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。

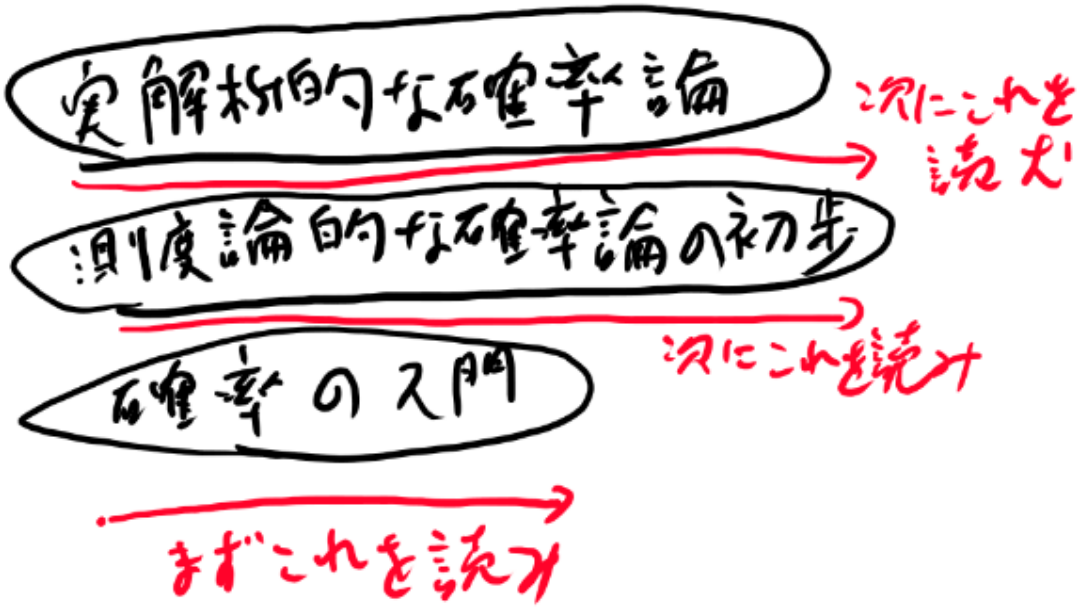

図 1: imgs/intro/0000.png

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

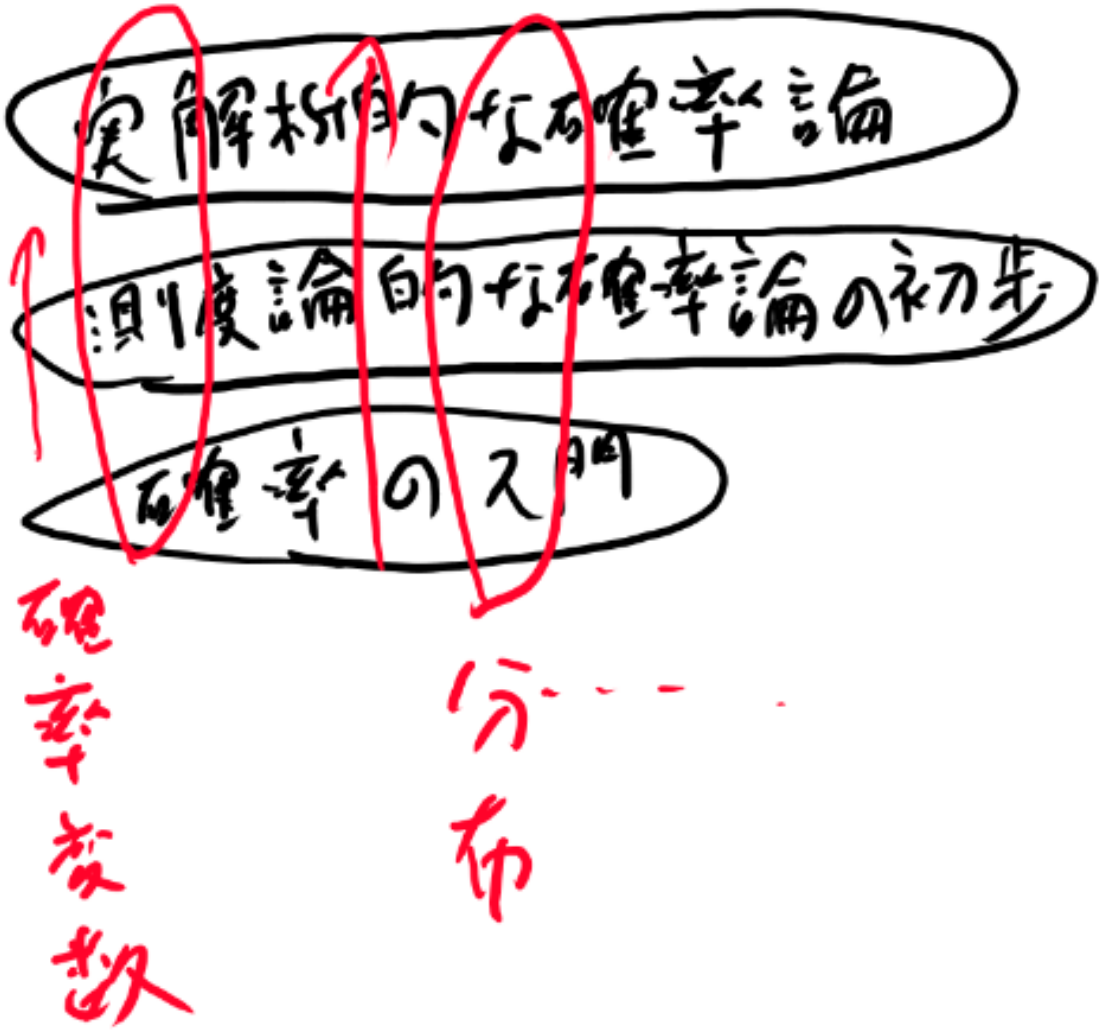

図 2: imgs/intro/0001.png

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のだが、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）ので、ここで 3 や 4 の話をしたい。

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。というのはボレル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。

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が、確率測度の事だ。

まずこの定義を見ていく事から始めよう。

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

$$\Omega$$

で表すので、

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

と書いておこう。

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。

事象族

さて、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。普通は「偶数の目が出る」などが事象の例となる。

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。

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 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。

確率測度の事なんだけど、測度が無い状態ではぼやっと定義される。なので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。

だが、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 B を引数として、その確率を返す関数だ。

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、それぞれ何を指しているかを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

なお、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、理論的には機械学習的な事は全部説明出来ると思う。説明には、本当は測度論とかは一切要らない。

ただ、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。

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「測度論」という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のだが、それでも一応この辺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

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知っている意味はある。

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よう。

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、だいたい同じ物である。測度論的な確率空間も以下の 3 つの要素からなる。

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

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 だ。

事象族はシグマ集合族であり、P は確率測度となる。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」の章で扱う。

ちょっと前後するが、一旦そちらを読んでから続きを読んでほしい。

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

$$\Omega$$

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、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小だけが重要。

だから例えば、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

$$\mu$$

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、1 で規格化されてる、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。

確率変数による確率空間の定義

さて、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ない。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。

確率変数はまた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）。可測関数と何かについては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、

1. 標本空間から R への関数
2. 可測関数

な物です。

まず 1 から、以下のように書けます。

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、

$$\Omega$$

上でのボレル集合族の元となっている」関数の事です。

ややこしいですね。もう少し説明します。

まず、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 3: imgs/p_space/0000.png

