情報理論

AVILEN

はじめに

Q なぜ機械学習で情報理論が必要なの？

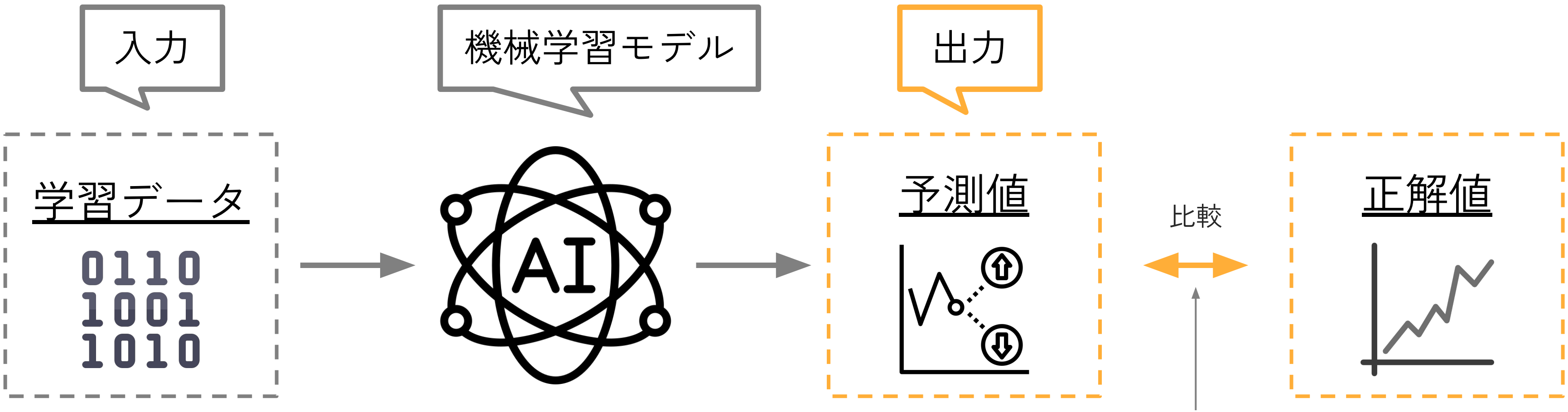
はじめに

Q なぜ機械学習で情報理論が必要なの？

A 損失関数の定義に必要だから！

教師あり学習では、損失関数の最小化を考える

教師あり学習における学習



損失関数を計算
→小さくなるように学習

損失関数と情報理論は密接な関係がある

情報理論と関わりの深い**損失関数**の例

- KLダイバージェンス
- クロスエントロピー
- JSダイバージェンス

→それぞれについて、定義や解釈のしかたを学んでいく。

まとめ

01

教師あり学習では、
予測と正解を比較する損失関数が必要

02

損失関数の中には、
情報理論と密接にかかわるものがある

はじめに

Q | 情報理論ではまず何を学ぶ？

はじめに

Q 情報理論ではまず何を学ぶ？

A 情報の価値を表す**情報量**について学ぶ！

情報の”価値”とは何か？

Q.　どちらのほうが驚きますか？

宝くじハズれた…　　宝くじ当たった！

OR

→「宝くじ当たった！」のほうが驚く

ありふれた出来事よりも珍しい出来事を
知るほうが、情報として”価値”がある

その”価値”を、**情報量(自己情報量)**と呼ぶ

情報量をどう定義すればよいだろうか？

おおまかに定義すると…

情報量＝事象の確率に応じて決まる量

情報量が最低限満たしてほしい条件

① 確率が同じ値なら情報量も同じ値

② 確率が小さいほど情報量は大きい

事象A

コインを投げたら
表が出た

確率

$$P(A) = \frac{1}{2}$$

事象B

コインを投げたら
裏が出た

確率

$$P(B) = \frac{1}{2}$$

確率が同じ
→情報量も同じに

事象C

サイコロを投げたら
1の目が出た

確率

$$P(C) = \frac{1}{6}$$

確率が小さい
→情報量は大きく

情報量を「確率の逆数」として定義してみる

例えば…
情報量Ｉを次のように定義してみる

$$I(X) = \frac{1}{P(X)}$$

事象Xの情報量を(仮に)
Xが起こる確率の逆数で定義する

このとき、前スライドの２条件を満たしている

① 確率が同じ値なら情報量も同じ値

$$I(A) = I(B)$$

② 確率が小さいほど情報量は大きい

$$I(B) < I(C)$$

しかし、この定義では値が大きくなりやすく扱いづらい

前スライドの定義にしたがうと…

「事象AとBが両方起きた」という事象がもつ情報量は

$$I(A,B)=\frac{1}{P(A,B)}=\frac{1}{P(A)P(B)}=4$$

同様に考えると…

「n個のコインが表、裏、裏、…、表だった」の情報量は

$$I(X_1,\cdots,X_n)=\frac{1}{P(X_1,\cdots,X_n)}=\frac{1}{P(X_1)\cdots P(X_n)}=2^n$$

事象A

コインを投げたら
表が出た

確率

$$P(A)=\frac{1}{2}$$

