

機械学習・ディープラーニングのための
基礎数学講座 確率・統計 Day 2

SkillUP AI

配布物

- 1_slide : スライド教材が入ったフォルダ
 - prob_and_stats_DAY2.pdf : このスライド
- revision_history.txt : 改訂履歴

参考文献（DAY1～DAY3を通して）

- 増補改訂版 語りかける中学数学
 - <https://www.beret.co.jp/books/detail/459>
 - 中学数学が怪しい方へ
- ライブ講義 大学1年生のための数学入門
 - <https://bookclub.kodansha.co.jp/product?item=0000275978>
 - 数学で出てくる記号の意味が怪しい方へ
- 流れるようにわかる統計学
 - <https://www.kadokawa.co.jp/product/301312000211/>
- 確率統計キャンパス・ゼミ 改訂6
 - <https://www.mathema.jp/product/確率統計キャンパス・ゼミ-改訂6/>
- 演習 確率統計キャンパス・ゼミ 改訂4
 - <https://www.mathema.jp/product/演習確率統計キャンパス・ゼミ-改訂4/>
 - 問題集形式

参考文献（DAY1～DAY3を通して）

- まなびのずかんシリーズ統計学の図鑑
 - <https://gihyo.jp/book/2015/978-4-7741-7331-3>
- Pythonで理解する統計解析の基礎
 - <https://gihyo.jp/book/2018/978-4-297-10049-0>
- なるほど統計学
 - <https://www.amazon.co.jp/dp/4875252102>
- 日本統計学会公式認定 統計検定 3級・4級 公式問題集
 - <https://jitsumu.hondana.jp/book/b496705.html>
 - 問題集形式
- データサイエンスのための統計学入門 第2版
 - <https://www.oreilly.co.jp/books/9784873119267/>
- 最短コースでわかる ディープラーニングの数学
 - <https://www.nikkeibp.co.jp/atclpubmkt/book/19/273470/>

本講座の全体の内容

Day 1 : 記述統計学の基礎

- ・ 内容 : 統計量・可視化
- ・ 修了演習 : データの前処理技術 (正規化と標準化) ・ 箱ひげ図を用いた外れ値検出

Day 2 : 確率

- ・ 内容 : 確率の基礎・条件付き確率・ベイズの定理・独立
- ・ 修了演習 : ナイーブベイズによるスパムメール判定

Day 3 : 確率分布

- ・ 内容 : 離散型/連続型確率分布
- ・ 修了演習 : ロジスティック回帰

本講座でやること / やらないこと

• やること

- 確率・統計分野の重要な概念・公式
- 各種公式を使った問題演習
- 機械学習 / 深層学習における上記概念・公式の利用方法の概要

• やらないこと

- 紹介する公式等の厳密な証明
- Python等を用いた実装方法
- 機械学習 / 深層学習の各種手法の詳細な説明

- 青字・下線付きは URL リンク付き文字です
 - PDFビューワ上で該当箇所をクリックすると参考ページに遷移することができます
 - 例) [スキルアップAI](https://www.skillupai.com/)
(スキルアップAIのトップページ <https://www.skillupai.com/> へ遷移)

講座に入る前に

- 本講座では機械学習 / 深層学習を学ぶための土台となる内容を学習します
そのため、目的意識を持って学び、アウトプットすることが重要です
- そこで、次の2点を必ず実施しましょう
 1. 事前にスライドに目を通し、予習を行いましょう
 - 漫然と目を通すだけでなく、どの部分を集中して聞くべきか自分の中で決めておきましょう
 2. 各 DAY ごとに振り返り・言語化の時間を取りましょう
 - 振り返り内容
 - この DAY で学んだ内容で参考になったことは？
 - 内容の簡単なサマリ
 - 重要な公式のまとめ など
 - 振り返りの結果は紙やテキストファイルにまとめましょう

Day 2

確率

目次

第1章：確率の基礎知識

第2章：条件付き確率



機械学習において必要な確率の知識を学ぼう

第3章：条件付き確率の発展

- ベイズの定理
- 条件付き独立



機械学習で多用されるベイズの定理と
ナイーブベイズのモデル構築に重要な
条件付き独立を学ぼう

第4章：修了演習

- ナイーブベイズによるスパムメール判定



本日学んだ内容を
機械学習に応用しよう！

第1章

確率の基礎知識

確率の基礎

試行

同じ状態のもとで繰り返すことができ

その結果が偶然によって決まる実験や観測（サイコロを1回投げる）

事象

試行の結果起こる事柄（サイコロを投げて3が出る）

標本空間

ある試行において起こりうる事象全ての集まり

コインを1回投げたときの標本空間は？

確率の基礎

試行

同じ状態のもとで繰り返すことができ

その結果が偶然によって決まる実験や観測（サイコロを1回投げる）

事象

試行の結果起こる事柄（サイコロを投げて3が出る）

標本空間

ある試行において起こりうる事象全ての集まり

コインを1回投げたときの標本空間は？ 表が出る・裏が出る

確率の基礎

確率の定義

$P(A)$ ：ある試行に対して事象Aが起こる確率

それぞれの事象はそれぞれ同様に確からしいとする

$$P(A) = \frac{\text{事象 } A \text{ が起こる場合の数}}{\text{起こりうる全ての場合の数}}$$

データ分析ではもう一つ重要な確率の ”見積もり方” がある

相対度数

$$\text{相対度数} = \frac{\text{度数}}{\text{データの個数}}$$

データの個数（標本サイズ）が大きくなり、
母集団に近づくにつれて相対度数は確率に近づく

日本人の身長

	度数	相対度数
160.0cm 未満	5	0.125
160.0~169.9cm	10	?
170.0~179.9cm	15	0.375
180.0cm以上	10	0.25
総計	40	1

データが全日本人の場合：確率 = 相対度数

データが全日本人ではない場合：確率 \approx 相対度数

日本人の身長

	度数	相対度数
160.0cm 未満	5	0.125
160.0~169.9cm	10	(10/40=)0.25
170.0~179.9cm	15	0.375
180.0cm以上	10	0.25
総計	40	1

