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。

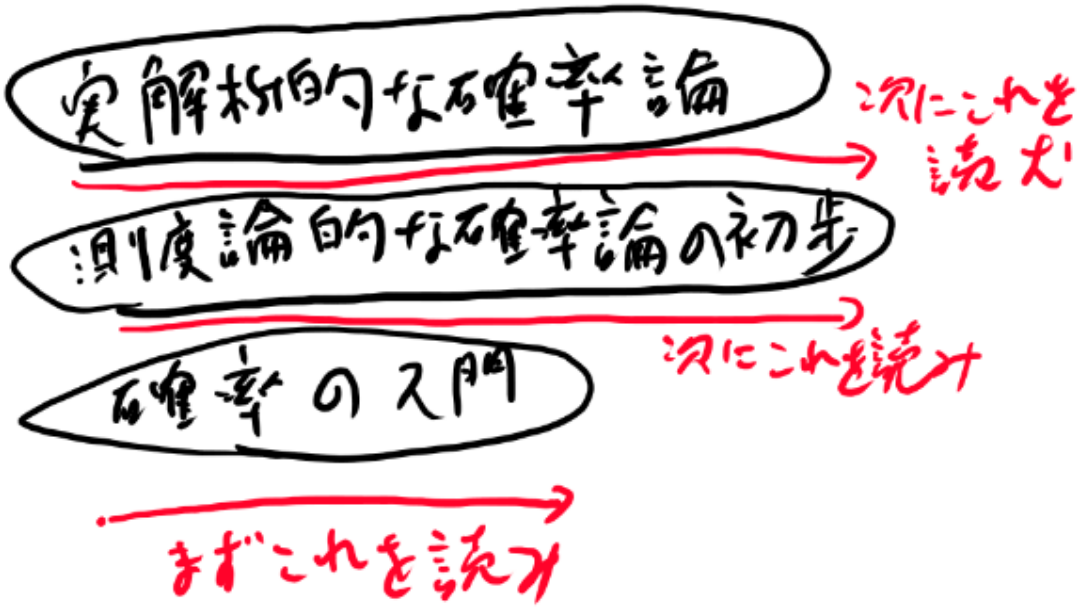

図 1: imgs/intro/0000.png

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る。

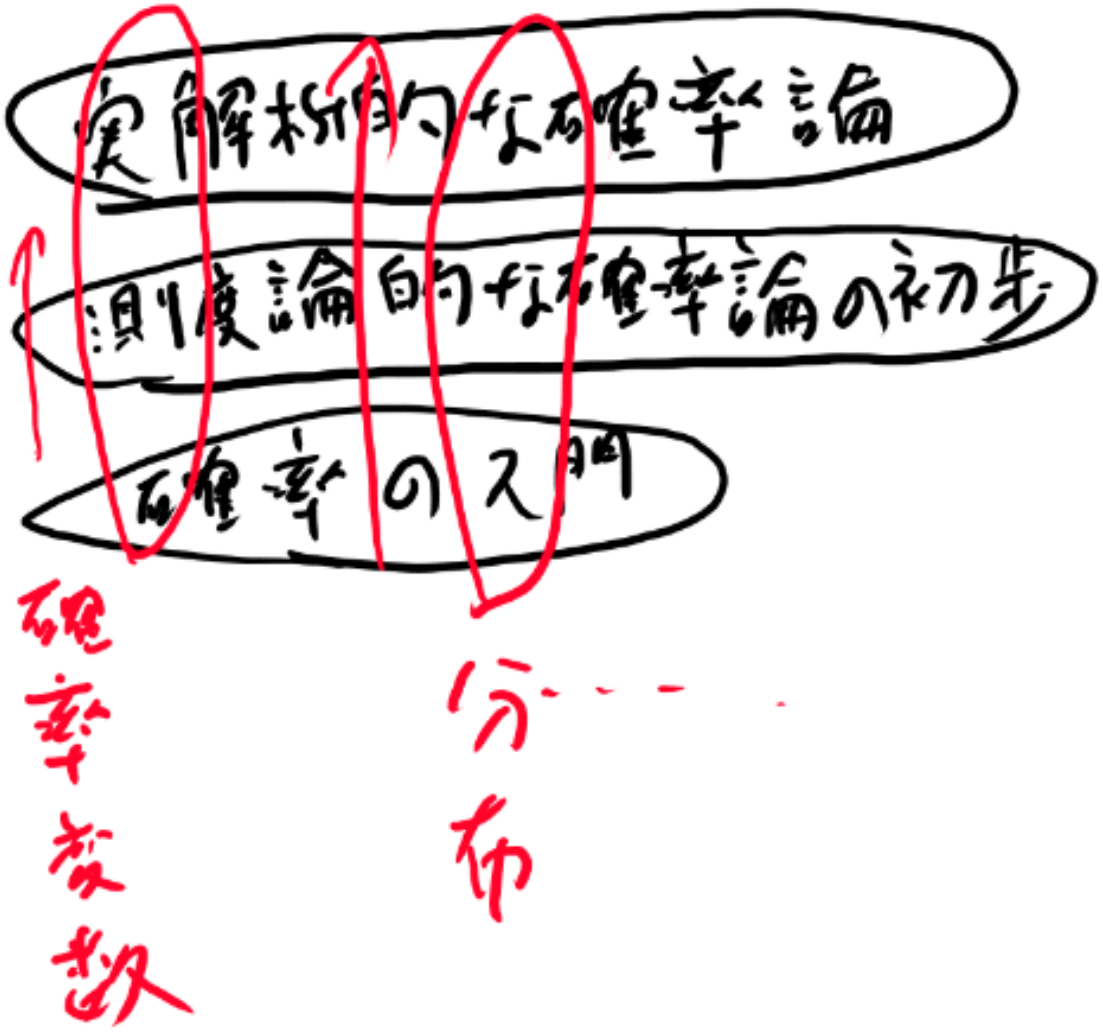

図 2: imgs/intro/0001.png

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、

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

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だ。この場合は

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

となる。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だ。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する。

事象族

さて、良くわからなくなるのが事象族だ。これは本質的にはボレル集合族の事なのに、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だ。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、集合である。例えば「偶数の目が出る」の場合は、

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

となる。

で、この事象を全部集めた物を事象族という。事象が集合なので、集合の集合は集合族と呼ばれるから、事象族と呼ぶ。

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 はボレル集合族だね。

花文字というのは下みたいな文字です。

$$\mathcal{B}$$

で、その事象族の要素となる事象は、普通の大文字 B で書く。

$$B \in \mathcal{B}$$

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、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が、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。確率測度の事だが、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ので調べない。どうせいい加減なのでどうでもいいんです。

だが、この P は割と具体的なので、定義は謎でも、感覚的には何なのかはわかりやすい。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、ボレル集合族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で、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。

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。

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

ちゃんと定義は出来てないが、それぞれ何かをちゃんと識別出来ておく必要はある。ぶっちゃけ古典的なこれらの定義をちゃんと理解しておけば、機械学習的な事は全部説明出来ると思う。 測度論とかは一切要らない。

ただ、誰もそんな事はしてくれないので、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。

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」
で、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、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ので、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。

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。

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

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。

事象族はボレル集合族であり、P は確率測度となる。このボレル集合族と測度は、このシリーズの中心的なトピックとなるので、次の章、「ボレル集合族

事象族とボレル集合族

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

$$\mathcal{B}$$

。

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

$$\mathcal{B}$$

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

$$\mathcal{B}$$

の要素となる。

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ボレル集合族が重要な位置を占める。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。

ボレル集合族とシグマ集合族の定義から

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。

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、サイコロの例でボレル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。