事象B

コインを投げたら
裏が出た

確率

$$P(B)=\frac{1}{2}$$

問題点 **nが大きくなると、値が爆発的に大きくなってしまう！**

情報量を「確率の逆数のlog」として定義してみる

かけ算されていくので爆発的に大きくなってしまう
→情報量が満たしてほしい条件を追加

③「独立な２事象が同時に起こった」事象がもつ情報量は、各事象の情報量の和でかける

条件③も満たすように、情報量の定義を改良

$$I(X) = \log_2\left(\frac{1}{P(X)}\right) = -\log_2 P(X)$$

条件③を満たしているか確認

$$\begin{aligned} I(A,B) &= -\log_2 P(A,B) \\ &= -\log_2 P(A)P(B) \\ &= -\log_2 P(A) - \log_2 P(B) \\ &= I(A) + I(B) \end{aligned}$$

よって、

$$I(A,B) = I(A) + I(B)$$

対数の底は何でもよい

情報量の定義

$$I(X) = -\log_{\underline{2}} P(X)$$

対数の底

対数の底は必ずしも2でなくてよい

→ 自然対数(底がe=2.718...)が用いられることも

情報量の単位

底が 2 のとき → **bit**（ビット）

底が e のとき → **nat**（ナット）

※一般的には底を2とすることのほうが多いため、以降の解説では底を2とした定義を採用します。

情報量を具体的に計算してみよう

問.　事象A~Cの情報量をそれぞれ計算せよ。

事象A
コインを投げたら
表が出た

$$P(A) = \frac{1}{2}$$

事象B
コインを投げたら
裏が出た

$$P(B) = \frac{1}{2}$$

事象C
サイコロを投げたら
1の目が出た

$$P(C) = \frac{1}{6}$$

答.

$$I(A) = -\log_2 \frac{1}{2} = \log_2 2 = 1$$

$$I(B) = -\log_2 \frac{1}{2} = \log_2 2 = 1$$

$$I(C) = -\log_2 \frac{1}{6} = \log_2 6 = 1 + \log_2 3 \fallingdotseq 2.585$$

例題

問.　宝くじの例でも、情報量をそれぞれ計算せよ。
　ただし、宝くじの当選確率は 1/2048 とする。

X「宝くじハズれた…」

Y「宝くじ当たった！」

答.

$$\begin{aligned} I(X) &= -\log_2 \frac{2047}{2048} \\ &= \log_2 2^{11} - \log_2 2047 \\ &\fallingdotseq 11 - 10.9993 \\ &= 0.0007 \end{aligned}$$

$$\begin{aligned} I(Y) &= -\log_2 \frac{1}{2048} \\ &= \log_2 2^{11} \\ &= 11 \end{aligned}$$