データが全日本人の場合：確率 = 相対度数

データが全日本人ではない場合：確率 \approx 相対度数

確率の基礎

和事象 $A \cup B$: 「 A または B が起こる」という事象

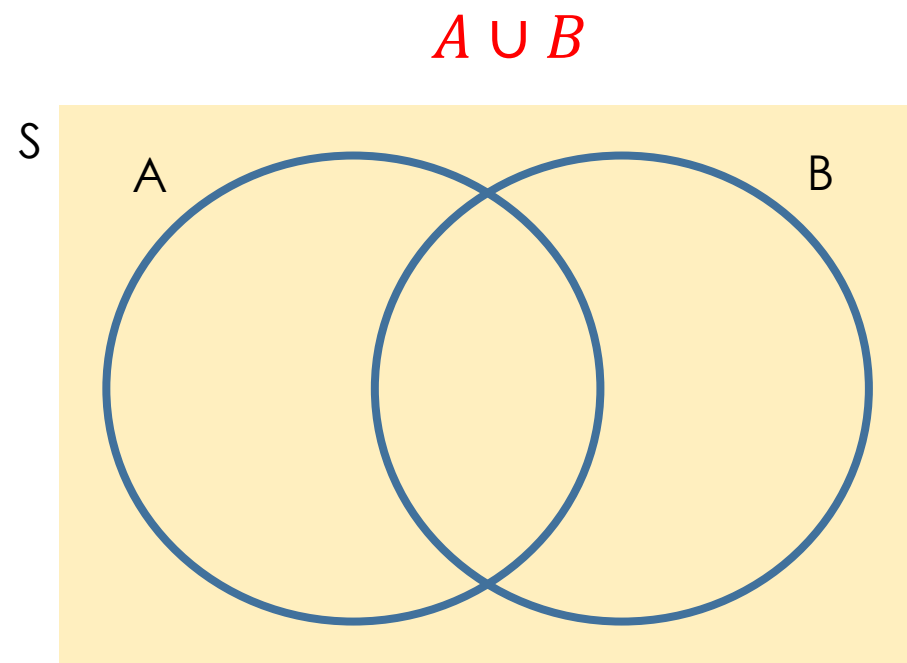
例)

1つのサイコロを1回振る

事象 A : 偶数が出る

事象 B : 1もしくは4が出る

和事象 $A \cup B$: ?



確率の基礎

和事象 $A \cup B$: 「 A または B が起こる」という事象

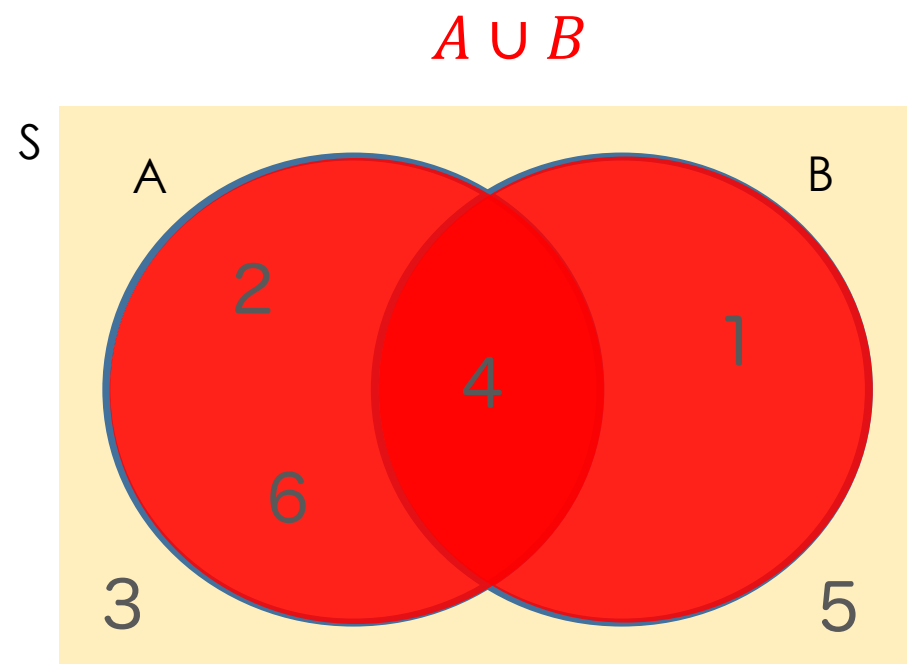
例)

1つのサイコロを1回振る

事象 A : 偶数が出る

事象 B : 1もしくは4が出る

和事象 $A \cup B$: 1, 2, 4, 6のいずれかが出る



確率の基礎

積事象 $A \cap B$: 「 A が起き、かつ B が起きる」という事象

例)