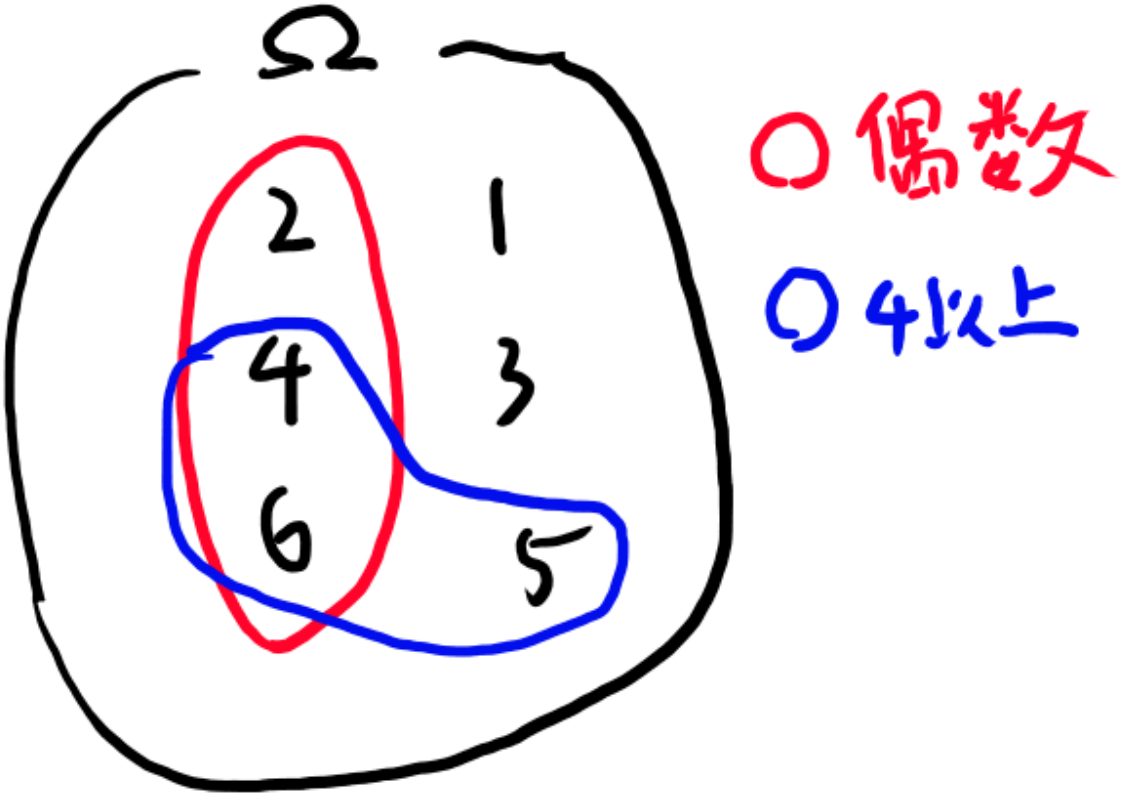

図 3: imgs/borel/0000.png

この時、ボレル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 B で表されるものだ。で、このいろいろな B を全部集めた物が