情報量の直感的な意味

情報量 1bit ってどれくらい？

||

「1枚のコインを投げて表が出た」
という事象がもつ情報量

$$-\log_2 \frac{1}{2} = 1 \ \ [\mathrm{bit}]$$

情報量は
「2択の事象が等確率で発生する」
というシンプルな場合を1単位としている

情報量は2進数の桁数を表す

コイン投げを数字に置き換える

表 ⟷ 数字の0

裏 ⟷ 数字の1

8回のコイン投げの結果は
2進数の数字8桁とみなせる

裏 表 表 表 表 裏 裏 表

＝1 0 0 0 0 1 1 0

情報量は $-\log_2 \frac{1}{2^8} = 8$ [bit]

→ 桁数に一致！

コンピュータは、情報を2進数で処理する

コンピュータの世界では
情報を0と1のみで表現する

この8bitをひとまとめにして
1 byteと呼ぶことが多い

1 byte（バイト） ＝ 8 bit

文字データ

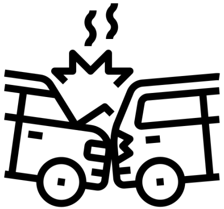
a → 10000110

2進数

この桁数が情報量に対応

例題

問.　Aさんは、交通事故が1日当たり平均5.0件発生する街に住んでいる。
下の情報M・Nでは、どちらのほうが情報量が多いか？



情報M「今日、この街では6件の交通事故が発生した」
情報N「今日、Aさんはこの街で交通事故に遭った」

ただし、Aさんが事故に遭う確率は0.002で、1日の事故件数Xはポアソン分布

$$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}\quad (k=0,1,2,\cdots)$$

に従うものとする。

例題の答え

答.

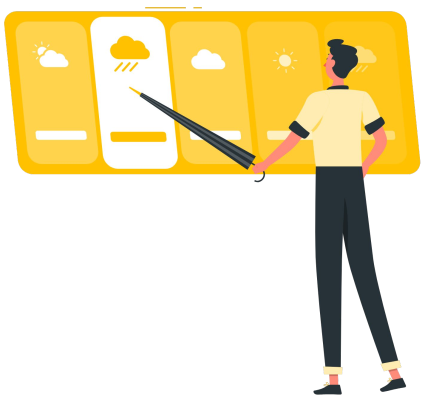
$$
\begin{aligned}
I(M) &= -\log_2 P(X=6) \\
&= -\log_2 \frac{e^{-5}5^6}{6!} \\
&= \log_2 6! - \log_2 e^{-5} - 6\log_2 5 \\
&\fallingdotseq 2.8
\end{aligned}
$$

$$
\begin{aligned}
I(N) &= -\log_2 0.002 \\
&= -\log_2 (2 \times 10^{-3}) \\
&\fallingdotseq 9.0
\end{aligned}
$$

よって、情報Nのほうが情報量が大きい。

天気予報にはどれくらいの情報量がある？

天気予報がどれくらい重要性を
持つかを定量的に知りたい



雨が多い国の天気

雨	晴れ
90%	10%

日本の天気

雨	晴れ
50%	50%

こっちのほうが天気予報のもつ
情報量が多そう？？

まずは、各事象について情報量を求めてみよう

「雨が降る」「雨が降らない」
それぞれの事象がもつ情報量を計算

雨が多い国の天気

90%　10%

「雨が降る」の情報量

$$-\log_2 0.9 \fallingdotseq 0.15$$

「雨が降らない」の情報量

$$-\log_2 0.1 \fallingdotseq 3.3$$

情報量の期待値をとったものを、”情報エントロピー”という

では、情報量の期待値はどうなる？

	雨が降る	雨が降らない
確率	0.90	0.10
情報量[bit]	0.15	3.3

期待値を計算すると…

0.90×0.15 ＋ 0.10×3.3 ≒ 0.47 bit

→ 情報量の期待値のことを
情報エントロピー*という

*平均情報量、**シャノンエントロピー**ともいう

情報エントロピーの定義式

一般には、確率分布Pに対する情報エントロピーHは次のように書ける！

・Pが<u>離散</u>確率分布のとき

$$H(P) = \sum_A P(A)I(A) = -\sum_A P(A)\log_2 P(A)$$

・Pが<u>連続</u>確率分布のとき

$$H(P) = \int_{-\infty}^{\infty} P(x)I(x)dx = -\int_{-\infty}^{\infty} P(x)\log_2 P(x)dx$$

