on.

Definition) probability space. For a sample space Ω and its σ-field $\mathcal{F}$ and probability measure μ, $(\Omega, \mathcal{F}, \mu)$ is called a probability space.

In the sequel, we denote a probability space over $\mathbb{R}$ as $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$, especially.

Exercise) probability space of coin toss. Define the probability space $(\Omega, \mathcal{F}, \mu)$ for one toss of a fair coin.

2 Random Variable

Definition) random variable. For probability space $(\Omega, \mathcal{F}, \mu)$, a mapping $X : \Omega \to \mathbb{R}$ is a random variable (or r.v.) if

$$\forall B \in \mathcal{B}(\mathbb{R})\ \Rightarrow\ \{w \in \Omega : X(w) \in B\} \in \mathcal{F}$$

In other words, X is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable. Denote $\{w \in \Omega : X(w) \in B\}$ as $(X \in B)$. The value of X at a point $w \in \Omega$ is called a realization (or observation) of X.

Note the followings:

- The measurable condition is natural since we define the probability of $B \in \mathcal{B}(\mathbb{R})$ via its inverse image by X.
- Elements of the sample space Ω can be thought of as all the different possibilities that *could* happen; while a realization $X(w) \in \mathbb{R}$ can be thought of as the value X attains when one of the possibilities *did* happen.

Exercise) rolling a dice. Define the probability space $(\Omega, \mathcal{F}, \mu)$ for one rolling of a dice. Then, consider a r.v. $X : \Omega \to \mathbb{R} : w \mapsto w\%2$.

To check that a given function $X : \Omega \to \mathbb{R}$ is a r.v., testing $(X \in B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$ is daunting. Hence, the following propositions are useful.

Proposition) sufficient conditions for random variable. For $X : \Omega \to \mathbb{R}$,

1. if $(X < a) \in \mathcal{F}\ \forall a \in \mathbb{R}$, then X is a r.v.. Here, $(X < a) := (X \in (-\infty, a))$.
2. if $(X \geq a) \in \mathcal{F}\ \forall a \in \mathbb{R}$, then X is a r.v..

Proposition) random variables induce random variables. For r.v.s $X, Y : \Omega \to \mathbb{R}$, and $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$-measurable (or simply called measurable) function $f : \mathbb{R} \to \mathbb{R}$,

1. $X + Y$ is a r.v..
2. aX is a r.v. for $a \in \mathbb{R}$.
3. XY is a r.v..
4. $f(X)$ is a r.v..

3 Distribution

Definition) distribution function. A function $F : \mathbb{R} \to \mathbb{R}$ is called a distribution function, or simply distribution, if

1. F is monotonically increasing, i.e., $x \leq y \ \Rightarrow\ F(x) \leq F(y)$
2. F is right continuous, i.e., $\lim_{b \to 0^+} F(x + b) = F(x)\ \forall x \in \mathbb{R}$
3. F has left limits, i.e., $\exists \lim_{b \to 0^-} F(x + b) := F(x^-)\ \forall x \in \mathbb{R}$
4. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

Furthermore, we define $F^{-1}(u) := \inf\{x : F(x) \geq u\}$ for $0 < u < 1$ called inverse transform or quantile function.

Theorem) A distribution function gives the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For a distribution $F : \mathbb{R} \to \mathbb{R}$, there exists the unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying

$$\mathbb{P}((a, b]) = F(b) - F(a)\ \forall a, b \in \mathbb{R},\ a < b$$

4 Random Variable and Distribution

Now, we are ready to understand the following: probability distribution is completely characterized by cumulative distribution function of a r.v..

Proposition) r.v. induces distribution function. For r.v. X from $(\Omega, \mathcal{F}, \mu)$, $F_X : \mathbb{R} \to \mathbb{R} : x \mapsto \mu(X \leq x)$ is a distribution function. $F_X(x) = \mu(X \leq x)$ is called cumulative distribution function (cdf).

Example) The meaning of a r.v. X follows distribution $\mathcal{D}$, i.e., $X \sim \mathcal{D}$. For a distribution $\mathcal{D}$ over $\mathbb{R}$, such as the standard normal distribution whose pdf is $x \mapsto \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$, a r.v. X follows the distribution $\mathcal{D}$, or $X \sim \mathcal{D}$, means F_X and cdf of $\mathcal{D}$ are the same.