$$\mathcal{B}$$

となる。

つまり以下のような式の話をしている。

$$B \in \mathcal{B}$$

さて、一つの B 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。

ボレル集合族の定義の、それぞれの意味を考える

さて、ボレル集合族の定義とは、だいたい任意の B の否定も

$$\mathcal{B}$$

の要素で、しかも、intersection も

$$\mathcal{B}$$

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

B の否定もまた

$$\mathcal{B}$$

となる、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。

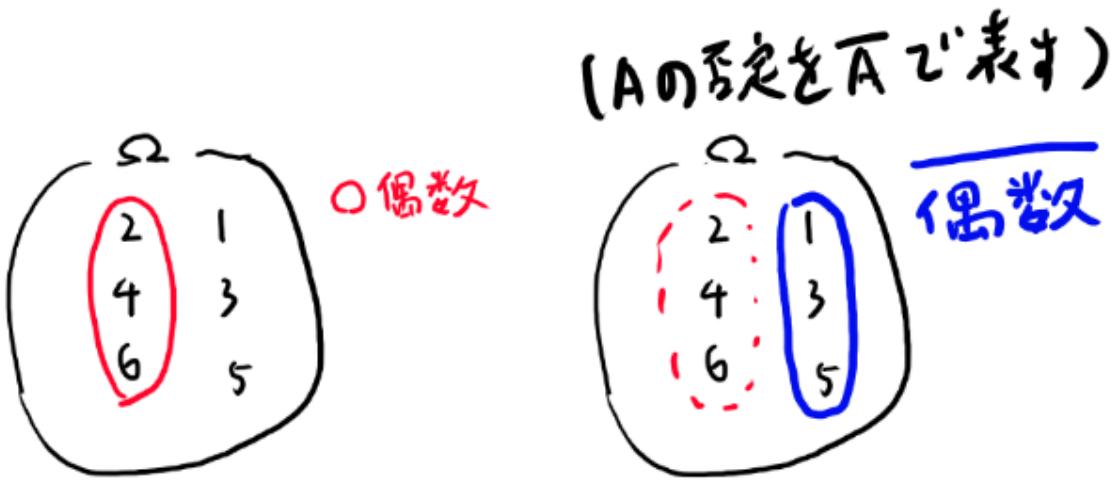

図 4: imgs/borel/0001.png

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

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

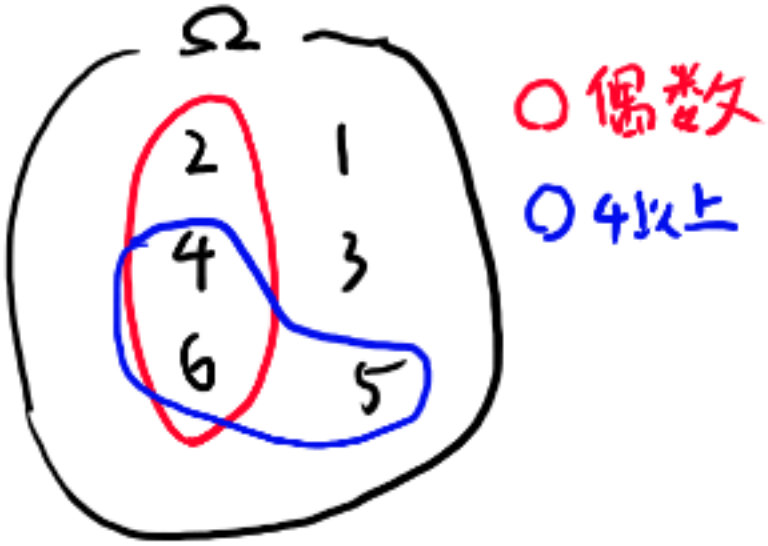

図 5: imgs/borel/0002.png

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。

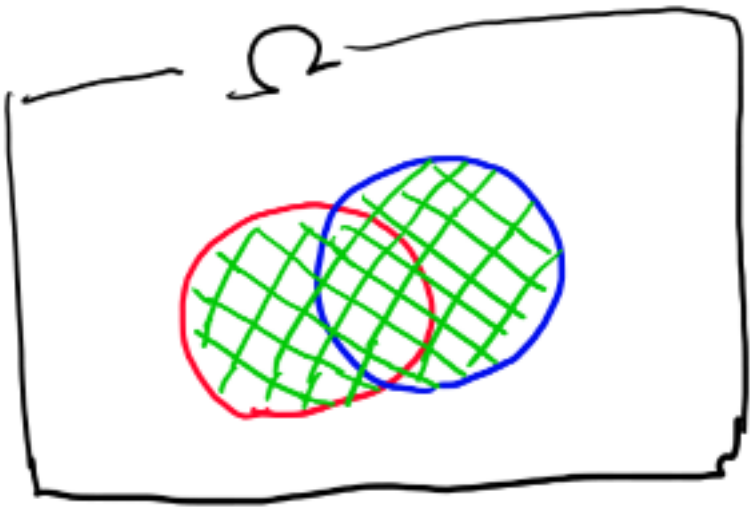