日本の天気の場合も同様に計算してみよう

日本の場合の情報エントロピーは？

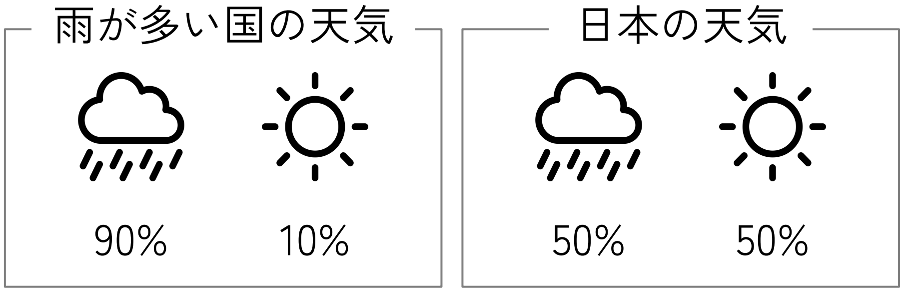
	雨が降る	雨が降らない
確率	0.50	0.50
情報量[bit]	1.0	1.0

期待値を計算すると…

$0.50 \times 1.0 + 0.50 \times 1.0 \fallingdotseq \underline{1.0\ \text{bit}}$

情報エントロピーの解釈

情報エントロピーは
日本のほうが大きくなった



0.47 bit **<** **1.0 bit**

日本の天気予報のほうが気になる！

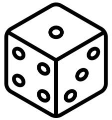
雨が多い国 ＝ ほぼ**確実に**雨が降る
→「明日は雨かどうか？」の情報価値：**小**

日本 ＝ 雨が降るかどうかが**不確か**
→「明日は雨かどうか？」の情報価値：**大**

情報エントロピーは
事象がどれくらい不確かなのかを表す指標

例題

問. 次の2つの場合について、それぞれ情報エントロピーを計算せよ。



A.　すべての目が等確率で出るサイコロを投げたとき

B.　それぞれの目が出る確率が下表のような、ゆがんだサイコロを投げたとき

出目	1	2	3	4	5	6
確率	0.90	0.02	0.02	0.02	0.02	0.02

例題の答え

答.

A(等確率)の情報エントロピー　：$-\sum_{i=1}^{6}\frac{1}{6}\log_2\frac{1}{6} \fallingdotseq 2.6$

B(歪みあり)の情報エントロピー：$-0.90\log_2 0.90-\sum_{i=2}^{6}0.02\log_2 0.02 \fallingdotseq 0.70$

(Aの情報エントロピー) > (Bの情報エントロピー)

→ Aの等確率で出るサイコロは、どの目が出るかがわからない
→ Aの結果の方が気になる(＝情報として価値がある)

