はじめに

この話を書こうと思ったきっかけ

世の中には機械学習に必要な数学の入門の話が溢れかえってる気がする。最低限必要な数学はこれですよ、みたいな。

でも、入門より先の話をする人が全然居ない気がする。測度論、とか言葉は出すけれど、その中身を語る人を見かけた事が無い。皆が話題にしている流行りの論文には結構難しい数学を前提としているのに、その前提の話をしている人はどこにも居ないように見える。

自分としては、最低限必要な話では無く、このくらいあれば十分、という方を知りたい。誰かに書いて欲しい気はするが、まずは自分が知る範囲で書いてみよう、と思った。

自分の知ってる範囲を書く

本当に書きたい事は、「機械学習に十分な確率はここまでです」という事を書きたいのだけれど、残念な事に私がそこまでは理解していない。

そもそもに十分、というのは個人差がある所で、例えば関数解析に詳しい人は関数解析的な議論を深める事で業界に貢献出来るし、幾何学に詳しい人は幾何学的な議論を深める事で、実解析に詳しい人は実解析的な議論を深める事で業界に貢献しているように見える。

そういう点からすると、皆が知らないような事を知っていると、それは武器になるという物であって、要らないという気はあまりしない。だから十分、というのは、最低限必要よりも定義が難しい。

そこで、自分が流行の論文を理解しようとした時に勉強した範囲を書いていこうと思う。ただ勉強した事を書くのでは無く、この論文のここを理解しようとしたらこれが必要と言われたのでこれを学んだ、というように、どこの論文から始まった話かを明確にしていきたい。

一応自分はプログラマとして機械学習に関わっている人間としては、標準的な程度の確率論の理解はあると思っている。しかも仕事でもそれを実際に使っているので、実務で実際に仕事をする場合の一サンプルにはなっているんじゃないか。

書く形式

確率論のトピックを幾つか、5 個か 6 個くらい選んで書いていく。例えば確率変数、とか。

確率変数とは何か、という事は、学ぶ数学の段階で定義が違うと思う。

- 入門的、古典的な確率論
- 測度論的な確率論の初歩
- 実解析的な確率論

これらは普通、別々の教科書になっていて、普通は順番に読んでいく必要がある。だから確率変数とは何かという事などを知りたいと思っても、それ以外の項目についての一段下の教科書を全部読んでおかないと、次に進めない、という事になっている気がする。

これをトピックごとに、縦につなげる側で話してみたい。

縦に話をする事で、それぞれの分野がどう違うのか、というのが、そんなに長い修行期間を経なくても分かるように出来るんじゃないか。

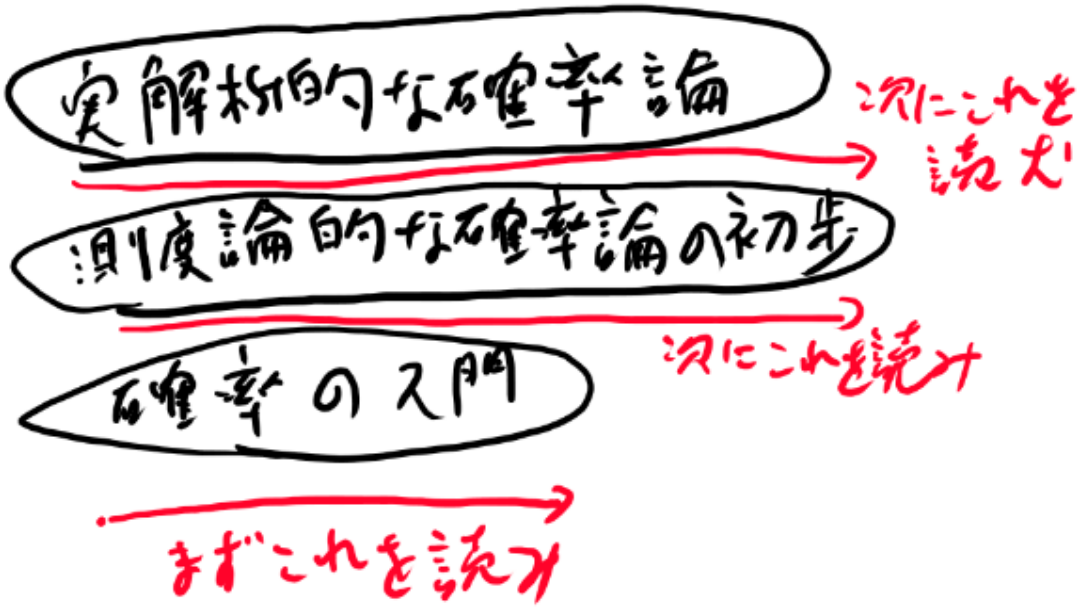

図 1: images/intro/0000.png

そしてそれらの違いから、どうして機械学習では実解析的な扱いや関数解析的な扱いが必要になりがちなのか、また逆に、それらを知らないで一段下のレベルの数学の理解でもどの位までは分かりそうか、みたいな話が出来たらなぁ、と思っている。

確率論の雑談を書いていきたい

数学の教科書を書きたい訳でも書く能力がある訳でも無いので、数学的な定義とかそういう話はあまり頑張っては書いていくつもりはありません。

個々の定義よりは、それらの定義と他の物との関係とか、機械学習ではどうやって出てくるかとか、どこが分かりにくいのかとか、どこが難しいのかとか、どこが自分には分からないのかとか、そういう雑談をしていきたい。