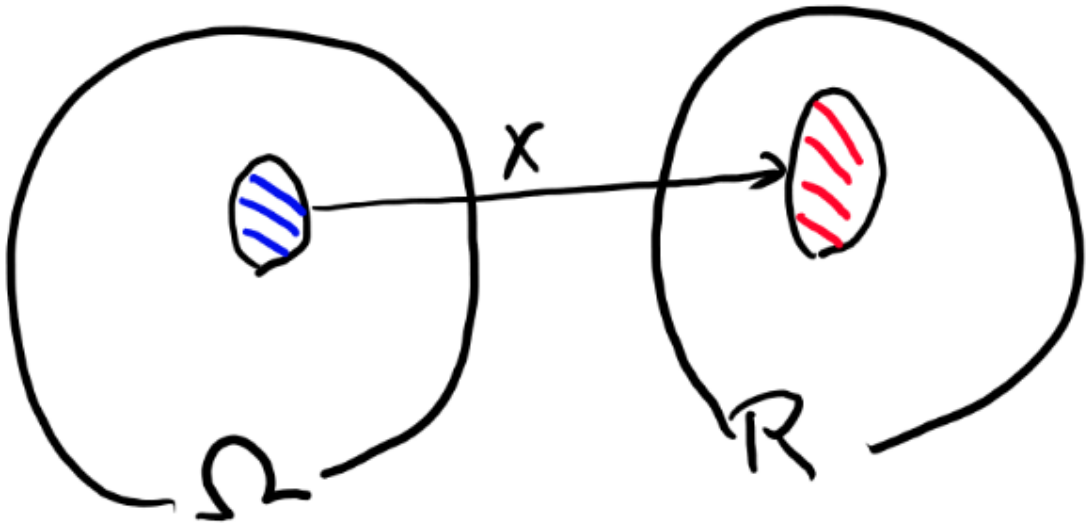

図 4: imgs/p_space/0001.png

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

$$\Omega$$

側を全部集めた物が、

$$\Omega$$

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

$$\Omega$$

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。

分布による定義

関数解析的には確率なんて物を持ち出さなくても、かなりの議論が出来る。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）。

ボレル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る。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る。この 2 つは似た物なのでここで話をしていきたい。

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る。

ここでは感覚的な話とか位相との関係とかを雑に話したい。

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だ。

無限回の intersection も許す所が理論的に難しい所だが、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だ。

