Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Govind Gopakumar
Instructor: Purushottam Kar
Date: January 6, 2018

LECTURE
1

Introduction to the course

Overview

In this lecture, we shall discuss some basic theory necessary for studying learning theory. We shall cover σ-algebras, measures (and probability measures), as well as provide basic definitions for expectation, independence etc. Some examples of random variables are discussed, and we conclude with a motivating example of basic results in probability and statistic help us analyze learning algorithms.

1 Motivation

With the rise of machine learning models in today's world, it becomes imperative that we are equipped with the right tools to analyze what we use. Various questions that arise during traditional algorithm design can be asked during designing and implementing machine learning models as well, and these are the questions we seek to address in this course. Irrespective of what task we are modelling, be it estimation, regression, or classification, we should be able to reason about, and prove guarantees of the models we come up with. As pointed out in Vapnik (1998), the theme that we wish to pursue when studying learning theory is the question,

"*Does there exist something common across machine learning models? Do they all implement a general principle to arrive at their results?*"

1.1 What is learning theory?

Modern machine learning generally works on a simple principle - we observe some training data, learn our model from this, and then hope that our model generalizes well to test data. This gives rise to many questions, among them the two main questions that we would like to know are -

- How do we ensure that our model minimizes error on the training data? (Algorithmic learning theory)
- How do we ensure that our trained model performs well even on the unknown data given to us during testing (Statistical learning theory)

The design of a learning algorithm and its subsequent analysis allows us to make claims (and hopefully prove them too) about how our model may do when given training data. For testing performance, we will aim to use statistical properties of the data as well as our learning methods to make generalization claims. Crucially, we shall make assumptions on how similar the training and testing datasets are, and then use properties of our methods to ensure similar, if not identical performance on training and test datasets.

1.2 Why study learning theory?

Apart from allowing us to answer the fundamental questions that one should naturally ask when using machine learning, learning theory is useful for a number of other reasons. Studying learning theory, understanding how to design learning algorithms and analyze them allows us to :

- Set our expectations and goals realistically, with respect to performance in the real world. It is critical that when using some tool, we are aware of its limits and capabilities. With the rise of neural networks and deep architectures in the present, it becomes even more important that we aim to use them only when we can at least have a rough idea of what to expect from them.
- Allow methods to be accepted in society and the scientific community. Acceptance of new methods is necessary for further development, and eventual usage in practice. If we can establish the performance of our algorithm and provide conditions for good performance, this would ease its usage as well as interest among others.
- Gives us insights into how they work. Understanding the theoretical basis for any algorithm's performance allows us to understand better its shortcomings as well as how it can be improved. While many a method may not be practically feasible, but have theoretical guarantees, they provide a starting point for models that do work in practice, and are not far away from being understood completely.
- It is important to note at the outset that most learning theoretic results and analyses make certain assumptions about data or the task and it may very well be that the algorithms/estimators, when actually executed on real data, do not agree with the theoretical predictions exactly. The practical performance may at times be better or worse than what theory predicts.
- The above may be because the data does not satisfy the assumptions, or satisfies even stronger assumptions than those made by theory. This may also be because the proof technique was feeble and was unable to exploit the assumptions completely or properly. This provides the genesis of an ongoing struggle in all learning theoretic efforts to make assumptions more and more realistic, and proof techniques more and more powerful, so that data actually satisfies those assumptions and practical observations match theoretical predictions.

Exercise 1.1. As a reading exercise, go over the lecture notes at (Pathak et al. (2016)).

2 Probability Theory

The definitions below are paraphrased from (Ash and Doleáns-Dade (2012)). Readers are advised to refer to the full text for a more rigorous and detailed explanation of the concepts discussed here.