数学読み物みたいな感じで。

ただなるべくちゃんとした記述へのポインタは示していきたいと思っている。だいたいは教科書のページ数とかへの参照となる予定。

確率空間

最初に測度とかボレル集合族とか可測とかの話をしておきたいので、確率空間について話す所から始めます。

確率空間は私の知る限り、

1. 古典的な確率空間の定義

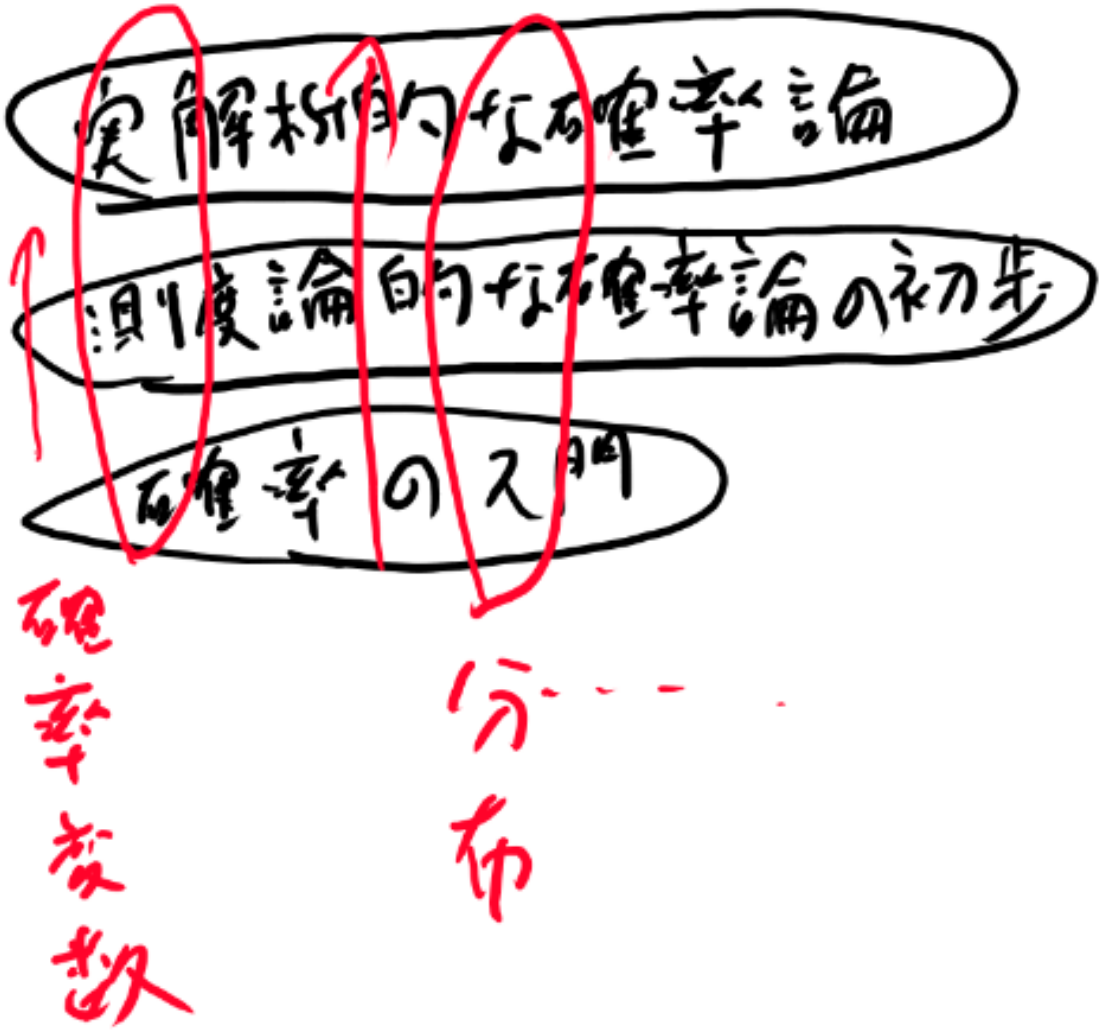

図 2: images/intro/0001.png

2. 測度論の入門的な確率空間の定義
3. 確率変数と law による定義（主に実解析でよく使う）
4. 分布による定義（主に関数解析でよく使う）

の 4 つの定義がある。

そして機械学習では 3 の定義が多くて論文でもだいたい 3 の定式化を使っていると思う。たまに 2 の定義もある。

機械学習ではよく使われる 3 の定義だけど、これは測度論の本ではあまり扱われてない事がある（特に入門書の場合）。自分は最初、3 を中心とした定式化で議論するというやり方を知らずにずいぶん混乱した。これは「測度論までやっておけば機械学習は十分」という神話の副作用に思う。

という事で、ここでは 3 や 4 の話をもっとしていきたい。

古典的な定義

普通、標本空間と事象と確率の話からぼんやりと確率空間の話をするのが古典的な確率論の入門書の始まりなのだが、これがなんだか良く分からない。というのはシグマ集合族と確率測度を出さずに、その話をしようとするからだ。

ここでは簡単に古典的な定義の話をして、それが全然わからん、という事から話を始めたい。

普通確率空間とは、$(\Omega, \mathcal{F}, P)$ の 3 つの構成要素からなる空間を言う。で、この 3 つは古典的には標本空間、事象、そして P と呼ばれる。

P には、古典的な世界ではたぶん名前が無いけれど、測度論的な用語で言えば確率測度の事。まずはこの 3 つの話を順番にしていきます。

標本空間

まず、サイコロを一つ振る、という事を考える。この時、標本空間とは出る可能性がある全てのサイコロの目の事です。この場合は

$$\{1, 2, 3, 4, 5, 6\}$$

となります。普通 Ω で表すので、

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

と書いておきましょう。

イメージとしては、確率的出来事のとりうる、全要素の事です。この標本空間からなにか一つの要素を取り出す事が、確率的な試行に対応します。

これは別段わからない事は無い。

事象族

さて、入門書で良くわからなくなるのが事象族です。これは本質的にはシグマ集合族の事なのに、入門書ではそれを持ち出さないでぼんやりと定義される。

事象というのは、確率を求めたい、標本空間の何らかの部分集合の事と言われる。なんじゃそりゃ。

で、そのあとに具体例が出てくる。例えば「偶数の目が出る」などが事象の例です、とか言ってくる。

事象は、標本空間の部分集合なので、集合です。例えば「偶数の目が出る」の場合は、

$$\{2, 4, 6\}$$

となります。この 3 つの要素を持った集合。

で、この事象を全部集めた物を事象族といいます。事象が集合なので、その事象を集めた物は、集合の集合という事になります。集合の集合は集合族と呼ばれるから、事象族と呼びます。

事象族の表記としては花文字 B とか花文字 F とかで書く。 F は F 集合族から来ているのか？B はボレル集合族ですかね。シグマ集合族の場合は F を使う場合が多い気がする。古典的な場合は event という事からか E を使う場合もある。

花文字というのは下みたいな文字の事です。

$$\mathcal{F}$$

で、その事象族の要素となる事象は、普通の大文字で書く。この場合は F。

$$F \in \mathcal{F}$$

なお、集合体も集合族と同じ意味。シグマ集合族はシグマ集合体と言っても良い。体をなしているかどうか、とか細かい話はあるかもしれないが、このシリーズでは細かい事は気にしない。