As a distribution function on $\mathbb{R}$ gives the probability measure on $\mathbb{R}$, we can conclude the following:

Proposition) r.v. induces distribution (=probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$). For a probability space $(\Omega, \mathcal{F}, \mu)$ and a r.v. $X : \Omega \to \mathbb{R}$, $F_X : x \mapsto \mu(X \leq x)$ is a distribution function and there exists the unique probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ s.t.

$$\mathbb{P}_X((a, b]) = \mu(X \in (a, b]) = F_X(b) - F_X(a) \ \forall a, b \in \mathbb{R}, \ a < b$$

Here, $\mathbb{P}_X$ is called a distribution of X.

Hence, probability distribution $\mathbb{P}_X$ is completely characterized by cumulative distribution function $F_X : x \mapsto \mu(X \leq x)$.

Intuitively, r.v. X push-forward the probability μ defined on $(\Omega, \mathcal{F})$ to the probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which we denote by $\mathbb{P}_X$. This can be generalized as follows.

Definition) distribution of general r.v.. For a r.v. $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, i.e., $\mathcal{F}/\mathcal{F}'$-measurable, and a probability measure μ on $(\Omega, \mathcal{F})$, the probability distribution of X is the pushforward measure $X_*\mu$, which is a probability measure on $(\Omega', \mathcal{F}')$ satisfying $X_*\mu := \mu X^{-1}$.

Again, for $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $X_*\mu = \mathbb{P}_X$.

Example) We care the distribution of a r.v. NOT its probability space. Consider a task to generate flower images by GAN.

1. Consider a sample space that is a set of all flowers and its probability space $(\Omega, \mathcal{F}, \mu)$. There is no need to define precisely as we do not care.
2. For a flower $w \in \Omega$, we have an *observation*, e.g., RGB-image taken by a camera. Formally, we have $X(w) \in \mathbb{R}^n$.
3. Hence, X gives the data distribution $\mathcal{D} := X_*\mu$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.
4. Meanwhile, the generator G of a GAN push-forward the measure $([0, 1]^k, \mathcal{B}([0, 1]^k))$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.
5. If the generator works well, the pushforward measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ by the camera $(=\mathcal{D})$ and that by G $(=F_G)$ are the same.
6. Note that original probability spaces of X and G are different.

4.1 Inverse Transform Sampling

This has nothing to do with probability theory so that the reader may skip this subsection.

Inverse transform sampling is a useful technique for generating random variables that follows desired distribution.

Definition) Inverse of cdf. For a cdf $F : \mathbb{R} \to [0, 1]$, we define its inverse F^{-1} as

$$F^{-1} : [0, 1] \to \mathbb{R} : u \mapsto \inf\{x \in \mathbb{R} : F(x) \geq u\}$$

This is called inverse distribution function or quantile function.

Proposition) Some properties of $F^{-1}(u)$. $F^{-1}(u)$ satisfies followings:

1. $u \mapsto F^{-1}(u)$ is non-decreasing.
2. $0 < u < 1 \;\Rightarrow\; -\infty < F^{-1}(u) < \infty$
3. $F^{-1}(u) \leq x \;\Leftrightarrow\; u \leq F(x)$
4. $F^{-1}(F(x)) \leq x$
5. $F(F^{-1}(u)) \geq u$

Exercise) Some properties of $F^{-1}(u)$. Check the above properties.

Theorem) Probability Integral Transform. The probability integral transform states that if X is a continuous random variable with cdf F_X, then the random variable $Y = F_X(X)$ has a uniform distribution on $[0, 1]$.

Proof. To show that $Y \sim Unif([0, 1])$, we need to check that the cdf of Y and that of $Unif([0, 1])$ are the same. Indeed,

$$\begin{aligned} F_Y(u) &= \mu(Y \leq u) \\ &= \mu(F_X(X) \leq u) \\ &= \mu(X \leq F_X^{-1}(u)) \quad (\because F^{-1} \text{ is non-decreasing and } X \text{ is continuous}) \\ &= F_X(F_X^{-1}(u)) \\ &= u \quad (\because X \text{ is continuous}) \end{aligned}$$

□

This theorem suggests inverse probability integral transform method for sampling random variables that follow a desired distribution **provided that we can compute the inverse distribution function.**

Inverse probability Integral Transform Given a continuous uniform variable $U \sim Unif([0,1])$ and an invertible cdf F_X, the random variable $X = F_X^{-1}(U)$ has distribution F_X. Hence, we can sample $X_1, \ldots, X_n$ as follows.