1つのサイコロを1回振る

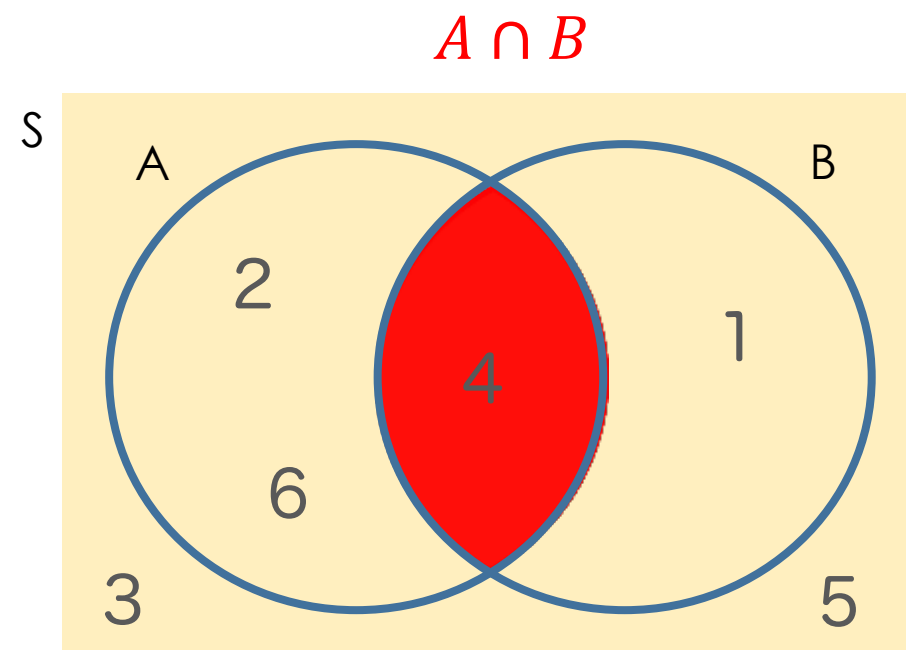
事象 A : 偶数が出る

事象 B : 1もしくは4が出る

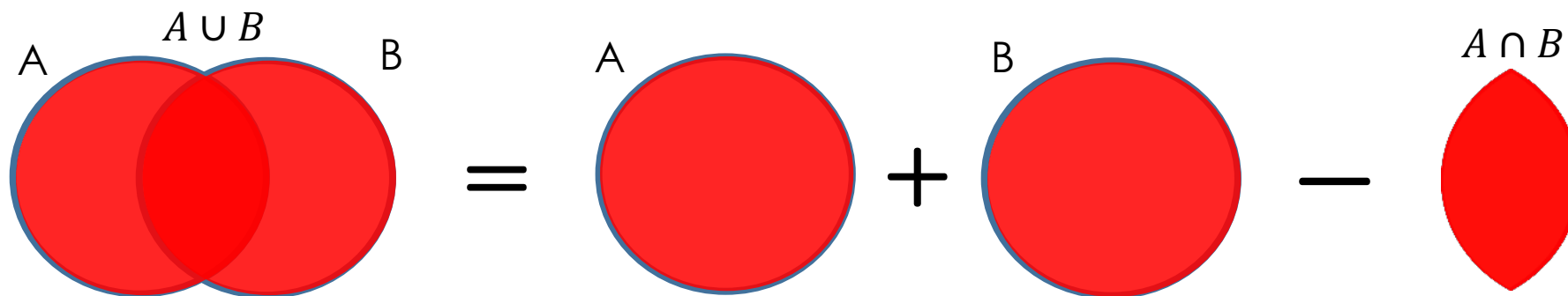
積事象 $A \cap B$: 4が出る

積事象の確率 $P(A \cap B)$ は

(A と B の) 同時確率とも呼ばれる



和事象の確率は **加法定理** を使って求められる



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

排反事象

「事象 A と事象 B が互いに排反」 $\Leftrightarrow P(A \cap B) = 0$

一つの事象が起こるともう一つの事象が絶対起こらない！

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

確率の基礎

排反事象の例)

試行：「1つのさいころを振り、出た目を観察する」

事象 A ：偶数の目が出る

A の排反事象：奇数の目が出る

事象 B ：1,2,3の目が出る

B の排反事象： ?

確率の基礎

排反事象の例)

試行：「1つのさいころを振り、出た目を観察する」

事象 A ：偶数の目が出る

A の排反事象：奇数の目が出る

事象 B ：1,2,3の目が出る

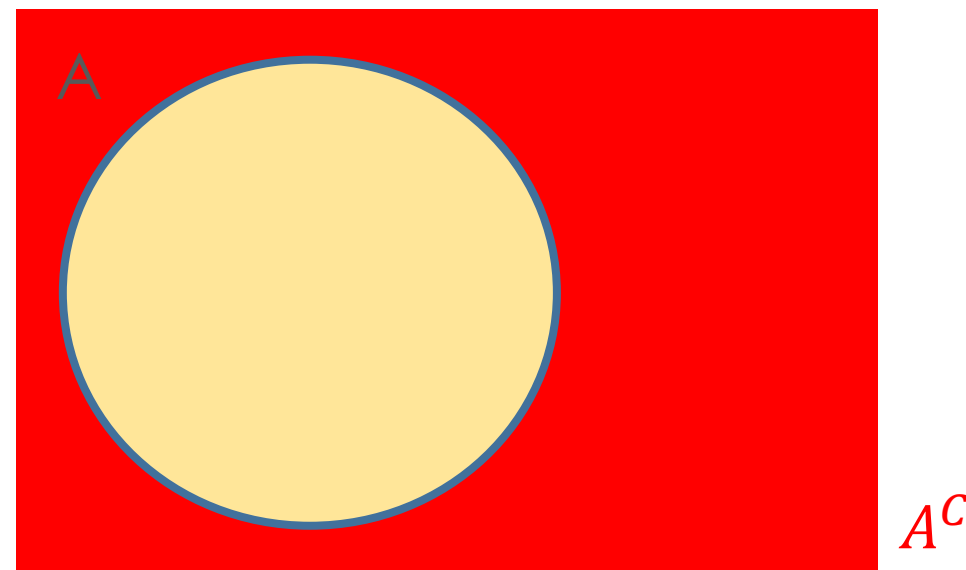
B の排反事象：4,5,6の目が出る

余事象

A の余事象 = 「 A が起こらない」という事象

S

- A^C で表す
- C はcomplementary (補足的な) の意味
- $P(A^C) = 1 - P(A)$



確率の基礎

余事象の例)

サイコロを2つふったとき、少なくとも片方の目が3になる確率

事象 A ：「少なくとも片方の目が3である」

余事象 A^C ： ?

$$P(A) = \quad ?$$

余事象の例)

サイコロを2つふったとき、少なくとも片方の目が3になる確率

事象 A ：「少なくとも片方の目が3である」

余事象 A^C ：「両方とも3の目ではない」

$$P(A) = 1 - P(A^C) = 1 - \frac{25}{36} = \frac{11}{36}$$

排反事象と余事象まとめ

- 排反事象とは？
- 余事象とは？

排反事象と余事象まとめ

- 排反事象とは？
 - 「1つの事象が起これともう一つの事象が絶対に起これない」
そういった2つの事象のことを排反事象という
- 余事象とは？
 - 「Aが起これない」という事象のこと

確率の基礎

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

$$\text{確率} \approx \text{相対度数} = \frac{\text{度数}}{\text{データの個数}}$$

$$P(\text{本社勤務で電車通勤ではない}) = \frac{5}{95} = \frac{1}{19}$$

$$P(\text{本社勤務である}) = \frac{25}{95} = \frac{5}{19}$$

確率の基礎

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A ：電車通勤である
事象 B ：本社勤務である

周辺確率：一つの事象についてのみ考えた確率

(周辺と言うのは念頭には同時確率 $P(A \cap B)$ があるから)

$$P(A) = P(\overset{A \cap B}{\text{電車通勤で本社勤務である}}) + P(\overset{A \cap B^c}{\text{電車通勤で本社勤務ではない}}) = \frac{20}{95} + \frac{30}{95} = \frac{10}{19}$$

$$P(B) = P(\overset{B \cap A}{\text{本社勤務で電車通勤である}}) + P(\overset{B \cap A^c}{\text{本社勤務で電車通勤ではない}}) =$$