花文字、手描きでうまく書けないからやめて欲しいのだけれど、業界の習慣なので仕方ない。

確率関数 P

古典的にはなんて呼ぶのか良く知らないけれど、事象を引数として、その事象が起こる確率を返す関数を P と呼ぶ。

測度論の用語で言えば確率測度の事なんだけど、古典的な入門書では測度が無い状態ではぼやっと定義される。そもそも定義もごまかしなので、それを正しくはなんと呼ばれるかとか全然興味湧かない。なので調べない。どうせこの辺はいい加減な誤魔化しなので、細かい事はどうでもいいんです。

だけど、この P は割と具体的なので、厳密な定義は入門書では謎でも、感覚的には何なのかはわかりやすい。だから入門者が入門書を読んでいる段階でも、あまり苦労は無いはず。

例えば、

$$P(\text{偶数の目}) = \frac{1}{2}$$

とか、そういうものだ。こういう風に、事象 F を引数として、その確率を返す関数の事を P と呼ぶ。

ただそもそも事象とは何かとかぼやっとしてるので、その対象に対する関数も古典的な世界ではあんまり細かくは議論出来ない。だからぼやっとそういうもんだ、とわかれば、このレベルでは十分と言える。

入門書は、確率測度を元とした定式化を分かっている人が、それを古典的な言葉に翻訳して書いてある。でも、測度の定義とかを出さないので、結局測度論を分かっている人だけが分かる自己満足な記述になってしまいがち。そんな物に、分かるはずの無い入門者は苦労する事になる。 酷い話だ。

という事でこの辺わからない人は、あんまりわからないと深く考えず、とっとと測度論に行くのがオススメです。

古典的な確率空間

さて、さっぱり定義出来ていない物を合わせて定義もクソも無いのだが、これら 3 つを合わせて確率空間と呼ぶ。

$$(\Omega, \mathcal{F}, P)$$

3 つなのでトリプレットとか言ったりもする。

ちゃんと定義は出来てないから理解は出来てなくて当然だが、それぞれが具体的には何を指しているかを、具体例でちゃんと識別出来ておく必要はある。

記号	意味
Ω	標本空間、$\{1,2,3,4,5,6\}$ の事
$\mathcal{F}$	事象族、$\{\{$偶数の目$\}, \{$4以上の目$\}, \{2,3,5\}$など$\}$ 標本空間の部分集合の集まり。
P	呼び方は知らないけど、事象を引数にその事象が起こる確率を返す関数

古典的な確率空間でだいたいすべてを説明出来る

古典的なこれらの定義が何を指しているかをちゃんと理解しておけば、機械学習に出てくる理論的な事は、原理的には全部説明出来ると思う。機械学習の話をするには、本当は測度論とかは一切要らない。

だからすごいこの辺詳しい暇な友達が居たら、Deep Learning でわからない事が出てくる都度「古典的な言葉に翻訳して説明してくれよ！」って頼めば、古典論だけでだいたい問題は無いと思います。

ただ、職場とかでは誰も古典的な言葉で説明なんてしてくれないので、一人分働くには測度論とかが要るのだ。誰か流行りの論文を全部古典的な言葉に翻訳してくれればいいのにねぇ。

この、「アイデアを伝達する為に皆が使っているから実務家もここから先の数学が必要」というのが、ほとんどの実務家にとっての数学の現実だと思う。

だから逆に言うと、変な性質を持ったゼロ測度の集合の時の振る舞いとか、ジャンプする関数の片側極限の話とか、そういう理論的に際どいところの証明とかは、実務家視点では要らない。説明をする為の言葉

とか、証明の為のパターンとか、そういうのだけが必要、と自分は思います（異論歓迎です）。

だから機械学習屋で、自分が理論系論文を書く訳では無い大多数の人は、この説明に使われる、「言葉としての実解析」をどうやって学んでいくか、という事を考える必要があると思うし、「難しい事は全部飲み込む、言葉としての実解析を学ぶ本」とかあったらすごい良いと思う。誰か書いて。

入門的な測度論的確率空間

古典的な話なんかしたくてこの文書を書いているのでは無いのです。という事で次の測度論的な定義に進みます。古典的な確率空間の次は「入門的な測度論的確率空間」。

2012 年とかその辺の時代なら、このセクションのタイトルに「入門的な」は要らなかったと思う。「測度論的な確率空間を理解すれば機械学習に必要な確率論は全て理解出来たと言って良い（キリッ」とか言えた。

で、分かってない人も、難しい数学の話は「ちゃんと知りたい人は測度論を勉強しましょう」とか測度論って単語を出してイキっておけば分かってるフリが出来ている、という事になっていた。

平和な時代だった…

もちろん今は測度論的な確率空間の初歩を知っている程度では、流行りの論文などさっぱり何を言っているか理解出来ない。それでも、一応この「入門的な」測度論的確率空間の事を知っている人なら、さわり位は分かるように書くのが論文とかのマナーとなっている気がする。だから 2018 年現在でも、入門的な測度論的確率空間をちゃんと勉強する意味はある。

という事で 2018 年現在ではもはや「入門的な」とつけなくてはいけない測度論的な確率空間の話を簡単にしてみましょう。

といっても、そもそもに古典的な確率空間はこの測度論的な確率空間を誤魔化して説明しているだけなので、だいたい同じ物です。測度論的な確率空間も以下の 3 つの要素からなります。

$$(\Omega, \mathcal{F}, P)$$

このうち、標本空間は古典的な物も測度論的な物も変わらない。

違うのは事象族と P です。

事象族はシグマ集合族で、P は確率測度となります。このシグマ集合族と測度は、このシリーズで重要なので、「シグマとボレル集合族と測度」の章で独立して扱う事にします。

ちょっと前後しちゃいますが、一旦そちらを読んでから続きを読んでください。

以下では確率空間というコンテキストに絞って、シグマ集合族と確率測度の話をしていきたい。

シグマ集合族

確率空間を構成する 3 つの文字の一つ、シグマ集合族について。

厳密な定義はおいといて、シグマ集合族が指している物がどんな物なのかイメージしておくのは大切です。特にこれが標本空間の部分集合の集まり、という事はちゃんと理解しておかないと、論文が読めない。