2.1 σ-fields and measures

Let us consider an experiment, in a general sense. Every experiment will have a set of outcomes, among which we may observe only one at every realization of the experiment. Consider the set of all possible outcomes the experiment could ever have, and denote this by Ω. Among this, when we wish to analyze, it might be useful to analyze a set of outcomes, instead of an individual outcome by itself. To this end, consider a collection of subsets of Ω, let us denote it by $\mathcal{F}$. Now, this represents the set of *events* that we are interested in, can make statements about, and are valid. For a reasonable collection of events, we can impose certain conditions. These conditions let us define the collection of events more formally as a σ-algebra or σ-field.

Definition 1.1 (σ-field)**.** A collection of subsets, $\mathcal{F} = \{E_1, E_2, \dots\}, E_i \subset \Omega$ is called a σ field if,

1. $\Omega \in \mathcal{F}$ (the sure event should always be a valid event)
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$ (the complement of a valid event should also be a valid event)
3. $A_1, A_2, \dots \in \mathcal{F} \implies \cup_{i=1}^{\infty} A_i \in \mathcal{F}$ (the union of countably many valid events should also be a valid event)

Remark 1.1. From these conditions follows that $\mathcal{F}$ is closed under finite/countable intersections.

Remark 1.2. The events $\Omega, \emptyset$ are often called *trivial* events since they are always present in any valid σ-field. The σ-algebra $\{\Omega, \emptyset\}$ is often called the *trivial* algebra since it contains only trivial events.

Definition 1.2 (Generated σ-fields)**.** Given a collection of subsets $\mathcal{G} = \{A_1, \dots\}$ where each subset $A_i \in \Omega$, the σ-field generated by $\mathcal{G}$ is the smallest set $\mathcal{F} \subset 2^{\Omega}$ such that $\mathcal{G} \subseteq \mathcal{F}$ and $\mathcal{F}$ is a valid σ-field. We often denote the generated σ-field as $\sigma(\mathcal{G})$.

Remark 1.3. Note that 2^{Ω} denotes the power set of Ω and is the collection of all subsets of Ω i.e. $2^{\Omega} = \{A : A \subseteq \Omega\}$. For example, if $\Omega = \{1, 2, 3\}$ then we have

$$2^{\Omega} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

Exercise 1.2. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. Find the σ-algebra generated by the collection $\mathcal{G} = \{\{1\}, \{2\}\}$.

Example 1.1. Let us consider the experiment of rolling a dice and recording the face that is shown. Naturally, the set Ω is given by $\Omega = \{1, 2, 3, 4, 5, 6\}$. Keeping in mind the definitions we have stated above, one particular σ field we could work with would be the set $\mathcal{F} = \{\{2, 3, 5\}, \{1, 4, 6\}, \emptyset, \Omega\}$. Here the only non-trivial "events" we are allowed to talk about is the dice showing a composite number, a prime number. Note that this algebra prohibits us from talking about the event that the outcome is a triangle number i.e. the event $\{1, 3, 6\}$ is not a valid event since this event is not present in the σ-field.

Another σ-field we could be interested in is $\mathcal{F} = \{\{1, 3, 5\}, \{2, 4, 6\}, \emptyset, \Omega\}$. Here, the outcomes we are are interested in is the dice showing an even or odd number, apart from the trivial events. We are not interested, nor can we talk about other specific events (like say, the dice showing up a 2, or 5). σ-algebras that prohibit certain events can be useful in special cases. It is important to note that the most "complete" σ-field $\mathcal{F} = 2^{\Omega}$ need not make sense when Ω is an infinite set.

Definition 1.3 (Measure)**.** When given a σ-field, we define by a *measure*, any non-negative real valued function μ on $\mathcal{F}$ such that :

1. $\mu(A) \geq 0, \forall A \in \mathcal{F}$
2. If $A_1, A_2, \dots$ are pairwise disjoint events i.e. all $A_i \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$, then any countable union of them must satisfy,

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

Definition 1.4 (Probability Measure)**.** A measure with the additional condition $\mu(\Omega) = 1$ is called a *probability measure.* We shall use $\mathbb{P}$ when referring to a probability measure.