まとめ

01 情報量は、ある事象が持つ情報の価値のこと

02 情報エントロピーは、情報量の期待値のこと

はじめに

Q

何のために確率分布を比較するの？

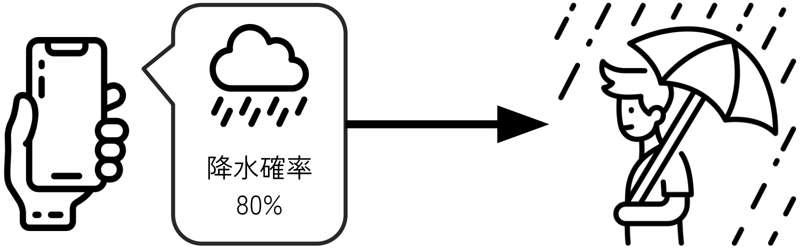
はじめに

Q 何のために確率分布を比較するの？

A 機械学習で、
予測と正解を比較するため！

天気予報と実際の天気の”差”を定量的に表したい

天気予報　　　実際の天気



降水確率80%と予報し、実際雨が降った

→感覚的には天気予報は当たっている

→予測と結果の差を定量的に表したい！

予報の確率分布

Q(X＝雨が降る)＝0.8

Q(X＝雨が降らない)＝0.2

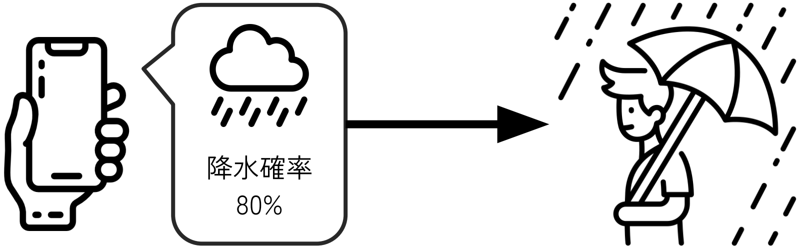
実際の確率分布

P(X＝雨が降る)＝1

P(X＝雨が降らない)＝0

クロスエントロピー

天気予報　　　実際の天気



予報の確率分布

Q(X＝雨が降る)＝0.8

Q(X＝雨が降らない)＝0.2

実際の確率分布

P(X＝雨が降る)＝1

P(X＝雨が降らない)＝0

「予測がどれくらい当たったか」を判断するためには、「予測が当たった」事象の持つ情報量の期待値を計算する

クロスエントロピー

(交差エントロピー)

$$H(P,Q) = -\sum_{X} P(X)\log_2 Q(X)$$

*連続確率変数の場合は、積分を使った定義となる。

KLダイバージェンス

KLダイバージェンス

$$D_{KL}(P||Q) = \underline{H(P,Q)} - \underline{H(P)}$$

情報エントロピー

クロスエントロピー
(交差エントロピー)

$$H(P,Q) = -\sum_X P(X)\log_2 Q(X)$$

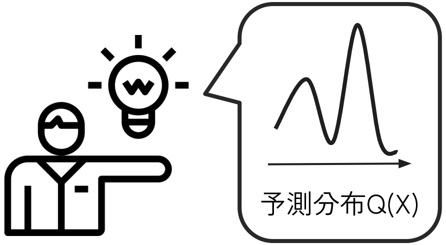
*連続確率変数の場合は、積分を使った定義となる。

「予測がどれくらい当たったか」だけではなく、実際の確率分布と予測の確率分布を比較したい場合にクロスエントロピーは使えない

KLダイバージェンスを使うと予測分布Qが、真の分布 Pに対してどれくらい正確かを知ることができる

KLダイバージェンスの解釈

分布をQ(X)と予測したが
実際はP(X)という分布だった



予測時点と観測時点で、
情報量はどれだけ減少する？

$$
\begin{aligned}
I(X) - I'(X) &= -\log_2 Q(X) - (-\log_2 P(X)) \\
&= \log_2\left(\frac{P(X)}{Q(X)}\right)
\end{aligned}
$$

この期待値を計算すると…

$$
\begin{aligned}
\sum_X P(X)\log_2\left(\frac{P(X)}{Q(X)}\right) &= -\sum_X P(X)\log_2 Q(X) + \sum_X P(X)\log_2 P(X) \\
&= H(P,Q) - H(P) \\
&= D_{KL}(P||Q)
\end{aligned}
$$