1. Generate random numbers $u_1, \ldots, u_n \sim Unif([0,1])$.
2. Find the inverse of the desired cdf, e.g., F_X^{-1}.
3. Compute $X_1 = F_X^{-1}(u_1), \ldots, X_n = F_X^{-1}(u_n)$.

Proof. For $U \sim Unif([0,1])$, we need to show that $F_X^{-1}(U)$ has F_X as its cdf. Indeed,

$$\begin{aligned} F_{F_X^{-1}(U)}(x) &= \mu(F_X^{-1}(U) \leq x) \\ &= \mu(U \leq F_X(x)) \quad (\because \text{refer to the above proposition 3}) \\ &= F_X(x) \quad (\because U \sim Unif([0,1])) \end{aligned}$$

□

5 Convergence

Definition) converge a.s.. For a probability space $(\Omega, \mathcal{F}, \mu)$, a seq of r.v.'s X_n converges almost surely to X, i.e., $X_n \xrightarrow{a.s.} X$, if

$$\exists C \in \mathcal{F} \text{ with } \mu(C) = 1 \text{ s.t. } \forall w \in C, X_n(w) \to X(w) \text{ as } n \to \infty$$

Definition) converge in prob. For a probability space $(\Omega, \mathcal{F}, \mu)$, a seq of r.v.'s X_n converges to X in probability, i.e., $X_n \xrightarrow{\mu} X$, if

$$\forall \epsilon > 0, \ \mu(|X - X_n| > \epsilon) \to 0 \text{ as } n \to \infty$$

Definition) convergence of measure. On a event space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, a seq of probability measures $\mathbb{P}_n$ weakly converges to $\mathbb{P}$, i.e., $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$, if

$$\forall x \in C(F) := \{x \in \mathbb{R} : F \text{ is continuous at } x\}, \ \mathbb{P}_n((-\infty, x]) \to \mathbb{P}((-\infty, x]) \text{ as } n \to \infty$$

Similarly, for a probability space $(\Omega, \mathcal{F}, \mu)$, a seq of r.v.'s X_n converges to X in distribution, i.e., $X_n \xrightarrow{d} X$, if $\mathbb{P}_{X_n}$ weakly converges to $\mathbb{P}_X$.

6 Expectation

Definition) simple function. A r.v. $X : \Omega \to \mathbb{R}$ is simple if X only takes a finite number of values. Equivalently, for disjoint sets $A_1, \ldots, A_k \in \mathcal{F}$, $X = \sum_{i=1,\ldots,k} a_i I_{A_i}$ where $I_A : \Omega \to \{0,1\}$ is an indicator function which is $x \mapsto 1$ if $x \in A$, otherwise 0.

Definition) expectation. For a probability space $(\Omega, \mathcal{F}, \mu)$, we define expectation of r.v. starting from simple function as follows.

1. We first define the expectation of indicator $I_A : \Omega \to \{0, 1\}$ as

$$\mathbb{E}I_A := \mu(A) =: \int_\Omega I_A d\mu$$

2. Then, we define the expectation of simple r.v. $X = \sum_{i=1,\dots,k} a_i I_{A_i}$ as

$$\mathbb{E}X := \sum_{i=1,\dots,k} a_i \mathbb{E}I_{A_i}$$

3. Next, we define the expectation of non-negative r.v. X as

$$\mathbb{E}X := \sup_{Y \le X, Y:\text{simple}} \mathbb{E}Y$$

4. Finally, we define the expectation of r.v. X as

$$\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^-$$

where

$$X^+ = \begin{cases} X & \text{if } X \ge 0 \\ 0 & \text{o.w.} \end{cases}$$

and $X^- = X^+ - X \ge 0$

We also denote $\mathbb{E}X$ as $\int_\Omega X d\mu$.