シグマ集合族が指しているのは、古典的な例の事象族、と言っていた物です。

事象族はサイコロの目の例なら「サイコロの目が偶数」といか、「サイコロの目が 4 以上」とかそういう物でした。書き方はいろいろだけど、最終的には必ず Ω の部分集合で表せる。

シグマ集合族は、ある数学の性質を持った厳密に定義されている集合族の事だけど、機械学習の実務家的には数学の性質はそんなに重要じゃない。

事象族をちゃんと定式化するとシグマ集合族の性質を持ってないとまずいらしくて、だからその性質が要請されるだけで、事象族の事を指していると思っておいて良い。詳細はシグマ集合族の章を見てください。

まぁ開集合みたいなもんですよ。

確率測度

測度というものについてはシグマ集合族と測度の章で扱うのだけど、ここでも簡単に話をしておく。

測度は、シグマ集合族の要素（つまり標本空間の部分集合）の大きさを測る関数です。絶対的な大きさはどうでも良くて相対的な大きさだけが重要。とにかく、部分集合の大きさを測る物、と思っておけば良い。

例えば、サイコロを一回振る、という例なら、サイコロの目の数、というのは立派な測度になります。普通一般の測度は P じゃなくて μ で書くのでそれに習うと、

- $\mu(\{1,2,3,4,5,6\}) = 6$
- $\mu(\{\text{偶数の目}\}) = \mu(\{2,4,6\}) = 3$
- $\mu(\{\text{4以上の目}\}) = 3$

こんな感じで要素の数を数えるような物が測度です。

ただ連続な場合は数えるというとよくわからないですが、だいたいは連続空間の中で要素が広がっている長さで良い。二次元なら面積で良い。実際そんなような測度には、ルベーグ＝スティルチェス測度という名前もついている。

で、測度と確率測度の違いは、確率測度は 1 で規格化されてるもの、というだけ。全集合の測度が 1 になるような測度、それが確率測度です。

とにかく、部分集合の大きさみたいなのを測る関数、というイメージを持っておくのが大切。

入門的な測度論的確率空間について

以上の 3 つで、入門的な測度論的確率空間の定義が出来た事になる。

- なんかの集合
- その部分集合族
- 部分集合の大きさを測る関数

の 3 つで確率空間となる訳です。

理論的にはもちろんこんないい加減な定義じゃだめで、教科書にはもっと真面目な定義がある訳ですが。

この部分集合族はどういう性質を持たなくてはいけないのか、その性質を持っているとどういう事がそこから証明出来るのか、というのを調べていく事で、シグマ集合族という物の理解が深まっていきます。

そして測度というのもどういう性質を持っていないといけないのか、その性質とシグマ集合族の性質から何が証明していけるか、という事を延々と見ていく。

これが測度論的な確率論の入門となります。

ただ、こういう話を突き詰めても、あんまり連続の実数の話、例えばガウス分布とかが出てきません。こういった物は確率変数という物を持ち出さないといけないのですが、これが測度論の入門的なところでは教科書的にあんまり入りきらないので、上記の基本的な確率空間周りの理解を深めたあたりで終わってしまう。

でも機械学習ではだいたい実数上の連続分布を gradient descent で近似していくので、確率変数の話を本格的に勉強してないといろいろ困る。

という事で、測度論の本の次の本、実解析の教科書を読む事になります。

確率変数による確率空間の定義

さて、上で定義した、入門的な測度論的な確率空間は、昨今では Deep Learning 系の論文ではほとんど使われていません。たまにあるけど。

最近はトリプレットの最後は確率測度では無くて、Random Variable、つまり確率変数で定義されています。

ここからがこの文書の本題。

確率変数もまた、確率空間と同様に数学のレベルに応じて何段階か定義があるところなので、独立した章で扱います（予定）。以下で出てくる可測関数とは何かについても確率変数の章で詳細を説明する事にしますが、ここでも簡単に説明しておきます。

まず、確率変数の大雑把な定義から。 X が確率変数であるとは、以下の 2 つの条件を満たす関数の事です。

1. 標本空間から R への関数
2. 可測関数

確率変数は変数という名前がついていながら、関数です。この確率変数とは何か、及びその周辺の事用語を一通り理解するのが「実解析の言葉を理解する」という表現で私が言っている事の実態で、2018 年現在で機械学習をやるプログラマが必要な数学という物の正体だと思っています。

普通、確率変数は大文字の X とか Y で表されます。

まず条件 1 から、確率変数 X は以下のように書けます。

図 3: images/p_space/0000.png

で、これが可測関数であるとは、「R 上でのボレル集合族の元の X による逆像が、Ω 上でのボレル集合族の元となっている」関数の事です。

さて、ボレル集合族が出てきました。ボレル集合族の詳細は「シグマとボレル集合族と測度」の章でやる事にします。

言葉は難しいのですが、ようするに R の上のシグマ集合族です。 R の上の、いろいろな開区間を集めた集合族、と思っておきましょう。

さて、上の定義の「逆像」というのを考える為には、X で特定の範囲が写されているような状況を考えます。以下みたいな感じ。

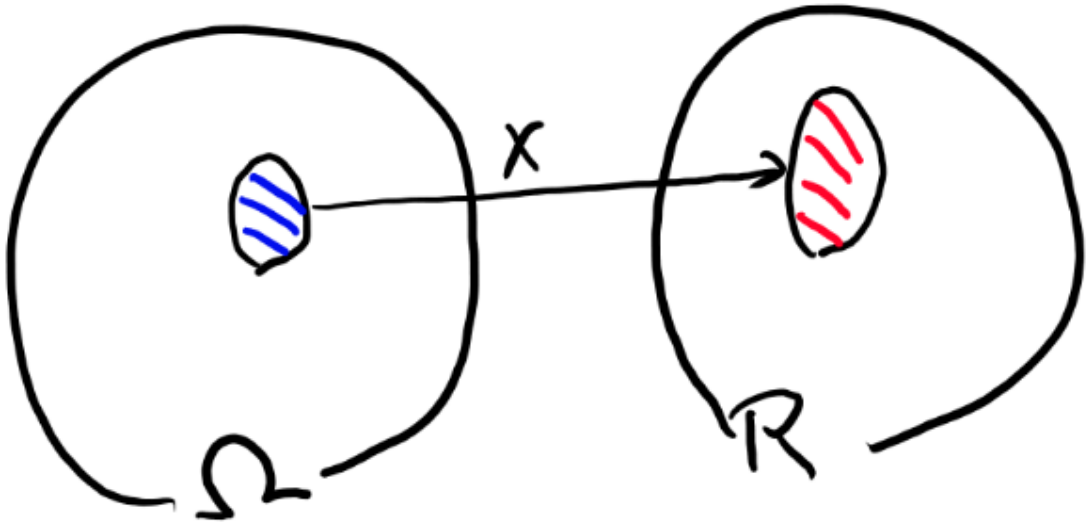

図 4: images/p_space/0001.png

この青い範囲の物を X で全部写すと、R 上の何かしらの部分集合になる訳です。それを赤で書いてます。

さて、可測関数というのは、先に赤い方を、シグマ集合族の元になるように選びます。この時、ここに X で来るような Ω 側を全部集めた物が、Ω 側でシグマ集合族になっている、という事です。