Example 1.2. Let us define $\Omega = \{1, 2, 3\}$. Now, let $\mathcal{F} = 2^\Omega$. Define $\mu(E) = |E|/3$, for any $E \in \mathcal{F}, E \subset \Omega$. This is a simple variation of what is called the *counting measure.* Readers may verify that this does indeed satisfy all the conditions imposed on a probability measure.

The example above fits well with the elementary definition of "probability of an event". In that, our usual understanding of probability stems from the general definition, $P(A) = |A|/|\Omega|$. We shall see that the notion of probability measure generalizes this idea to arbitrary sets, of which cardinality may not be so easily computable.

Since we restrict ourselves to probability measures, we shall use $\mathbb{P}$ to indicate the same.

Definition 1.5 (Probability space)**.** The three objects we have so far defined give us what is called a *probability space.* A probability space is the 3-tuple defined by $(\Omega, \mathcal{F}, \mathbb{P})$

2.2 Random Variables

When working with a probability space, it may be convenient to further define a map from events to say, the real line. This allows us to condense the information contained within the event space. For instance, when rolling a pair of dice, the outcome space consists of all possible pairs of the form $(i, j), i, j \in \{1, \ldots 6\}$. If we are interested merely in the sum of the two faces, we could define a mapping from each of the outcomes to a value. So the outcome $(3, 5)$ would get mapped to the real value 8.

Definition 1.6 (Random Variable)**.** A real-valued *random variable* on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function from Ω to $\mathbb{R}$, i.e. that maps outcomes to real numbers[1][2]. We may have vector valued random variables as well in which case the random variable would map outcomes to vectors.

For a real valued random variable $X : \Omega \to \mathbb{R}$ and constants $a, b \in \mathbb{R}$, we will use the notation $E = \{X \in [a, b]\}$ to denote the event $E = \{\omega \in \Omega : X(\omega) \in [a, b]\}$. Note that $E \in 2^\Omega$. The discussion in the footnotes below indicate that random variables always ensure that not just $E \in 2^\Omega$, but in fact $E \in \mathcal{F}$ i.e. it is a valid event according to the σ-field. Similarly $\{X \leq t\}$ will denote the event $\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}$.

Definition 1.7 (Distribution function)**.** One manner of using the induced distribution is to define a distribution function for the random variable X. The *distribution function* (often termed cdf as well) is given by,

$$F(x) = \mathbb{P}\left[w : X(w) \leq x\right]$$

Definition 1.8 (Density function)**.** If there exists a nonnegative real valued function on $\mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, x \in \mathbb{R}$$

we term such an $f(t)$ the *density function* of the random variable X.

[1] For formal reasons, this function must be *measurable.* A real random variable X is measurable iff we have, for any $t \in \mathbb{R}$, the following guarantee $\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}$ or in shorthand, $\{X \leq t\} \in \mathcal{F}$ i.e. the set of outcomes on which the random variable assigns a value less than or equal to t must be a valid event, and this must happen for all $t \in \mathbb{R}$.

[2] We note that the above measurability requirement is actually there to make life very convenient for us. Since the probability measure is compelled to assign a probability value to every event in $\mathcal{F}$, this means we can most definitely assign a probability value to events of the kind $\{X \leq t\}$. We do not have to make any extra effort to assign such events a probability. Since events are closed under countable unions and negations, it can be verified that events of the form $\{X \in [a, b]\}$ are also assigned a valid probability value for any $a, b \in \mathbb{R}$.

The induced probability distribution is essentially all that we require when dealing with random variables. Often, we shall describe a random variable with just the distribution function. Here, there need not be any particular underlying probability space that needs to be defined along with the random variable.

Example 1.3 (Indicator random variable)**.** The *indicator random variable* is defined with respect to a particular event. Consider $E \in \mathcal{F}$. For this, we define the appropriate indicator random variable as,

$$\mathbb{I}_E(\omega) = \mathbb{I}(\omega \in E) = \begin{cases} 1; \text{ if } \omega \in E \\ 0; \text{ if } \omega \notin E \end{cases}$$

where $\mathbb{I}$ is the indicator function. This random variable takes values in the range $\{0, 1\}$. We shall see how these random variables are useful in our analysis.