確率の基礎

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A ：電車通勤である
事象 B ：本社勤務である

周辺確率：一つの事象についてのみ考えた確率

(周辺と言うのは念頭には同時確率 $P(A \cap B)$ があるから)

$$P(A) = P(\overset{A \cap B}{\text{電車通勤で本社勤務である}}) + P(\overset{A \cap B^c}{\text{電車通勤で本社勤務ではない}}) = \frac{20}{95} + \frac{5}{95} = \frac{25}{95} = \frac{5}{19}$$

$$P(B) = P(\overset{B \cap A}{\text{本社勤務で電車通勤である}}) + P(\overset{B \cap A^c}{\text{本社勤務で電車通勤ではない}}) = \frac{20}{95} + \frac{30}{95} = \frac{50}{95} = \frac{10}{19}$$

第1章：理解確認

(1) メールデータが30000件ある

このうちスパムメールが3000件であった

メールを一つ得た時、それがスパムメールではない確率 $P(S_{\text{no}})$ を
相対度数を元に見積もれ

(2) 3枚の硬貨を同時に投げるとき、表が2枚出る確率を求めよ

(3) 3枚の硬貨を同時に投げるとき、少なくとも1枚表が出る確率を求めよ

第1章：理解確認

$$(1) \quad P(S_{\text{no}}) = \frac{27000}{30000} = 0.9$$

$$(2) \quad {}_3C_2 \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{3-2} = \frac{3}{8}$$

${}_3C_2$: 3枚のコインから面になる2枚のコインを選ぶ組み合わせ

$\left(\frac{1}{2}\right)^2$: 表が2枚出る確率

$\left(1 - \frac{1}{2}\right)^{3-2}$: 裏が1枚出る確率

$$(3) \quad 1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}$$

少なくとも1枚表が出るという事象は、表が1つも出ないという事象の余事象

(参考) 組み合わせ

n, k を非負の整数($0, 1, \dots$)とする

組み合わせとは、異なる n 個の中から k 個を順番をつけずに選ぶ場合の選び方のこと

$${}_nC_k = \frac{n!}{k!(n-k)!} \quad {}nC_k \text{は} \binom{n}{k} \text{と表記することもある}$$
$$n! = n \times (n-1) \times \dots \times 2 \times 1$$

例)

A, B, C, D, Eの5つから3つ選ぶ

$${}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 10$$

第2章

条件付き確率

B という条件下における A の条件付き確率

$$P(A|B)$$

- 読み方： A ギブン B （A given B）
- $P(A|B) \neq P(A \cap B)$ であることに注意
- 事象 B が起きたという条件のもとで
事象 A が起こる確率のこと

条件付き確率の例

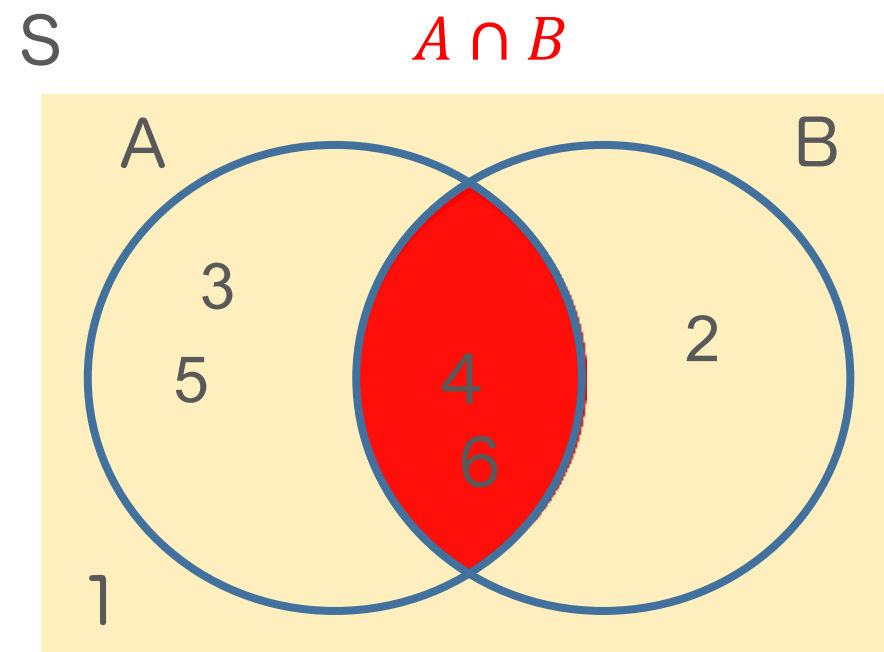
例) A : サイコロで3以上の目が出る

B : サイコロで偶数の目が出る

$$P(A \cap B) = \frac{2}{6} = \frac{1}{3}$$

$$P(A|B) = \frac{2}{3}$$

事象Bが起きたことを知ったのであれば
起こりうる事象の候補は全事象Sから
事象Bの範囲内に絞り込めるというイメージ



条件付き確率の例

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

例) ランダムに選んだ人が本社勤務だった場合、その人が
電車通勤の確率は $P(\text{電車通勤である} | \text{本社勤務である})$ で表せられる
本社勤務は25人でその内電車通勤が20人

$$P(\text{電車通勤である} | \text{本社勤務である}) = \frac{20}{25} = \frac{4}{5} (= 0.8)$$

条件付き確率の例

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

例) ランダムに選んだ人が電車勤務でなかった場合、その人が本社通勤ではない確率は $P(\text{本社勤務ではない} | \text{電車通勤ではない})$ で表せられる
電車通勤ではない人は45人でその内本社勤務ではない人が40人