で、右側の R でのシグマ集合族としては、ボレル集合族が使われます。ボレル集合族は実数上の自然なシグマ集合族として使われるもので、だいたい開区間を集めた物です（その拡張だけど）

だから感じとしては、写った先である R 側のシグマ集合族の元を好きに選んでも、それを X が来る場所に戻すと、もとの方のシグマ集合族に入っている、という事です。

さて、X は可測関数なので、R の上のボレル集合族の元、つまり右側の赤いやつには、必ず Ω の方の青いやつが対応します。

なので、青い方に確率測度が定義されていれば、R の赤いヤツを指定するとそれのもととなる青い奴の測度を一意に求める事が出来ます。

数式で言うと、逆像を X^{-1} と書くと、

$$P \circ X^{-1}$$

という関数は、R 上のボレル集合族の測度となる。

これは測度論の入門書とかだと分布、と呼ばれていて、機械学習でもあんまり厳密な話をしない人は分布、となんとなく使ってる気がする。だが、実解析とかの方に行くと分布って累積分布関数の事を指すようになるので、最初から分布とは言わない方が良いと思う。分布＝累積分布関数と脳に負荷をかけずに解釈出来るように慣れておかないと、実解析の教科書読む時に本当に辛い思いをする事になるので...

さて、実解析とかでは、これは X の law と呼ばれたりして、$\mathcal{L}(X)$ とか書く。という事でこのシリーズでも law と呼ぶ事にする。 law の日本語は知らない。まぁここまで来たらもう日本語はいいでしょう。

呼び方はいいとして、この測度というのは、概念的には P が使われているのだけど、一方で実数のボレル集合族上で測度が定義されていれば、それがもともとは P から出来ている、という事なんて知らなくても良い。

実際、law を一つ決めると、それに対応した P は一意に決まったはず（TODO: あとで厳密な条件を調べる）

そういう訳で、X と law を指定する事と P を指定する事は等価なので、X と law を指定する確率空間の定式化が可能となる。これが機械学習で一番使われている、確率変数による確率空間の定式化だと思う。

確率変数と law をもとにしたトリプレット

確率変数と law を決めると確率空間が定義出来る、という事は、トリプレットとしては以下のような物を考えている、という事になります。

$$(\mathbb{R}, \mathcal{B}, \mu)$$

Ω のかわりに $\mathbb{R}$ になる訳ですね。これは普通にガウス分布とか考える時にみんな暗黙のうちに考えている標本空間なのですが、それは実は確率変数をもとにした確率空間の定義で考えていた、という事なのです。

で、その $\mathbb{R}$ の部分集合としてボレル集合族である $\mathcal{B}$ を確率を測る対象とする訳です。 $\mathbb{R}$ もその上の $\mathcal{B}$ も、問題によらず世界に一つなので、特にことわりを入れなければこれについて考えている、というのが機械学習の議論のお約束だと思います（ただし論文では明示するのが普通）。

で、唯一問題によるのが最後の μ。確率変数の law は確率測度なので入門的な定式化と揃えるなら P と書く所なのに、なぜか μ と書く事が多いですね。なんで P じゃないんでしょうか？law だという事を明示的に表す為ですかね。

ここで注目して欲しいのは、このトリプレットの中には、肝心の確率変数 X が居ない、という所です。 law を決めてしまえばボレル集合族に対する測度は定義出来るので、もう X も P も必要無くなってしまう。

また、確率変数とその law を指定すれば、確率空間は決まってしまう、という事にも注意が必要です。だからこの 2 つだけ決めればその上での議論は出来るし、そういう風に始めるのは Deep Learning では一般的です。

確率変数と law による確率空間の定義を理解している必要性

さて、確率空間として何を考えているのか、というのは、どの位理解している必要があるのでしょうか？例えば「確率変数 X の分布 p を推計する」みたいな話をしている人が居た時、これをどのくらい正確に理解している必要があるのか、という話だと思います。 (なお上記の表現は、ちゃんと分かっている人には p が何を指しているのかはかなり曖昧です)。

自分の見解としては、登場人物か確率変数と p だけなら必要無いと思います。ただ、この p を等価な関数空間で探索する場合には、この辺の事情を正確に把握していないと辛い。具体的には確率密度の関数空間の探索に置き換える（WGAN）とか、確率変数の和がどういう law に従うのか、という事を議論する必要がある時、などです。

また、極限定理などの収束を議論する必要がある時も、law の話になっていくので、この辺の構造をちゃんと理解している必要があります。機械学習の言葉で言うなら、gradient descent とかの optmize 周辺の理論を議論したいとなると、ちゃんと確率空間を意識して読んでいく必要が出てきます。

分布による定義

機械学習的には確率変数と law による定義でだいたい事足りますが、もっと関数空間を本格的に分析しよう、と思ったら、もう一段抽象的な定義をするのが普通のようです。これは関数解析で確率論の話をする時によく用いられる確率空間の定義になります。

関数解析的には、確率なんて物を持ち出さなくても、確率に関わる関数空間の構造について、かなりの議論が出来ます。この場合、中心になるのは non-decrease な関数で、最大値が 1 のもの、みたいなすごく一般的な定義で分布と言われる物が最初に決まる。

ジャンプがどういう物が許されるか、とかすごく細かい話が続くのだけど、基本的にはこのレベルでは確率的な要素は特に無い。

ただ、この分布でかなりの部分の確率論の話が出来てしまう。

確率変数同士の距離とか、距離自身が確率分布する場合を扱おうとするとこちらの定義が主流となる。だが、自分の知る限り、機械学習ではこちらの定式化が使われる事はあまり無い（私は見た事無い）。