Example 1.4 (Rademacher random variables)**.** The Rademacher random variable follows a discrete distribution, where it takes the values $\{-1, 1\}$ with equal probability. It is defined as,

$$R = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

Example 1.5 (Bernoulli random variable)**.** The Bernoulli random variable takes values in the discrete set $\{0, 1\}$. It is defined by the bias parameter p, as follows :

$$B_p = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability 1-p} \end{cases}$$

Remark 1.4. The set of all random variables forms a vector space. The addition operation arises naturally from the addition on the reals, and the additive identity is simply the $0(\omega) = 0$ random variable.

2.3 Expectation and Moments

When using random variables, we shall find it useful to also work with what is called the moment of a random variable. These capture the variation present in the random variable.

Definition 1.9 (Expectation)**.** The *expectation* of a random variable corresponds to our idea of mean. We shall see that it is easier to define an expectation for simpler, discrete random variables. For the case where X takes discrete values, the expectation can be obtained by a direct sum. Let us say that $\Omega = \{k_1, k_2, \ldots k_n\}$ (Implicit is the assumption $\mathcal{F}$ is the powerset of Ω). Thus, the expectation can be found by a direct sum,

$$\mathbb{E}X = \sum_{i=0}^{n} k_i \cdot \mathbb{P}\left[\omega : X(\omega) = k_i\right]$$

We note that when notational ambiguity is not an issue, we often use the notation $\mathbb{E}X$ to denote $\mathbb{E}[X]$. However, when delimiting the scope of the expectation operator (yes it is an operator from the vector space of random variables to reals, for real valued random variables) is important we will keep using parentheses to remove any ambiguity.

For cases where X is a real valued random variable, when X has a distribution function $f_X(x)$, we shall use the form,

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) dx$$

Remark 1.5. The formal defintion of expectation is arrived at by dividing the random variable into a positive random variable, and a negative random variable. For a positive random variable, the expectation is defined as,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\left[X \geq t\right] dt$$

Remark 1.6. While the above definition is simple, there are examples where the expectation does not exist (in contrast with being infinite, which is still a valid expectation). Interested readers may find details of this in the book (Ash and Doleáns-Dade (2012)) or even in (Chung (2001)).

Definition 1.10 (Moment)**.** Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. For $k > 0$, the quantity $\mathbb{E}\left[X^k\right]$ is called the *kth moment* of X. Note that the case where $k = 1$ is simply the expectation of the random variable.

Definition 1.11 (Variance)**.** We define the *variance* of a random variable as its second moment, when appropriately centered or shifted.

$$\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$$

Exercise 1.3. Show that $\mathbb{E}\left[X^2\right] \geq (\mathbb{E}X)^2$

2.4 Independence

The notion of independence of events is motivated by the idea that there can be a pair of events such that a statement concerning the occurence of any one event does not give us any information about the occurence of the other.

Definition 1.12 (Pairwise independence of events)**.** Two events A, B are said to be *pairwise independent* if and only if $\mathbb{P}\left[A \cap B\right] = \mathbb{P}\left[A\right] \cdot \mathbb{P}\left[B\right]$

This can also be extended to a set of events considered together. We define independence of a set of events as,

Definition 1.13 (Independent events)**.** Let I be an arbitrary index set, and let $E_i, i \in I$ be events in a given probability space. The events are *mutually independent* if and only if, for all finite collections of distinct indices from I, say, $\{i_1, \ldots, i_k\}$, and for all $k > 0$ we have

$$\mathbb{P}\left[E_{i_1} \cap \cdots \cap E_{i_k}\right] = \prod_{n=1}^{k} \mathbb{P}\left[E_{i_n}\right]$$

Note that restricting the above property for only $k = 2$ gives us the pairwise indpendence property.

Remark 1.7. Note that pairwise independence does not imply independence of events. Readers are encouraged to find examples of events where the former holds but the latter does not.