図 6: imgs/borel/0003.png

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、それがボレル集合族、と思っておいて実用上は OK です。

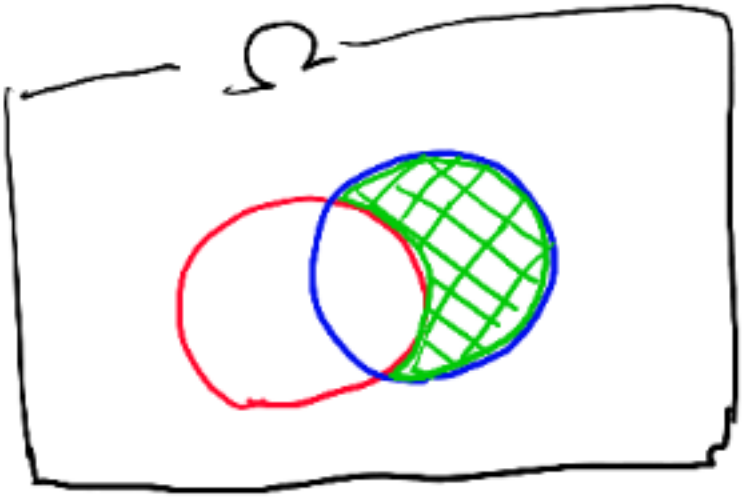

図 7: imgs/borel/0004.png

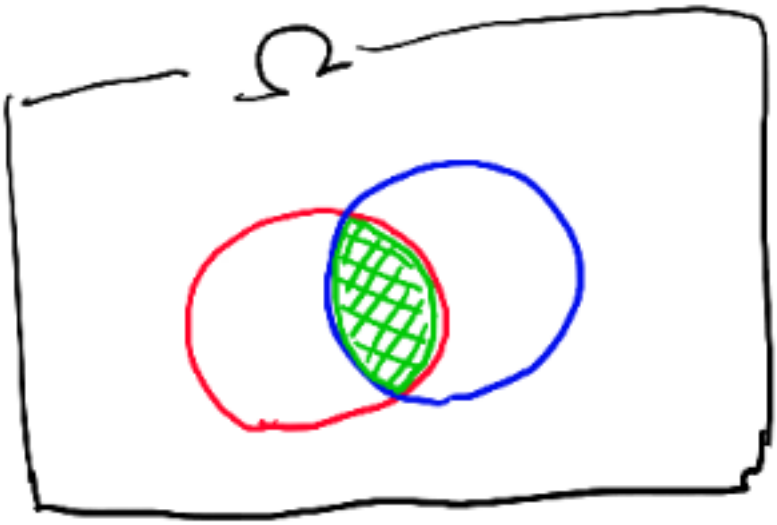

図 8: imgs/borel/0005.png

ボレル集合族と確率測度

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。

確率測度は全集合が 1 になるような測度の事だ。

で、ボレル集合族の 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ボレル集合族だ。

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。

事象族とボレル集合族

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

$$\mathcal{B}$$

。

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

$$\mathcal{B}$$

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

$$\mathcal{B}$$

の要素となる。