シグマ集合族ってなんなのさ

測度論の入門的な確率空間ではシグマ集合族が重要な位置を占めます。そして機械学習とか実解析ではボレル集合族が重要な位置を占めるようになります。

このシリーズとしては本題では無いのだけど、やはりこの辺の事は知っておく必要があるので、この章ではこの 2 つにまつわる話をしていきたい。

ボレル集合族でググると測度論で苦しむ若者たちのメモのような物をたくさん読む事が出来ます。ここでさらにもう一つ似たようなメモを増やすよりは、このシリーズではもっといい加減な雑談をしていきたい。

感覚的な話とか位相との関係とかを雑に話していきます。

シグマ集合族とボレル集合族の定義から

厳密な定義としては

ルベーグ積分から確率論 (共立講座 21 世紀の数学) https://www.amazon.co.jp/dp/4320015622/

の p15 あたりからを見てもらうとして、ここでは雑な話を。

シグマ集合族とは、大雑把には

- その要素の not
- その要素の intersection

もまたシグマ集合族に属すような集合族の事です。

無限回の intersection も許す所が理論的に難しい所だけど、感覚的にはある部分集合の、否定をとっても intersection をとってもその集合族に属す、と思っておけば機械学習的には十分。

で、ボレル集合族は開集合を含む最小のシグマ集合族の事、と定義される。現実的には離散的な集合の話をする時はシグマ集合族、$\mathbb{R}$ の上の話をする時はボレル集合族、と思っておけばだいたい OK。

対象となる集合	事象族として選ばれる集合族
離散集合 Ω	シグマ集合族
$\mathbb{R}$	ボレル集合族

ボレル集合族は実数上の自然なシグマ集合族として重要で、これは確率変数が実数への可測関数として定義される事から、確率変数中心の定式化ではボレル集合族が主役となります。

機械学習屋としては、定義よりも、それが自分が今取り組んでる実際の問題の、何に対応しているかを知る事が大切です。という事で具体的には何か、という話をしていきたい。

まずは大雑把に、サイコロの例でシグマ集合族と確率測度を考える

サイコロを一回振った時の、偶数の目が出る、という事象と、4 以上の目が出る、という事象について考えよう。

図示すると以下のようになる。

この時、シグマ集合族というのは赤とか青の丸で書いた物だ。

厳密に言えば赤い丸が一つの要素、青い丸がもうひとつの要素となる。文字としては F で表されるものだ。で、このいろいろな F を全部集めた物が $\mathcal{F}$ となる。

つまり以下のような式の話をしている。

$$F \in \mathcal{F}$$

さて、一つの F としては、例えば偶数の目、というのは、

$$\{2, 4, 6\}$$

という集合を表す。これはいつも Ω の部分集合だ。

測度というのはこの F の大きさを表す物だ。確率測度は Ω 全体を測ると 1 になる物、という決まりがあるが、確率測度じゃないただの測度ならその辺には特に決まりが無い。

だから絶対的な大きさにはあまり意味が無くて、相対的な大きさにしか意味が無い。

具体的に考えよう。普通確率じゃない測度は μ で表す事が多い気がするので、ここでもそうしよう。

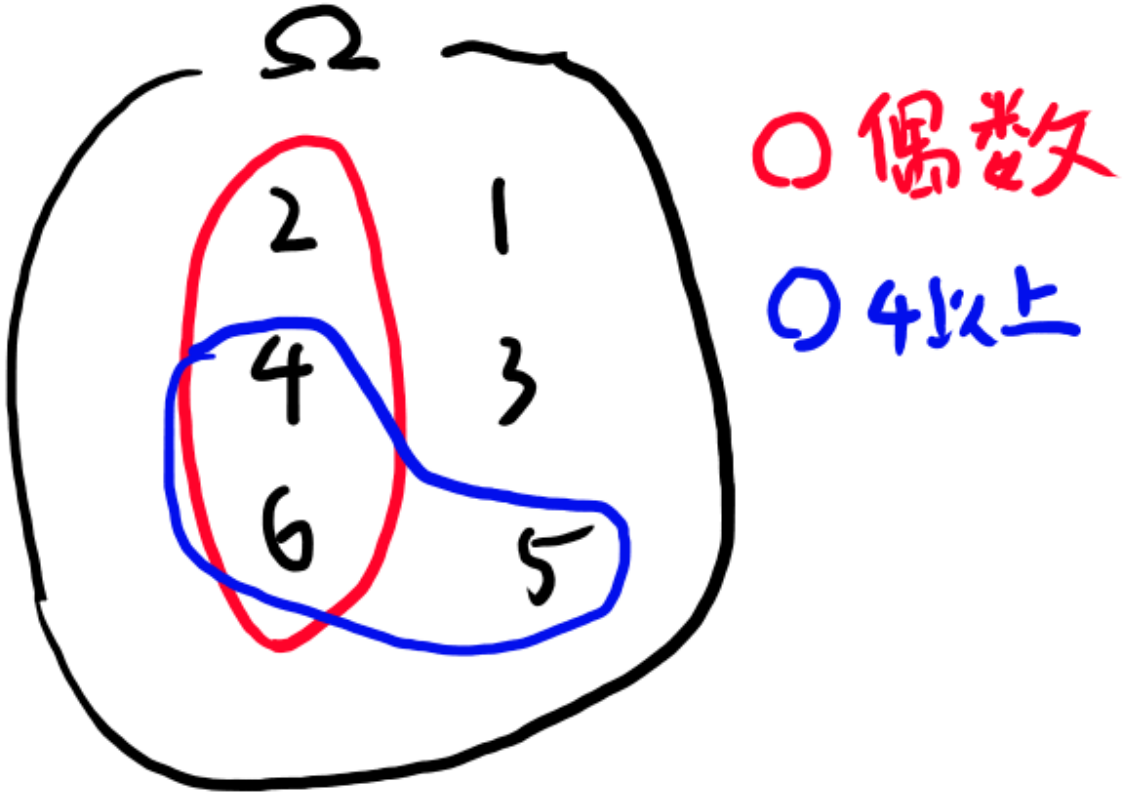

図 5: images/borel/0000.png

今回は要素の数を測度としよう。

つまり、

$$\mu(\text{偶数の目}) = \mu(\{2,4,6\}) = 3$$

となる。

機械学習屋としては、測度が大きさを測る関数で、$\mathcal{F}$ がその大きさを測る対象だ、という事をしっかり覚えておく事が大切。

シグマ集合族の定義の、それぞれの意味を考える

さて、シグマ集合族の定義とは、だいたい任意の F の否定も $\mathcal{F}$ の要素で、しかも、intersection も $\mathcal{F}$ に入る、というのがおおまかな物だ、と言った。