The same idea can be extended to random variables. However, doing so takes some care since we wish to gracefully extend the notion of independence we developed above for events. Below we give a usable notion of independence of random variables by reducing the question of independence of random variables to questions of independence of events which we already have.

Definition 1.14 (Pairwise Independent random variables)**.** A collection of real valued random variables $X_1, \dots X_n$ defined over a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be pairwise independent if for any $i, j \in [n]$ and any real values $s, t \in \mathbb{R}$ the events $\{X_i \leq s\}$ and $\{X_j \leq t\}$ are independent. This should be the case for any real values s, t and any pair of random variables. In particular, this ensures that we have, for any $i, j \in [n]$ and any real values $s, t \in \mathbb{R}$

$$\mathbb{P}\left[\{X_i \leq s\} \wedge \{X_j \leq t\}\right] = \mathbb{P}\left[X_i \leq s\right] \cdot \mathbb{P}\left[X_j \leq t\right]$$

We will often use the shorthand $\mathbb{P}\left[X_i \leq s\right] \equiv \mathbb{P}\left[\{X_i \leq s\}\right]$ to avoid excessive notation.

Definition 1.15 (k-wise Independent random variables)**.** A collection of real valued random variables $X_1, \dots X_n$ defined over a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be k-wise independent if for any real values $s_1, \dots, s_k \in \mathbb{R}$ the collection of the events $\{\{X_1 \leq s_1\}, \dots, \{X_n \leq s_n\}\}$ is k-wise independent. This should be the case for any set of real values $s_1, \dots, s_n$. In particular, this ensures that we have, for any $i_1, \dots, i_k \in [n]$ and any real values $s_1, \dots, s_k \in \mathbb{R}$

$$\mathbb{P}\left[\bigwedge_{j=1}^{k} \left\{X_{i_j} \leq s_j\right\}\right] = \prod_{j=1}^{k} \mathbb{P}\left[X_{i_j} \leq s_j\right]$$

Remark 1.8. The above combined with closure properties of σ-fields, should tell us that we should also have, for the setting described above, and for any $s_i, t_i \in \mathbb{R}$,

$$\mathbb{P}\left[\bigwedge_{j=1}^{k} \left\{X_{i_j} \in [s_j, t_j]\right\}\right] = \prod_{j=1}^{k} \mathbb{P}\left[X_{i_j} \in [s_j, t_j]\right].$$

Remark 1.9. A collection of real valued random variables $X_1, \dots X_n$ will be said to be mutually independent if it is k-wise independent for all $k \leq n$.

Remark 1.10. A more rigorous definition of the independence of random variables actually proceeds by first constructing σ-algebras induced by these random variables (a more careful treatment constructs something called a π-system which is a simpler construction than a σ-algebra) and then defines independence of random variables in terms of independence of those σ-algebras. For a real valued random variable, the σ-algebra associated with the random variable is the one generated by the following collection (recall the notion of generated σ-fields).

$$\mathcal{G}_X = \{\{\omega : X(\omega) \leq t\} : t \in \mathbb{R}\},$$

i.e. we consider the σ-field $\mathcal{F}_X := \sigma(\mathcal{G}_X)$. If we have two random variables X, Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then it is easy to verify that $\mathcal{F}_X, \mathcal{F}_Y \subset \mathcal{F}$ i.e. the σ-algebras corresponding to the random variables are *sub-algebras*. The random variables X, Y are then said to be independent if $\mathcal{F}_X, \mathcal{F}_Y$ are independent, which happens when for all $E_1 \in \mathcal{F}_X, E_2 \in \mathcal{F}_Y$, we have $\mathbb{P}\left[E1 \wedge E_2\right] = \mathbb{P}\left[E_1\right] \cdot \mathbb{P}\left[E_2\right]$. Note that since $E_1, E_2 \in \mathcal{F}$, all three quantities $\mathbb{P}\left[E1 \wedge E_2\right], \mathbb{P}\left[E_1\right], \mathbb{P}\left[E_2\right]$ make sense and are well defined.