Theorem) Markov inequality. If $X \ge 0$ is a r.v. and $a > 0$, then

$$\mu(X \ge a) \le \frac{\mathbb{E}X}{a}$$

Proof. A simple function $aI_{X \ge a}$ is less than and equal to X, i.e., $aI_{X \ge a} \le X$. □

6.1 Computation of Expectation

So far, we have defined $\mathbb{E}X$. The immediate question is how to compute expectation?

6.1.1 Formal Computation

First, let's compute expectation formally. As for Riemann integration, we use anti-derivatives of given functions, i.e. fundamental theorem of calculus. As for the expectation, we typically use probability density function defined as follows:

Definition) pdf. A r.v. X with values in the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gives a measure $X_*\mu$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then, the density of X with respect to a Lebesgue measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the Radon–Nikodym derivative:

$$f_X = \frac{dX_*\mu}{d\lambda}$$

That is, f_X is any measurable function with the property that:

$$\begin{aligned}\mu(X \in B) = X_*\mu(B) &= \int_B d(X_*\mu) \quad \text{(are expressed differently but have the same meaning)} \\ &= \int_B f_X d\lambda = \int_B f_X(x)dx\end{aligned}$$

for any measurable set $B \in \mathcal{B}(\mathbb{R})$.

We also need the following theorem for the computation:

Theorem) change of variables. For a probability space $(\Omega, \mathcal{F}, \mu)$ and an event space $(\Omega', \mathcal{F}')$, let $X : \Omega \to \Omega'$ be a $\mathcal{F}/\mathcal{F}'$-measurable function, i.e., r.v.. Then, for any borel function $f : \Omega' \to \mathbb{R}$,

$$\int_\Omega f(X)d\mu = \int_{\Omega'} f d(X_*\mu)$$

We also denote $\int_{\Omega'} f d(X_*\mu)$ as $\int_{\Omega'} f dF_X$ since $X_*\mu$ and F_X are essentially the same.

Therefore, $\mathbb{E}[f(X)]$ can be computed by either (1) the expectation of $f(X) : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F}, \mu)$ or (2) the expectation of $f : \mathbb{R} \to \mathbb{R}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*\mu)$.

Now, we can compute expectation as follows:

Proposition) computation of expectation. With pdf and the above theorem,

1. Put $f : x \mapsto x$ and $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in the above theorem. Then,

$$\begin{aligned}\mathbb{E}(X) &= \int_\Omega X d\mu \\ &= \int_{\mathbb{R}} x dX_*\mu(x) = \int_{\mathbb{R}} x dF_X(x) = \int_{\mathbb{R}} x d\mathbb{P}_X(x) \quad \text{(different notations)} \\ &= \int_{\mathbb{R}} x \frac{dX_*\mu}{d\lambda} d\lambda(x) = \int_{\mathbb{R}} x f_X(x)dx\end{aligned}$$

   Note that we also denote $d\lambda(x)$ by dx ($\because \int_a^b d\lambda(x) = b - a = \int_a^b dx$).

2. Put an arbitrary measurable function f and $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in the above theorem. Then,

$$\begin{aligned}\mathbb{E}(f(X)) &= \int_\Omega f(X)d\mu \\ &= \int_{\mathbb{R}} f(x)dF_X(x) \\ &= \int_{\mathbb{R}} f(x)\frac{dF_X}{dx}dx = \int_{\mathbb{R}} f(x)f_X(x)dx\end{aligned}$$

Again, we do not care about the original probability space $(\Omega, \mathcal{F}, \mu)$. To hide μ, we prefer to use $\mathbb{P}_X$ or F_X instead of $X_*\mu$.

The meaning of $\mathbb{E}_{X\sim D}[f(X)]$ Practically, we have no knowledge about the original probability space $(\Omega, \mathcal{F}, \mu)$. Hence, the exact computation of $\mathbb{E}[f(X)]$ is impossible. However, if there is an assumption $X \sim \mathcal{D}$ with the distribution $\mathcal{D}$ whose pdf $f_{\mathcal{D}}$ is well-known, we can compute $\mathbb{E}f(X)$ as follows:

$$\mathbb{E}[f(X)] = \mathbb{E}_{X\sim\mathcal{D}}[f(X)] = \int_{\mathbb{R}} f(x) f_{\mathcal{D}}(x) dx$$

6.1.2 Empirical Computation

You may be heard of Monte Carlo integration or Law of Large numbers (LLN). We can empirically approximate the exact value of expectation via Monte Carlo integration. Monte Carlo integration works because of LLN.