という事で、それぞれの定義の意味を実例を元に考えてみよう。

否定が含まれるとは

F の否定もまた $\mathcal{F}$ に含まれる、という事の意味を、先程のサイコロの例で考えてみよう。

まず、偶数の目、という部分集合を考える。その否定というのは以下になる。

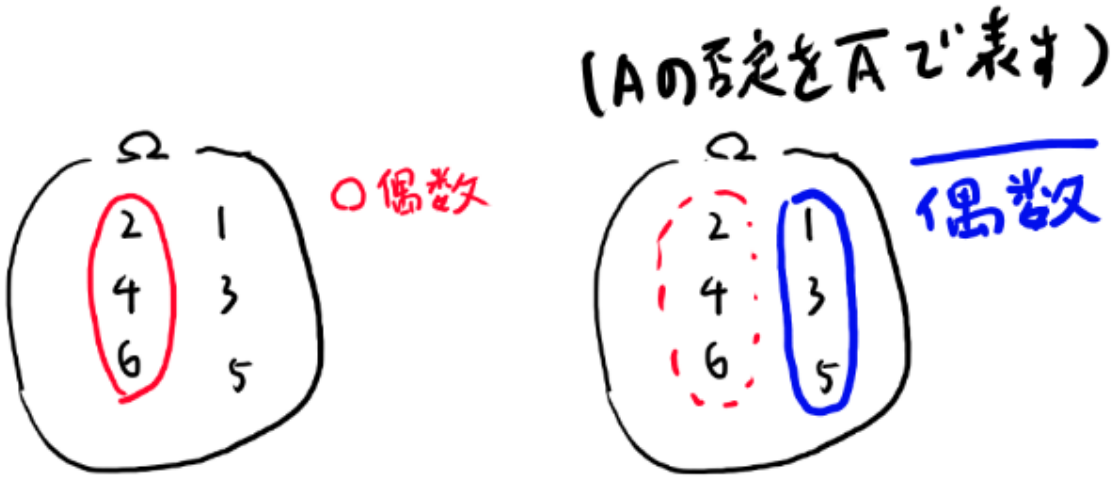

図 6: images/borel/0001.png

つまり $\{2,4,6\} \in \mathcal{F}$ なら、$\{1,3,5\} \in \mathcal{F}$ でも無くてはいけない、という事だ。

なんでこれが大切かといえば、確率というと

$$P(A) = 1 - P(\bar{A})$$

とか、そういう関係式が成り立って欲しい訳だが、この時右辺がいつも成立する為には、右辺も大きさを測る対象である必要がある。つまり、$\mathcal{F}$ の中に入っていないと困る、という事だ。

intersect が含まれる事

サイコロを一回振った時の、偶数の目が出る、という事象と、4 以上の目が出る、という事象について考えよう。

図示すると以下のようになる。

$$\{\text{偶数の目} \cap 4\text{以上の目}\} \in \mathcal{F}$$

とは、この場合は

$$\{4,6\} \in \mathcal{F}$$

という意味となる。 この intersect がまた $\mathcal{F}$ に入る、というのは、先程の否定が入る事と合わせると、良くあるような確率の対象を表す事が出来る訳です。