$$P(\text{本社勤務ではない} | \text{電車通勤ではない}) =$$

条件付き確率の例

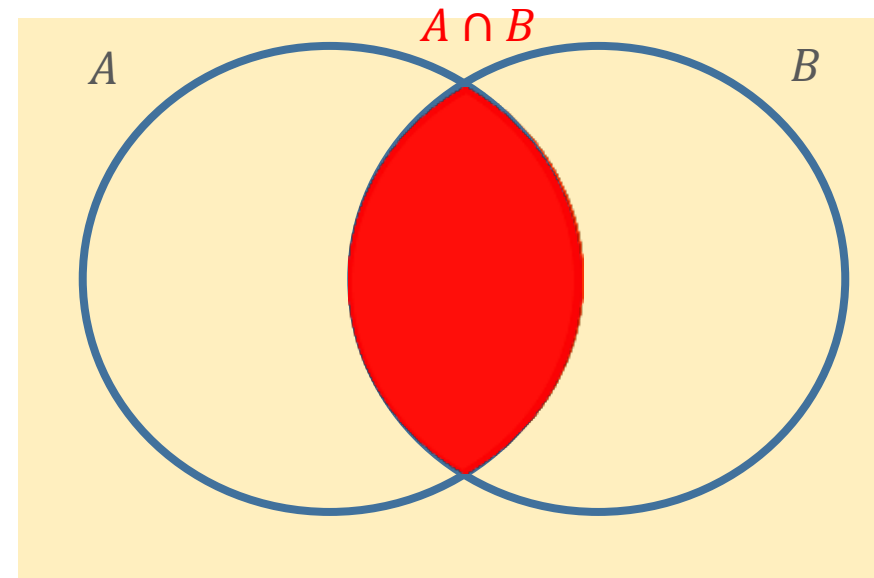
	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

例) ランダムに選んだ人が電車勤務でなかった場合、その人が
本社通勤ではない確率は $P(\text{本社勤務ではない} | \text{電車通勤ではない})$ で表せられる
電車通勤ではない人は45人でその内本社勤務ではない人が40人

$$P(\text{本社勤務ではない} | \text{電車通勤ではない}) = \frac{40}{45} = \frac{8}{9}$$

条件付き確率の公式

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$



条件付き確率の公式

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A ：電車通勤である
事象 B ：本社勤務である

ランダムに選んだ1人が電車通勤であった

このときその人が本社勤務である確率を条件付き確率の公式を用いて求めよ

条件付き確率の公式

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A ：電車通勤である
事象 B ：本社勤務である

ランダムに選んだ1人が電車通勤であった

このときその人が本社勤務である確率を条件付き確率の公式を用いて求めよ

解答)

電車通勤で本社勤務の人が20人、電車通勤の人が50人なので

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{20/95}{50/95} = \frac{20}{50} = \frac{2}{5}$$

第2章：理解確認

(1) 3本当たり5本はずれのくじを、A君とB君が順に引く

引いたくじはもとに戻さないとする

事象 A , B をそれぞれ

A : A君が当たりを引く

B : B君が当たりを引く

とするとき $P(B|A)$ を求めよ

(2) A : サイコロaを振り、1の目が出る

B : サイコロbを振り、2の目が出る

$P(B|A)$ を求めよ

第2章：理解確認

- (1) A君が当たりを引いたことが確定しているので
くじの中には2本の当たりと5本のはずれが入っている

よってB君が当たりを引く確率 $P(B|A) = \frac{2}{7}$

もしくは条件付き確率の公式を用いて $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{3}{8} \cdot \frac{2}{7}}{\frac{3}{8}} = \frac{2}{7}$

$$(2) \quad P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6}} = \frac{1}{6}$$

もしくは、Aが発生してもBの発生には全く関係がないと考え

$$P(B|A) = P(B) = \frac{1}{6} \text{ でもよい}$$

第3章

条件付き確率の発展

ベイズの定理

$P(A|B)$ と $P(B|A)$ を条件付き確率の定義を用いて書き下すと？



$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$P(A \cap B)$ を通して2つの式を結べそう！

ベイズの定理

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

ただし $P(B) \neq 0$

$P(A)$ ：事前確率

$P(A|B)$ ：事後確率

$$P(A|B) = \frac{P(B|A)}{P(B)} \times P(A)$$

修正項

事象 B が起きたことで事象 A の発生確率が何倍大きくなったか

もし事象 B が事象 A と独立なら $P(B|A) = P(B)$ となり更新されない
 $P(B)$ よりも $P(B|A)$ が大きいなら、事象 B は事象 A にも何らかの影響を及ぼしているはず

$P(A)$ という確率を B という情報を用いて更新している

この確率の更新方法を「ベイズ更新」と呼ぶ

(修了演習で再登場)

独立と条件付き独立

独立と条件付き確率

事象Aと事象Bが独立

Aが起きることとBが起きることは全く関係ない

$$P(A|B) = P(A) \quad \text{または} \quad P(B|A) = P(B)$$

A：サイコロaを振り、1の目が出る

B：サイコロbを振り、2の目が出る

$$P(B|A) = P(B) = \frac{1}{6}$$

$P(B|A)$ を求めよ

(第2章理解確認問題)

独立と条件付き確率

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A : 電車通勤である

事象 B : 本社勤務である

例) 事象 A と事象 B は独立？

$$P(A|B) =$$

$$P(A) =$$

独立と条件付き確率

	電車通勤	電車通勤ではない
本社勤務	20	5
本社勤務ではない	30	40

事象 A : 電車通勤である

事象 B : 本社勤務である

例) 事象 A と事象 B は独立 ?

$$P(A|B) = \frac{4}{5}$$

$$P(A) = \frac{10}{19}$$

$P(A|B) \neq P(A)$ より事象 A と事象 B は独立ではない

独立と条件付き確率

知っておきたい「独立」に関する知識①

事象 A と事象 B が互いに独立のとき、同時確率が簡単に求まる

$$P(A \cap B) = P(A)P(B|A) \rightarrow P(A \cap B) = P(A)P(B)$$

A : サイコロaを振り、1の目が出る

B : サイコロbを振り、2の目が出る

$P(A \cap B)$ を求めよ

$$P(A \cap B) = P(A)P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

独立と条件付き確率

知っておきたい「独立」に関する知識②

事象 C が起こったという条件を追加して

以下が成り立つことを条件付き独立と呼ぶ

条件付き独立

$$P(A \cap B|C) = P(A|C)P(B|C)$$

修了演習で登場！

独立と条件付き確率

ランダムに人を選んだ時に...