KLダイバージェンスの解釈

したがって、KLダイバージェンスとは…

予測Qを立てた後に、実際はPだと知ったときに減少する情報量の期待値

KLダイバージェンスが小さい

→ 減少する情報量が小さい

→ 真の情報を知っても
情報量があまり変化しない

→ 予測は妥当だった

KLダイバージェンスが大きい

→ 減少する情報量が大きい

→ 真の情報を知ると
情報量が大きく変化する

→ 予測は外れていた

KLダイバージェンスはやや使いづらい

問題点①

PとQを入れ替えても同じ値とは限らない

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

「P→Qの距離」≠「Q→Pの距離」
ということになり、不自然！

問題点②

分母が0になる可能性がある

$$D_{KL}(P||Q) = \sum_X P(X) \log_2 \frac{P(X)}{\boxed{Q(X)}}$$

予測Qの時にはありえない（確率0）と
思っていた事象が実際には発生しうる
場合に計算できない！

KLダイバージェンスの欠点を補うJSダイバージェンス

JSダイバージェンス

$$D_{JS}(P||Q) = \frac{1}{2}\left\{D_{KL}\left(P\middle|\middle|\frac{P+Q}{2}\right)+D_{KL}\left(Q\middle|\middle|\frac{P+Q}{2}\right)\right\}$$

JSダイバージェンスなら…

① PとQを入れ替えても同じになる

② 分母がゼロになることはない
(期待値計算でPとQが同時にゼロになることはない)

例題

問. 天気予報と実際の確率分布に対するJSダイバージェンスを計算せよ。

予報の確率分布

Q(X＝雨が降る)＝0.8

Q(X＝雨が降らない)＝0.2

実際の確率分布

P(X＝雨が降る)＝1

P(X＝雨が降らない)＝0

例題の答え

問. 天気予報と実際の確率分布に対するJSダイバージェンスを計算せよ。

予報の確率分布

Q(X＝雨が降る)＝0.8

Q(X＝雨が降らない)＝0.2

実際の確率分布

P(X＝雨が降る)＝1

P(X＝雨が降らない)＝0

$$D_{KL}\left(P \,\middle\|\, \frac{P+Q}{2}\right) = \sum_X P(X)\log_2\frac{2P(X)}{P(X)+Q(X)} = 1\cdot\log_2\frac{2\cdot 1}{1+0.8} = \log_2\frac{10}{9} = \log_2 10 - \log_2 9$$

$$D_{KL}\left(Q \,\middle\|\, \frac{P+Q}{2}\right) = \sum_X Q(X)\log_2\frac{2Q(X)}{P(X)+Q(X)} = 0.8\cdot\log_2\frac{2\cdot 0.8}{1+0.8} + 0.2\cdot\log_2\frac{2\cdot 0.2}{0+0.2}$$

$$= 0.8\cdot\log_2\frac{8}{9} + 0.2\cdot\log_2 2 = 2.4 - 0.8\log_2 9 + 0.2 = 2.6 - 0.8\log_2 9$$

$$\therefore D_{JS}(P \,\|\, Q) = \frac{\log_2 10 - \log_2 9 + 2.6 - 0.8\log_2 9}{2} = 0.5\log_2 10 - 0.9\log_2 9 + 1.3 \fallingdotseq 0.11$$

機械学習への応用

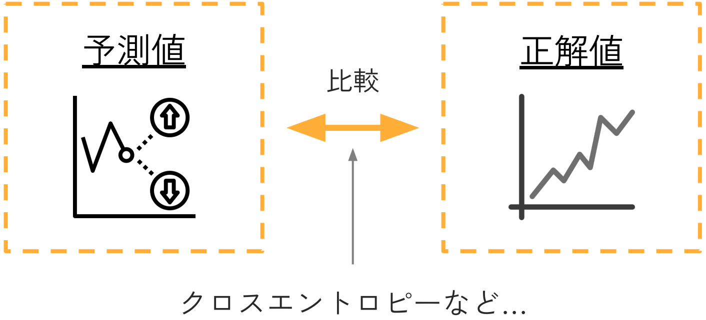
ここでは、2つの確率分布に対する
”距離”のようなものを学んできた

ここで学んだもの

KLダイバージェンス
クロスエントロピー
JSダイバージェンス

予測 Q と正解 P の差を計算するのに利用
＝**損失関数**の計算

→ この差が小さくなるようにパラメータを調整
→ 精度の良い予測を出す



まとめ

01

2つの確率分布の差を表すには、
KLダイバージェンスやJSダイバージェンスを使う！

02

KLダイバージェンスやJSダイバージェンスは
機械学習の損失関数に使える！