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。

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る。

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だ。という事で具体的には何か、という話をしてみよう。

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。

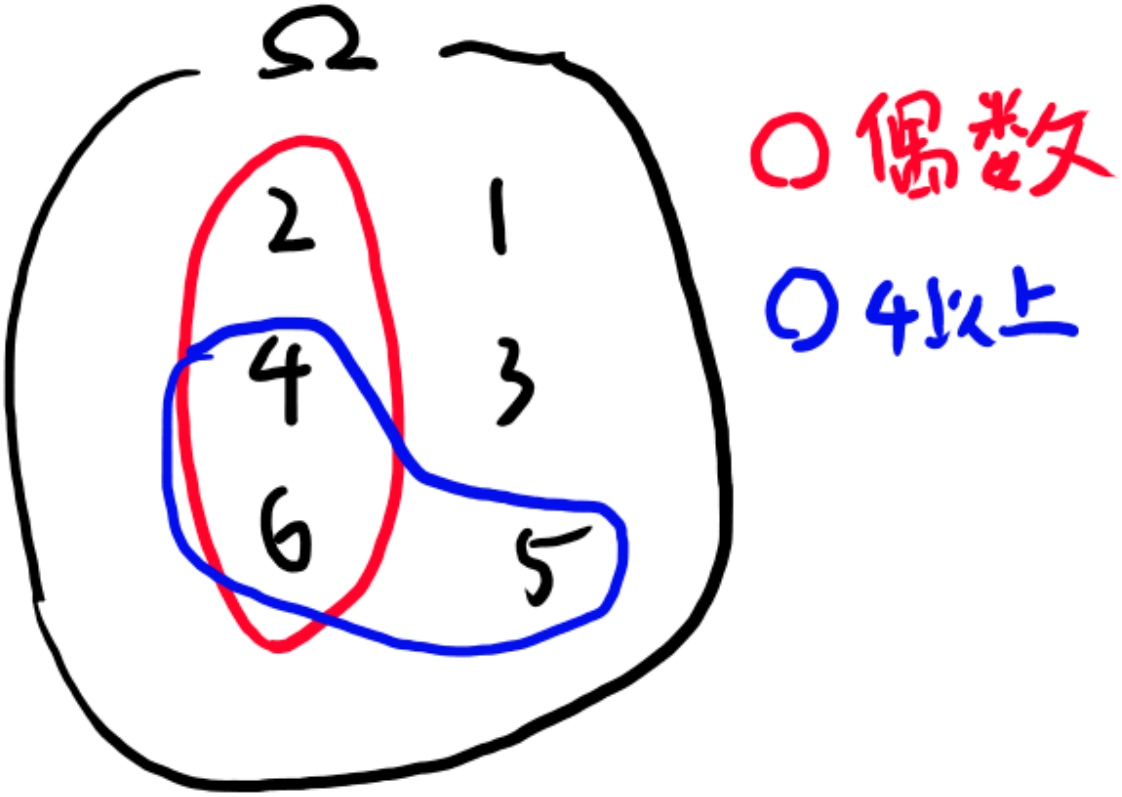

図 5: imgs/borel/0000.png

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

$$\Omega$$

の部分集合だ。

測度というのはこの B の大きさを表す物だ。確率測度は

$$\Omega$$

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

$$\mu$$

で表す事が多い気がするので、ここでもそうしよう。

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、

$$\mathcal{B}$$

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。

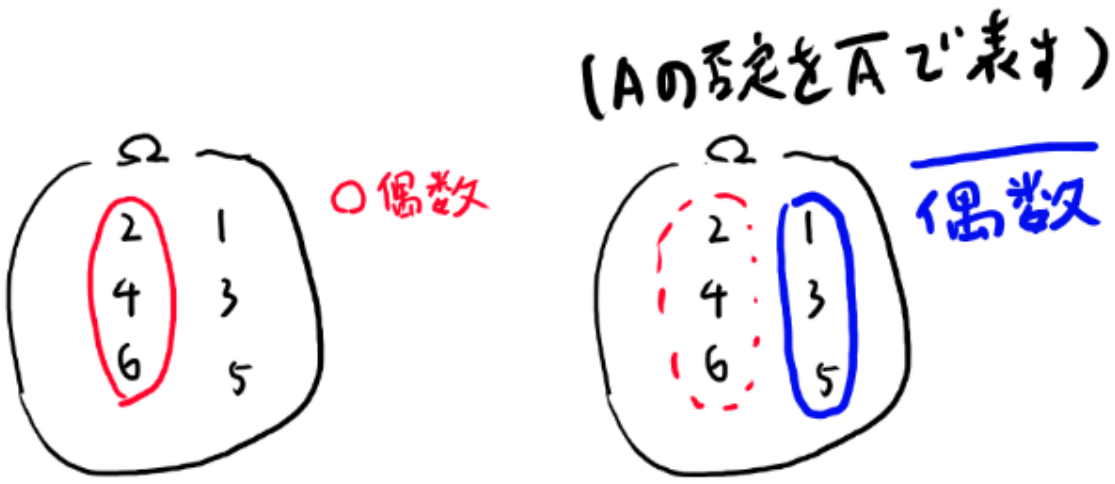

図 6: imgs/borel/0001.png

つまり

$$\{2,4,6\} \in \mathcal{B}$$

なら、