事象A：言葉を100個以上知っている

事象B：身長が160cm以上である

事象C：20歳男子である

身長が高い人ほど年齢が高い傾向がある
言葉を100個以上知っている可能性が高そう！

AとBは互いに影響している（独立ではない）

Cで条件付けすると？

独立と条件付き確率

ランダムに人を選んだ時に...

事象A：言葉を100個以上知っている

事象B：身長が160cm以上である

事象C：20歳男子である

事象Cの条件下では年齢が20歳だと分かっている
身長と知っている言葉の数に依存関係はなさそう！

「Cという条件下で」 AとBは互いに独立

第3章：理解確認

(1) $P(B|A)$ をベイズの定理を用いて書き下せ

(2) 3本当たり5本はずれのくじを、A君とB君が順に引く
引いたくじはもとに戻さないとする

事象 A, B をそれぞれ

A : A君が当たりを引く

B : B君が当たりを引く

とするとき $P(A|B)$ を求めよ

第3章：理解確認

$$(1) \quad P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad P(A) \neq 0$$

$$(2) \quad P(B) = \text{A君が当たりを引いてB君も当たりを引く確率} + \text{A君が外れを引いてB君が当たりを引く確率}$$

$$= \frac{3}{8} \cdot \frac{2}{7} + \frac{5}{8} \cdot \frac{3}{7} = \frac{21}{56} = \frac{3}{8}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{2}{7} \cdot \frac{3}{8}}{\frac{3}{8}} = \frac{2}{7}$$

第4章

修了演習

前提知識：パイ計算

パイ記号（ Π ）は数列の総乗を表す

Πa_i のように「上付き・下付き文字なし」の場合には a_i 全ての積を意味する

$$\prod a_i = a_1 \times a_2 \times \cdots \times a_n$$

例) データセット：(1, 2, 3, 4)

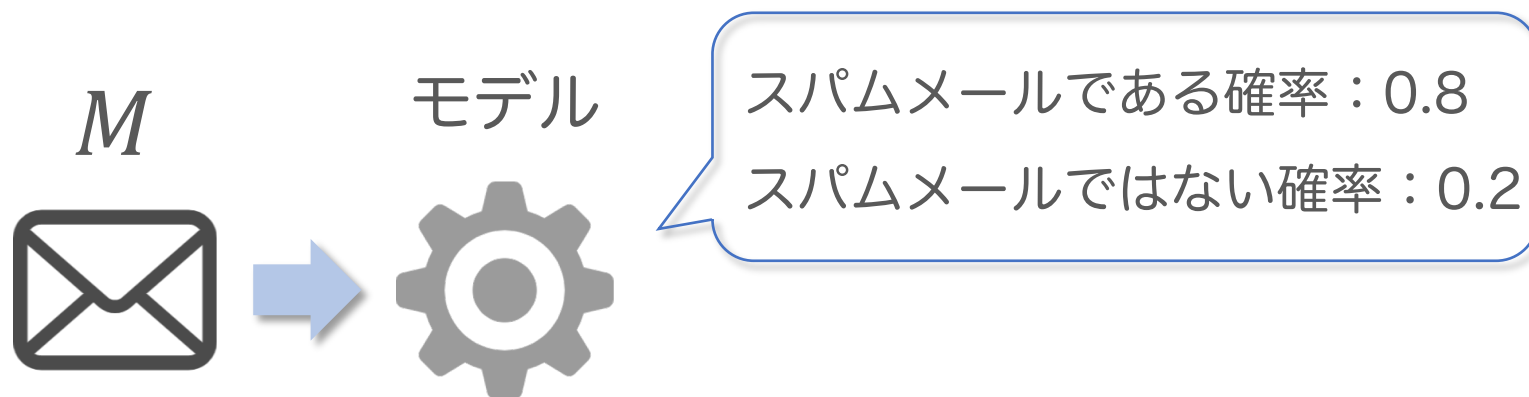
$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ とおく

$$\prod x_i = x_1 \times x_2 \times x_3 \times x_4 = 24$$

今回はこのパイ計算がたくさん出てきます

ナイーブベイズによるスパムメール判定

新しくメールが与えられた時にそのメールが
「スパムである」か「スパムではない」を判定したい！



S_{yes} : メールがスパムである

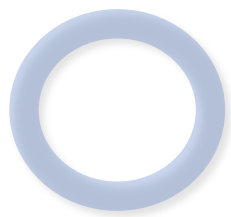
S_{no} : メールがスパムではない

M : 新しいメールが与えられる

まずはモデルに出力させたい確率を数式で表そう！

ナイーブベイズによるスパムメール判定

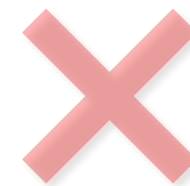
モデルに出力させたいもの



$$P(S_{\text{yes}}|M)$$

$$P(S_{\text{no}}|M)$$

メールが与えられたときに
それがスパムである（ではない）確率



$$P(S_{\text{yes}})$$

$$P(S_{\text{no}})$$

メール全体からメールを一つ取り出した時
それがスパムである（ではない）確率
（スパムメールの割合に対応する）

ナイーブベイズによるスパムメール判定

まずは $P(S_{yes}|M)$ を求めていこう！



本日習った定理で書き下すと？

ナイーブベイズによるスパムメール判定

まずは $P(S_{yes}|M)$ を求めていこう！



本日習った定理で書き下すと？

$$P(S_{yes}|M) = \frac{P(M|S_{yes})P(S_{yes})}{P(M)}$$

ベイズの定理！

ナイーブベイズによるスパムメール判定

$P(S_{\text{no}}|M)$ も同様に、ベイズの定理を用いて求めると

$$P(S_{\text{yes}}|M) = \frac{P(M|S_{\text{yes}})P(S_{\text{yes}})}{P(M)}$$

$$P(S_{\text{no}}|M) = \frac{P(M|S_{\text{no}})P(S_{\text{no}})}{P(M)}$$

右辺を求めればOK！

ただし、右辺の中で「求めなくても目的が達成できる」ものがある！

ナイーブベイズによるスパムメール判定

$P(M)$ は求めなくて良い！

$$P(S_{\text{yes}}|M) = \frac{P(M|S_{\text{yes}})P(S_{\text{yes}})}{P(M)} \propto P(M|S_{\text{yes}}) P(S_{\text{yes}})$$

$$P(S_{\text{no}}|M) = \frac{P(M|S_{\text{no}})P(S_{\text{no}})}{P(M)} \propto P(M|S_{\text{no}}) P(S_{\text{no}})$$

