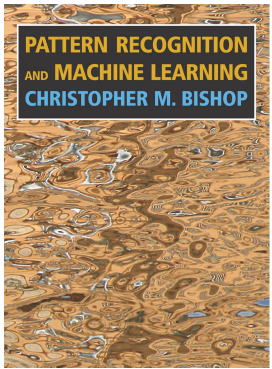
Introduction

正田 備也
masada@rikkyo.ac.jp

Contents

参考書 (1)

- C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.
- 日本語訳もあります。



参考書 (2)

- 鈴木譲『渡辺澄夫ベイズ理論 100 問 with Python/Stan』共立出版, 2024.
  - https://www.kyoritsu-pub.co.jp/book/b10084053.html

- 私はまだちゃんと読んでいませんが・・・
  - 「統計モデリング 1」の内容は、第 1 章に近いです。
  - 「統計モデリング 2」の前半の内容は、第 2 章に近いです。

Contents

統計的推測 (statistical inference) とは

- 我々が関心を持つデータは**ある分布**に従う。
  - この分布は、よく「真の分布」と呼ばれる。
- その分布に従うとみなせるデータ集合が**手元にある**。
  - このデータ集合は、よく「観測データ」と呼ばれる。

- そこで、**ある定められた手続き**にしたがって・・・
- 観測データから真の分布を**推測する (infer)**。
- これを、**統計的推測 (statistical inference)** と呼ぶ。

この授業で説明する統計的推測の方法

- 最尤法 (maximum likelihood estimation)
  - 「最尤推定」とも言う。
- 事後確率最大化法 (maximum a posteriori probability estimation)
  - 「MAP 推定」とも言う。
- ベイズ法 (Bayesian inference)
  - 「ベイズ推測」とも言う。(「ベイズ推定」や「ベイズ推論」と訳されることもある。)

統計学は不良設定問題を扱う学問

- 前のスライドに示したどの方法を使っても・・・
- 推測した分布 $\neq$ 真の分布
- つまり、推測は常に間違う！

- では、真の分布を推測しても無意味なのか？

- 推測がいつも間違うとしても・・・
- どのくらい間違っているかを、統計学で明らかにできる！

統計モデリングとは

- 真の分布は、未知。
- そこで我々は、データが従う分布を、自由に設定する。

- そして、推測された分布と、真の分布とが、どのくらい違っているかを、数理的に明らかにする。

- 「どのくらい違っているか」＝汎化誤差
  - 真の分布を $q(x)$、統計的推測によって得た予測分布を $p^\star(x)$ とする。
  - このとき、$K(q||p^\star)$ を汎化誤差という。
  - $K(\cdot||\cdot)$ は KL 情報量で、$\int q(x) \log(q(x)/p(x))dx$ と定義される。
  - 予測分布については、いずれ説明します。

ベイズ的統計モデリングの実験の手続き

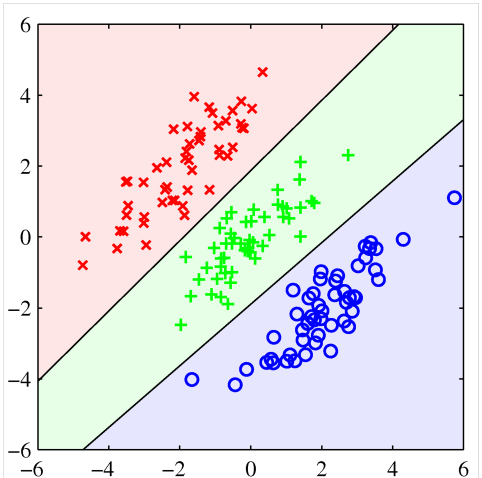
- 真の分布に従うとみなせるデータ集合を入手する。
  - つまり、観測する、ということ。
- モデルを設定する。
  - データをモデリングする分布と、事前分布とを、決める。
- 事後分布を（近似的に）求める。
  - 事後分布は、近似的にしか求められないことが多い。
- 予測分布を（近似的に）求める。
- テストデータ上で汎化誤差を（近似的に）求める。
  - モデル間で汎化誤差を比べれば、どのモデルが良いか、分かる。

Contents

機械学習の３区分

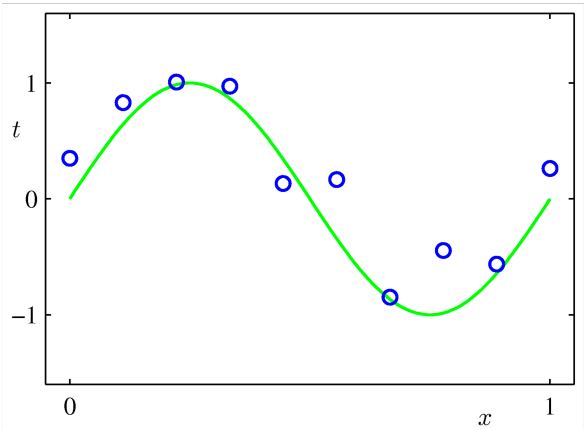
- 教師あり学習 (supervised learning)
  - classification
  - regression
- 教師なし学習 (unsupervised learning)
  - clustering
  - dimensionality reduction
- 強化学習 (reinforcement learning)
  - （ここでは触れません。）

classification のイメージ



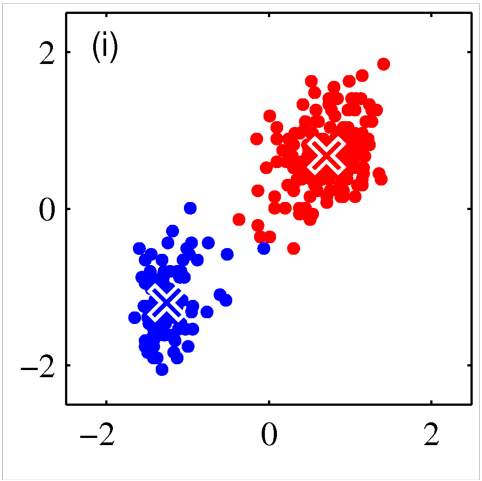
出典: C. Bishop. Pattern Recognition and Machine Learning. 2006.