例えば以下の緑の網掛けみたいな感じのが表現出来ます。

逆にこういう良くあるようなパターンも確率測度の対象となるような物を全部集めた物、それがシグマ集合族、と思っておいて実用上は OK です。

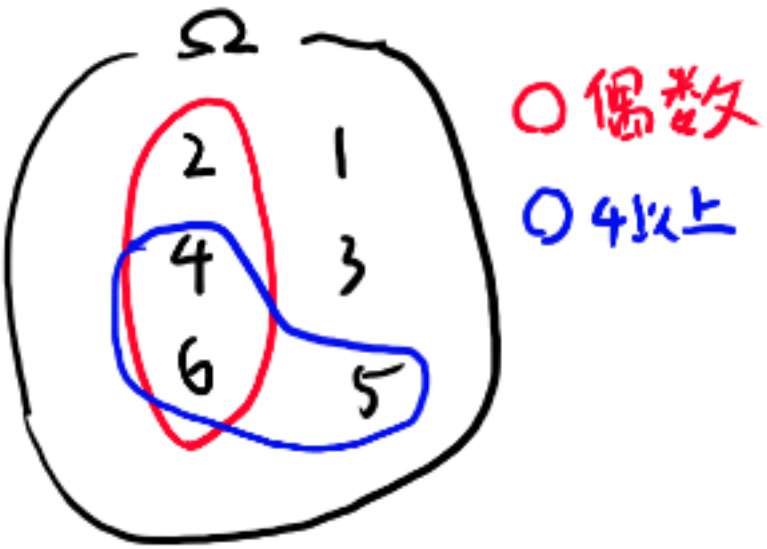

図 7: images/borel/0002.png

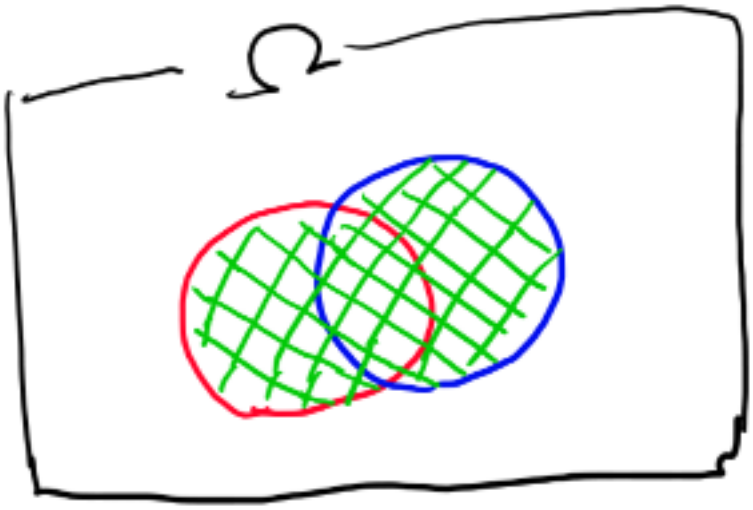

図 8: images/borel/0003.png

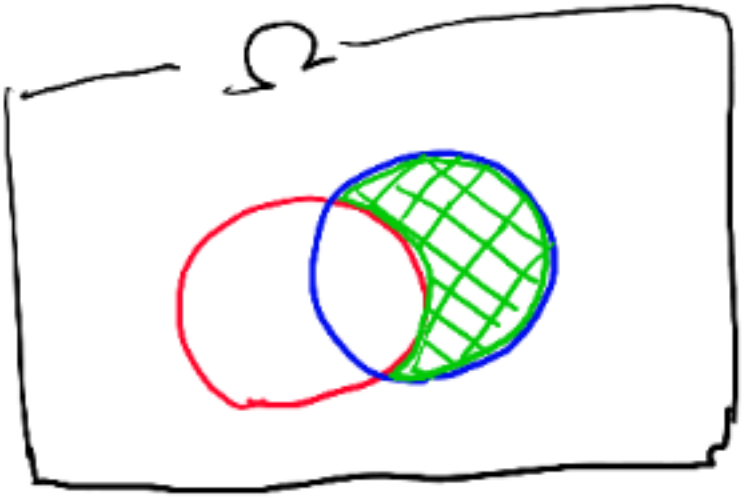

図 9: images/borel/0004.png

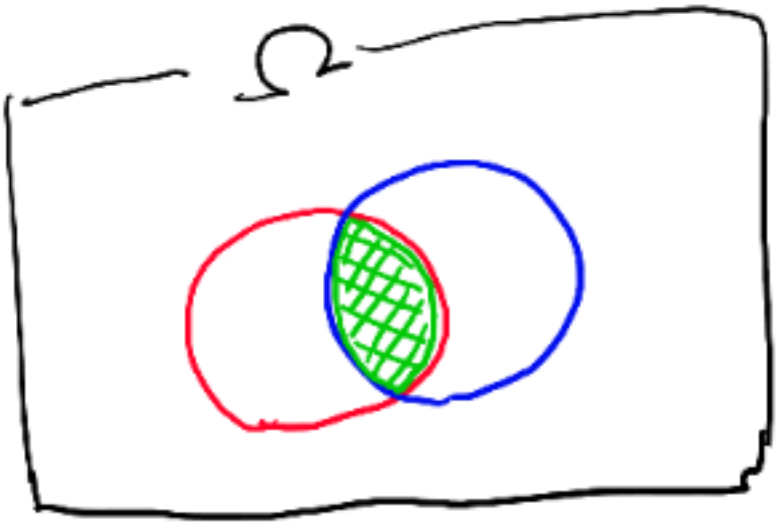

図 10: images/borel/0005.png

シグマ集合族と確率測度

測度というのは、雑な言い方をすれば集合の大きさだ。その集合がどのくらいの範囲を占めているのか、という、集合の大きさを表す。

確率測度は、測度のうち全集合が 1 になるような測度の事だ。

で、シグマ集合族の not と intersection が自身に含まれる、というのは、確率の基本的な公式を満たすのに必要となる。

例えばある事象 A と B があった時に、

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

というような確率論の基本的な公式を議論する為には、$A \cup B$ とか $A \cap B$ が事象、つまり P の定義域である必要がある。

これらが全て P の定義域として閉じてますよ、という要請を追加した開集合のような物がシグマ集合族だ。

そして測度の定義も上のような類の式が成り立つような何か、という事になる。感覚的にはやはり部分集合の大きさを測る物で、分解したら個々の要素の和が全体の和と等しくなる、みたいな感じの性質の物と思っておけば良い。

事象族とシグマ集合族

確率空間を構成する 3 つの文字の一つ、事象族について。

厳密な定義はおいといて、事象族というのがどんな物なのかイメージしておくのは大切だ。特にこれが標本空間の部分集合の集まり、という事はちゃんと理解しておかないと、論文が読めない。

事象というのは、確率測度で大きさが図れる物、という事だ。確率測度は部分集合の大きさを測るもので、もっといえば全集合との大きさの比率を測ることになる（確率測度は定義により全体の測度が 1 なので）。

これはつまり P の定義域になる、

P は集合の大きさを測る物だったので、事象も集合、正確には部分集合となる。 P が対象とするような部分集合を全部集めた物、それが $\mathcal{F}$。

それらから好きに要素、A, B を取り出したら、$A \cup B$ とか $\bar{A}$ とか $A \cap B$ とかも $\mathcal{F}$ に含まれる、という事が保証されているだけ。

具体例としては、「サイコロの目が偶数」と「サイコロの目が 4 以上」という 2 つの事象があった時に、この not とか intersection も事象、つまり $\mathcal{F}$ の要素となる。

TODO: 続きはあとで書く

- 可測関数の話
- 開集合とシグマ集合族の比較
- 連続関数と可測関数の比較

確率変数の話

みんな大好き確率変数。

だってそもそもに機械学習って入力 X が与えられた時に、Y の分布を

$$P(Y|X)$$

という条件つき確率として推計する物でしょ？とか言ったりする。 Y の分布ってなんだよ、とか、この Pってなんだよ、という事はあんまり考えないでこういう物言いをしたりしがち。

そんなふうに機械学習においては普遍的に登場すると言っても良いくらい頻繁に使われる確率変数だけど、ここの定義はかなりへんてこでちゃんと勉強してないと「何を言っているんだ、お前は」という感じになりがち。

しかも英単語としては random variable。 ランダムな数って事か。なんだ、わかりそうじゃん、という事で、良く分かってない人も結構いい加減に使う。

ただ、理論的な話をする時には、確率変数、確率密度、分布、law の区別をちゃんとしてないと何の話をしているかさっぱりわからない。最近の Deep Learning の論文ではこれらの空間の間を行ったり来たりして議論するので、random variable をランダムな数でしょ？と思ってると壊滅的な事態となる。

そんな訳でハマりがちな確率変数について、この章では扱っていきたい。

古典的な定義

確率空間と同様に、確率変数も古典的な定義と測度論的な定義がある。で、古典的な定義が意味がわからないのも同様。

良くある確率変数の定義は、