$P(M)$ は共通！

※ \propto : 比例
 $A \propto B$ なら A は B と
比例関係にあるという意味

$P(S_{\text{yes}}|M)$ と $P(S_{\text{no}}|M)$ の
比/大小関係に影響無し！

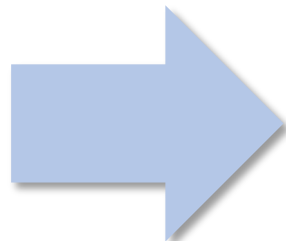
ナイーブベイズによるスパムメール判定

$P(S_{\text{yes}})$ と $P(S_{\text{no}})$ を求めよう！

メールデータセット
 N 件



...

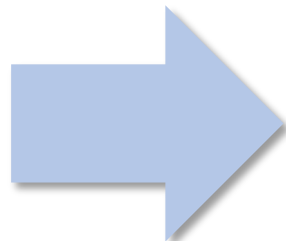


これを用いれば求まりそう！


ナイーブベイズによるスパムメール判定

$P(S_{\text{yes}})$ と $P(S_{\text{no}})$ を求めよう！


メールデータセット
 N 件



スパムメール

 N_{yes} 件 $P(S_{\text{yes}}) \approx \frac{N_{\text{yes}}}{N}$

スパムではないメール

 N_{no} 件 $P(S_{\text{no}}) \approx \frac{N_{\text{no}}}{N}$

これを用いれば求まりそう！

※ \approx : 近似

$P(S_{\text{no}}) \approx \frac{N_{\text{no}}}{N}$ なら、確率 $P(S_{\text{no}})$ を度数 $\frac{N_{\text{no}}}{N}$ で近似するという意味

ナイーブベイズによるスパムメール判定

$P(M|S_{\text{yes}})$ と $P(M|S_{\text{no}})$ を求めるのが難所

まずは $P(M|S_{\text{yes}})$ を考える

$P(M|S_{\text{yes}})$ とは

「スパムメールの世界において、そのメールが出現する確率」

のこと

「メール」というデータをもっと噛み砕く

“メール” とは何？！



ナイーブベイズによるスパムメール判定

$P(M|S_{\text{yes}})$ と $P(M|S_{\text{no}})$ を求めるのが難所

まずは $P(M|S_{\text{yes}})$ を考える

$P(M|S_{\text{yes}})$ とは

「スパムメールの世界において、そのメールが出現する確率」

のこと

「メール」というデータをもっと噛み砕く

“メール” とは何？！   単語の集まり！

ナイーブベイズによるスパムメール判定

単語の集合！



=

W_1 : 本日は W_3 : 数学

...

W_2 : 以上

W_K : 宜しく

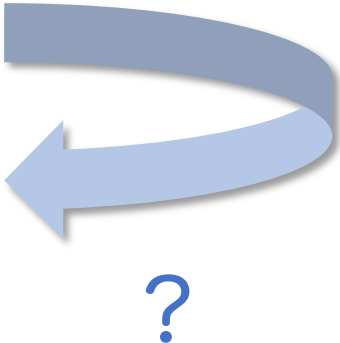
$P(M|S_{\text{yes}}) = P(W_1 \cap W_2 \cdots \cap W_K|S_{\text{yes}})$ と捉え直す！

W_i : 単語 i が発生するという事象

K : 単語の総数

ナイーブベイズによるスパムメール判定

計算が簡単にできるように、さらにもう一工夫！

$$\begin{aligned} P(M|S_{\text{yes}}) &= P(W_1 \cap W_2 \cdots \cap W_K | S_{\text{yes}}) \\ &= P(W_1 | S_{\text{yes}}) P(W_2 | S_{\text{yes}}) \cdots P(W_K | S_{\text{yes}}) \\ &= \prod_{i=1}^K P(W_i | S_{\text{yes}}) \end{aligned}$$


Π : 総乗. 全て掛け合わせるという意味

ナイーブベイズによるスパムメール判定

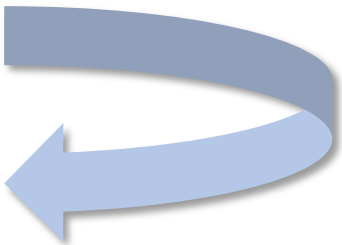
計算が簡単にできるように、さらにもう一工夫！

$$P(M|S_{\text{yes}}) = P(W_1 \cap W_2 \cdots \cap W_K | S_{\text{yes}})$$

$$= P(W_1 | S_{\text{yes}}) P(W_2 | S_{\text{yes}}) \cdots P(W_K | S_{\text{yes}})$$

$$= \prod_{i=1}^K P(W_i | S_{\text{yes}})$$

\prod : 総乗. 全て掛け合わせるという意味



条件付き独立
を仮定！

= ある単語の発生は他の
単語には依存しないと仮定

あとは、各単語 W_i について $P(W_i | S_{\text{yes}})$ を求めれば良い！

ナイーブベイズによるスパムメール判定

$P(W_i|S_{\text{yes}})$ ：スパムメールデータの中で単語 W_i が得られる確率

$W_i = \text{振込}$ とすると



振込



振込



$$N_{\text{yes}} = 3$$

$$N_{W_i} = 2$$

$$P(W_i|S_{\text{yes}}) = \frac{N_{W_i}}{N_{\text{yes}}} \quad \text{相対度数で見積もれる！}$$