7 Conditional Expectation

Definition) Conditional Expectation w.r.t. a sub-σ-algebra. Consider the following:

- $(\Omega, \mathcal{F}, \mu)$ is a probability space.
- $X : \Omega \to \mathbb{R}^n$ is a random variable on that probability space with finite expectation.
- $\mathcal{H} \subset \mathcal{F}$ is a sub-σ-algebra of $\mathcal{F}$.

Since $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{F}$, the function $X : \Omega \to \mathbb{R}^n$ is usually not $\mathcal{H}$-measurable, thus the existence of the integrals of the form $\int_H X \, d\mu|_{\mathcal{H}}$, where $H \in \mathcal{H}$ and $\mu|_{\mathcal{H}}$ is the restriction of μ to $\mathcal{H}$, cannot be stated in general.

However, the local averages $\int_H X \, d\mu$ can be recovered in $(\Omega, \mathcal{H}, \mu|_{\mathcal{H}})$ with the help of the conditional expectation. A conditional expectation of X given $\mathcal{H}$, denoted as $\mathbb{E}(X \mid \mathcal{H})$, is any $\mathcal{H}$-measurable function $\Omega \to \mathbb{R}^n$ which satisfies:

$$\int_H \mathbb{E}(X \mid \mathcal{H}) d\mu = \int_H X d\mu$$

for each $H \in \mathcal{H}$.

(existence) The existence of $\mathbb{E}(X \mid \mathcal{H})$ can be established by noting that $\nu^X : A \mapsto \int_A X d\mu$ for $F \in \mathcal{F}$ is a finite measure on $(\Omega, \mathcal{F})$ that is absolutely continuous with respect to μ, i.e., $\mu(A) = 0 \Rightarrow \nu^X(A) = 0$. If h is the natural injection from $\mathcal{H}$ to $\mathcal{F}$, then $\nu^X \circ h = \nu^X|_{\mathcal{H}}$ is the restriction of ν^X to $\mathcal{H}$ and $\mu \circ h = \mu|_{\mathcal{H}}$ is the restriction of μ to $\mathcal{H}$. Furthermore, $\nu^X \circ h$ is absolutely continuous with respect to $\mu \circ h$, because

$$\mu \circ h(A) = 0 \;\Leftrightarrow\; \mu(h(A)) = 0$$

implies (since $\nu^X \ll \mu$)

$$\nu^X(h(A)) = 0 \;\Leftrightarrow\; \nu^X \circ h(A) = 0.$$

Thus, we have

$$\mathbb{E}(X \mid \mathcal{H}) = \frac{d\nu^X|_{\mathcal{H}}}{d\mu|_{\mathcal{H}}} = \frac{d(\nu^X \circ h)}{d(\mu \circ h)}.$$

Definition) Conditional Expectation w.r.t. a r.v.. For r.v.s X, Y on a probability space $(\Omega, \mathcal{F}, \mu)$, we define the conditional expectation of X w.r.t. Y as follows:

$$\mathbb{E}(X \mid Y) := \mathbb{E}(X \mid \sigma(Y)).$$

where $\sigma(Y) = Y^{-1}(\mathcal{B}(\mathbb{R}))$.

Definition) Conditional Probability. For an event $A \in \mathcal{F}$, we define its conditional probability as the conditional expectation of I_A.

8 Appendix

8.1 Why we use Borel-field, instead of the set of all subsets in $\mathbb{R}$?

Note that measure is the generalization of Lebesgue measure λ. So, let's first consider a Lebesgue measure on the set of all subsets in $\mathbb{R}$. Then, it is natural to impose the Lebesgue measure to have the following intuitive properties:

1. The measure of an interval is its length, i.e., λ(a interval) = the length of the interval.
2. Measure is translation invariant, i.e., $\lambda(A + x) = \mu(A)$ for $A \subset \mathbb{R}$ where $A + a = \{a + x \mid a \in A\}$
3. Measure is countably additivity over countable disjoint unions of sets,

UNFORTUNATELY, SUCH MEASURE λ DOES NOT EXISTS. Hence, we need the domain of measure smaller than the set of all subsets in $\mathbb{R}$. As expected, that domain is Borel field.

References