Exercise 1.4. Show that if X, Y are independent real, random variables, then $\mathbb{E}[X \cdot Y] = \mathbb{E}X \cdot \mathbb{E}Y$. You may assume that both random variables are supported on the same set, their expectations exist and are finite.

Exercise 1.5. Show that if X, Y are independent random variables and $f : \mathbb{R} \to \mathbb{R}$, then $W = f(X)$ and $Z = f(Y)$ are independent too. Does this hold if f is identically a constant?

Exercise 1.6. Show that if X, Y are independent random variables and $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, then $Z = f(X)$ and $W = g(Y)$ are independent too.

Exercise 1.7. Construct a setting where X, Y are not independent but for some function f, $f(X) = Z$, $f(Y) = W$, Z, W are independent.

3 A toy learning problem

Our lecture closes with a motivating example of how we can analyze even simple learning methods, and what sort of results we shall seek in this course. Consider the following game between two players, `Alice` and `Bob`. They are given two coins, say C_1, C_2 with unknown biases p_1, p_2. At each turn, `Alice` tosses one of the coins. Depending on the outcome,

$$\text{Heads} \rightarrow \texttt{Alice} \text{ gives } \texttt{Bob} \text{ ₹}1$$

$$\text{Tails} \rightarrow \texttt{Bob} \text{ gives } \texttt{Alice} \text{ ₹}1$$

This game will be repeated, say a million times and it is clear that even if the coin being used is slightly biased, one of the players is expected to go home a very rich person! For instance, even if the coin used had a bias of 0.45 (i.e. landed heads with only 45%), Alice is expected to go home around ₹50000 richer than Bob at the end of the million trials.

Before the game starts, `Alice` is allowed to inspect both the coins and do whatever she wishes with them (apart from of course, modifying them). Devise a strategy for `Alice` to ensure that she does not lose too much money, or if possible, gain more money over the course of multiple rounds than she gives `Bob` (for example, if both coins have large biases say $p_1 = 0.8$ and $p_2 = 0.9$ then there is no hope for Alice to actually make money from this game – all she can hope for is to choose a coin so that she the loses least amount of money).

Note that one strategy that immediately comes to mind is to ask Alice toss both coins C_1, C_2 multiple times, observe them, and then decide upon one of them as the more *advantageous* coin, i.e. the one with the smaller bias. `Alice` can then choose to use this coin, hence minimizing her loss, or making the most profit in the situation, as the case may be.

Our goal will be to design an algorithm that lets `Alice` discover the better coin, and prove how much loss/profit she could make if she played with that coin. Also, it is in our interest to bound how much error we could be making in discovering this optimal coin, and if there exists a number of trials beyond which we can be very certain that our estimates are correct.

Simple as this toy problem may sound, it actually describes pretty much the goal of all machine learning. Think of coins as learning models, and the biases of the coins as the probability of those models making an error on randomly chosen data. The whole point of machine learning is to discover the models that have the smallest error rates, which is really what `Alice` is trying to do with her toy coin problem. Studying this toy problem in some depth will allow us to build a sound base for investigations into realistic learning problems.

Here, our ideas of confidence and error motivate the study of our learning algorithm. We would like to ensure that the error in our estimation of the coin biases, say ϵ is low, with very high probability. Our probability of making this error should then be low, denoted as δ. We shall see how this central idea of making low errors in high probability arise again and again over the course of the semester.

References

Robert Ash and Catherine Doleáns-Dade. *Probability and Measure Theory.* Academic Press, Elsevier, 2012.

Kai Lai Chung. *A Course in Probability Theory.* Academic Press, Elsevier, 2001.

Kalpant Pathak, Bhaskar Mukhoty, and Purushottam Kar. Probability Theory, Lecture 3. *CS 773 Online Learning and Optimization*, 2016.

Vladimir Vapnik. *Statistical Learning Theory.* Wiley, 1998.