ナイーブベイズによるスパムメール判定

新しいメール M が単語 $W_1 \sim W_K$ で構成されているとき

$$P(S_{\text{yes}}|M) \propto P(S_{\text{yes}})P(M|S_{\text{yes}}) = P(S_{\text{yes}}) \prod_{i=1}^K P(W_i|S_{\text{yes}}) = \frac{N_{\text{yes}}}{N} \prod_{i=1}^K \frac{N_{W_i}}{N_{\text{yes}}}$$

$$P(S_{\text{no}}|M) \propto P(S_{\text{no}})P(M|S_{\text{no}}) = P(S_{\text{no}}) \prod_{i=1}^K P(W_i|S_{\text{no}}) = \frac{N_{\text{no}}}{N} \prod_{i=1}^K \frac{N_{W_i}}{N_{\text{no}}}$$

$P(S_{\text{yes}}|M) > P(S_{\text{no}}|M)$ ならばスパムメールと判定する！

ナイーブベイズによるスパムメール判定

ナイーブベイズ

「条件付きの仮定」 + 「ベイズの定理」を用いた分類器

ナイーブ（単純）：

「条件付き独立」という強い仮定（複雑なことを考えない単純なアイデア）を導入するという意味

ナイーブベイズによるスパムメール判定

ベイズ更新による見方をしてみよう

$$P(S_{\text{yes}}|M) \propto \frac{N_{\text{yes}}}{N} \prod_{i=1}^K \frac{N_{w_i}}{N_{\text{yes}}}$$

ナイーブベイズによるスパムメール判定

ベイズ更新による見方をしてみよう

$$P(S_{\text{yes}}|M) \propto \frac{N_{\text{yes}}}{N} \prod_{i=1}^K \frac{N_{W_i}}{N_{\text{yes}}}$$

$P(S_{\text{yes}})$ M の内容に依存せず
 $P(S_{\text{yes}})$ の確率でスパムという意味

$P(S_{\text{yes}})$ を M という情報を用いて修正し $P(S_{\text{yes}}|M)$ を取得

データセット中のスパムの割合に加えて「メールの中身」まで考慮している

ナイーブベイズによるスパムメール判定

(1) 文章のカテゴリ分類をナイーブベイズ を用いて行いたい

D : 文章

$W_1 \sim W_K$: 文章 D に含まれる単語

C_{IT} : カテゴリがITである

C_{Art} : カテゴリが芸術である

N : 学習データの数

N_{IT} : ITカテゴリの文章

N_{Art} : 芸術カテゴリの文章

$N_{W_i, IT}$: ITカテゴリの記事における単語 W_i の出現回数

$N_{W_i, Art}$: Artカテゴリの記事における単語 W_i の出現回数

$P(C_{IT}|D)$ と $P(C_{Art}|D)$ を導け

ナイーブベイズによるスパムメール判定

(2) 文章 D の中に単語 W' が含まれているとする

この単語 W' が学習データの中に含まれていなかった場合

どのような不都合が起こるか

またその不都合はどのように解消すべきだろうか

(3) ナイーブベイズでは独立性の仮定を行う

例えば文章 D に対して単語の独立性を仮定することは

各単語の出現確率についてどのような仮定を置くことを意味するだろうか

ナイーブベイズによるスパムメール判定

(1)

$$P(C_{IT}|D) = \frac{P(D|C_{IT})P(C_{IT})}{P(D)} \propto P(D|C_{IT})P(C_{IT})$$

$$P(C_{Art}|D) = \frac{P(D|C_{Art})P(C_{Art})}{P(D)} \propto P(D|C_{Art})P(C_{Art})$$

$$P(D|C_{IT})P(C_{IT}) = \frac{N_{IT}}{N} \prod_{i=1}^K \frac{N_{W_i; IT}}{N_{IT}} \quad P(D|C_{Art})P(C_{Art}) = \frac{N_{Art}}{N} \prod_{i=1}^K \frac{N_{W_i; Art}}{N_{Art}}$$

ナイーブベイズによるスパムメール判定

- (2) $P(D|C_{IT})$ や $P(D|C_{Art})$ の中に「掛ける0」が発生し
値が0となってしまう

$$P(D|C_{IT})P(C_{IT}) = \frac{N_{IT}}{N} \prod_{i=1}^K \frac{N_{W_i; IT}}{N_{IT}} = \frac{N_{IT}}{N} \left(\frac{N_{W_1; IT}}{N_{IT}} \cdot \frac{N_{W_2; IT}}{N_{IT}} \cdot \dots \cdot \frac{N_{w'; IT}}{N_{IT}} \cdot \dots \cdot \frac{N_{W_K; IT}}{N_{IT}} \right)$$

$= 0$

通常の実装では、全ての出現頻度に対して以下のような修正を施す

$$\frac{N_{W_i; IT} + 1}{N_{IT}} : \text{スムージングと呼ばれる}$$

ナイーブベイズによるスパムメール判定

- (3) 「単語の並びは単語の出現確率（＝文章の出現確率）に影響を与えない」という仮定を置いていることを意味する

例えば「以上宜しくお願い致します」という文章が出現する確率は

$$p(\text{以上}) p(\text{宜しく} | \text{以上}) p(\text{お願い} | \text{以上, 宜しく}) p(\text{致します} | \text{以上, 宜しく, お願い})$$

と表現する方が自然である

「以上」の後には「宜しく」が出現しやすくなると考えられるので

$$p(\text{以上}) p(\text{宜しく} | \text{以上}) \neq p(\text{以上}) p(\text{宜しく})$$

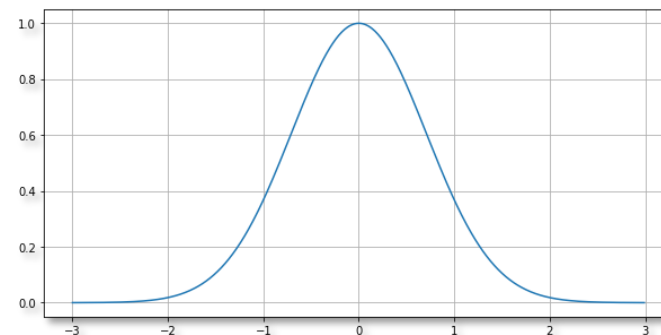
従って、ナイーブベイズでは文章の出現確率を少し雑に見積もることになる
ただこの仮定を置いても実用的な性能が出ることが知られている

次回予告

次回予告

次回は確率分布を学びます！

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (x \in \mathbf{R})$$



確率分布の知識を用いて

「ロジスティック回帰モデル」 「正規分布を用いた異常検知」

についても学びます！

お疲れ様でした！