$$\{1,3,5\} \in \mathcal{B}$$

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、

$$\mathcal{B}$$

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。

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

とは、この場合は

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

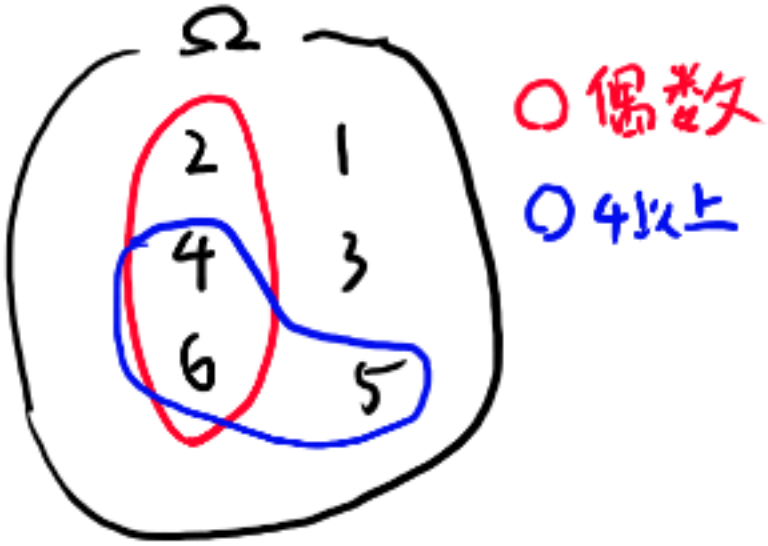

図 7: imgs/borel/0002.png

という意味となる。 この intersect がまた

$$\mathcal{B}$$

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。

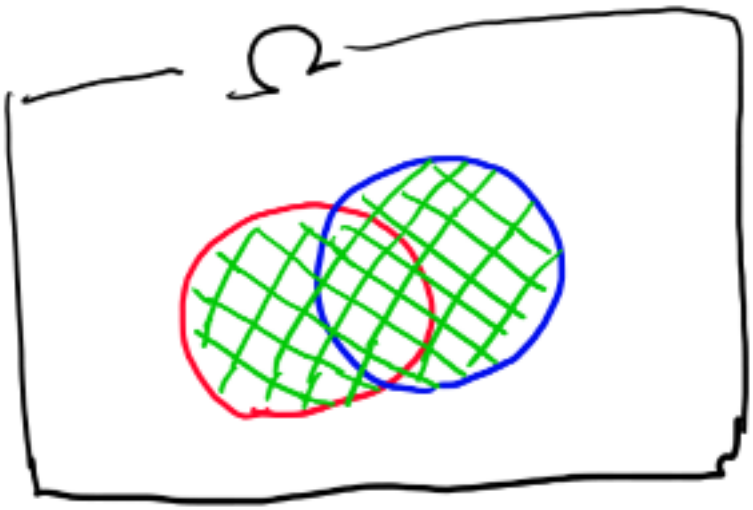

図 8: imgs/borel/0003.png

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。

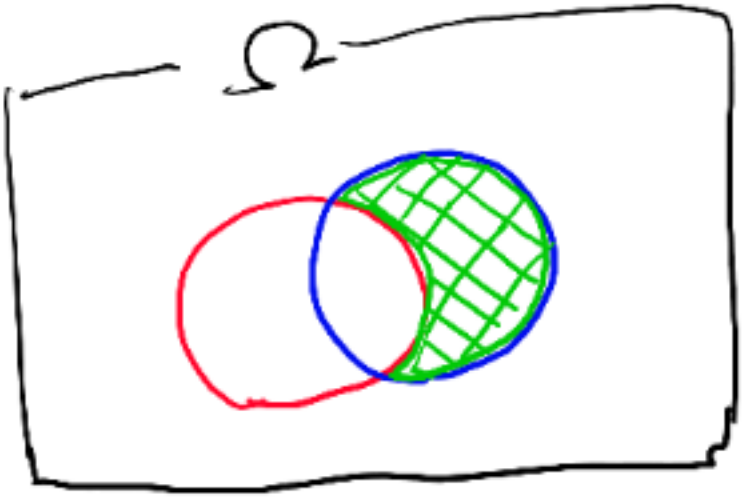

図 9: imgs/borel/0004.png

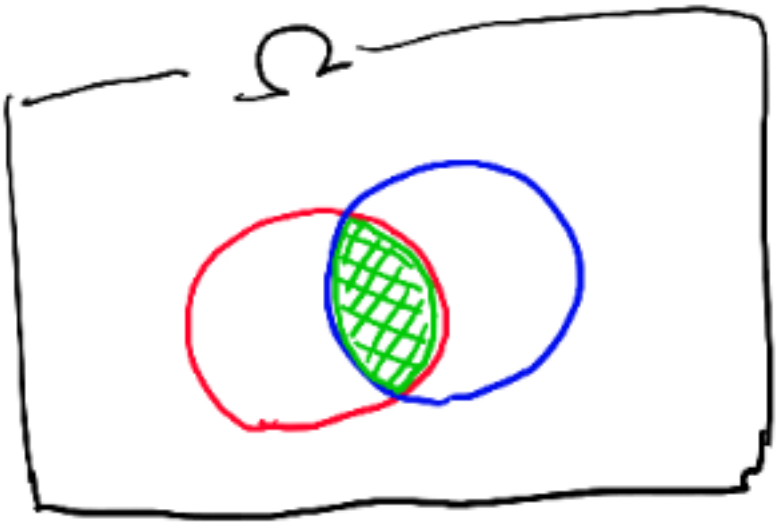

図 10: imgs/borel/0005.png

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。