regression のイメージ



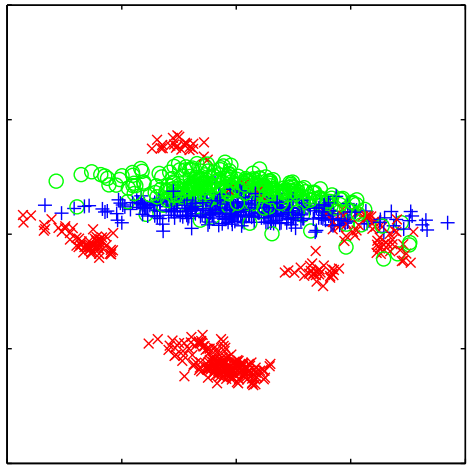
出典: C. Bishop. Pattern Recognition and Machine Learning. 2006.

clustering のイメージ



出典: C. Bishop. Pattern Recognition and Machine Learning. 2006.

dimensionality reduction のイメージ



出典: C. Bishop. Pattern Recognition and Machine Learning. 2006.

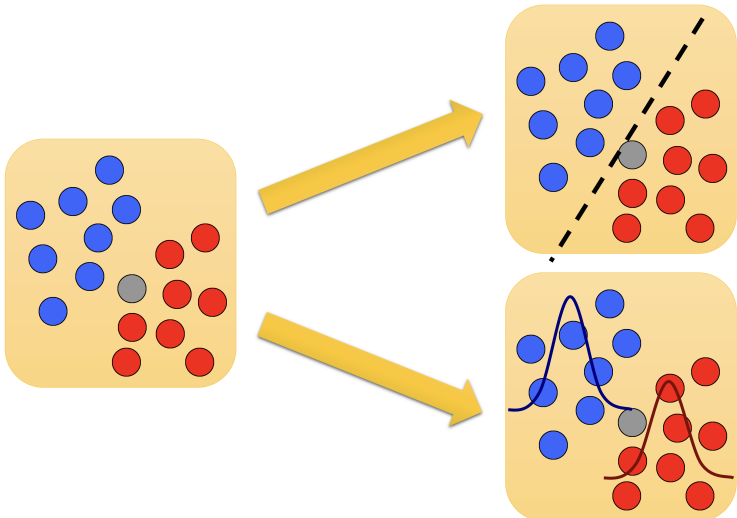
機械学習の別の2区分

- 識別的アプローチ
  - クラスを識別する境界を見つける。
  - 境界から遠いところにあるデータは気にしない。
- 生成的アプローチ
  - クラスが従う分布（＝クラスを生成する分布）を推測する。
  - 境界はこの分布の違いから派生する。

出典: http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

- 今はあまりこの区分について議論しなくなった。なぜなら・・・
- ほとんどの場合、生成的アプローチを使っているから。
  - それどころか我々は、観測データの分布に従う（と思われる）インスタンスを、いわゆる「生成 AI」を使って生成している。

discriminative vs generative のイメージ



Contents

generative approaches (1/2)

> *Generative models in machine learning posit that there is some underlying (···) process that is generating the data you are observing and aim to use the data to infer the parameters of that underlying process, which then lets you classify the data.*

出典: http://www.forbes.com/sites/quora/2015/02/12/what-is-the-future-of-machine-learning/

generative approaches (2/2)

In my mind, if you succeed in solving a generative model, you have "understood" the data and the problem.'

'Unfortunately for me, discriminative models tend to work better than generative models to solve lots of machine learning problems

出典: http://www.forbes.com/sites/quora/2015/02/12/what-is-the-future-of-machine-learning/

データを生成する確率分布を推定することの難しさ

- 2010 年前後は以下のような問題意識が共有されていた。

> 「統計的機械学習のほとんどの課題は、データの生成確率分布の推定を介して解決することができる。しかし、確率分布の推定は機械学習における最も困難な問題の一つとして知られているため、現実的には分布推定を回避しながら対象となる課題を解決することが望ましい。」

出典: http://ibisml.org/archive/ibisml001/Sugiyama.pdf

生成的アプローチの現在

- 今やすっかり状況は変わり・・・
- 我々は普通に・・・
  - テキストデーター般が従う分布を推定している。(ex. 言語モデル)
  - 画像データー般が従う分布を推定している。(ex. 拡散モデル)
- 深層学習が、非常に複雑な分布の推定を可能にした。
- しかし、やっていることは、昔と同じ。
  - 観測データの確率をできるだけ高くする分布を推定している。(後で説明する。)

Contents

確率の復習

- 確率変数 random variable
- 同時確率 joint probability
- 周辺化 marginalization
- 条件付き確率 conditional probability
- ベイズ則 Bayes rule

確率変数 random variable

- どの値をとるかが不確か(uncertain)な変数。
- $p(x = v)$ は、確率変数 x が値 v をとる確率。
- 確率なので $\sum_x p(x) = 1$ を満たす。
  - つまり、x がとりうる値を $\{v_1, \ldots, v_N\}$ とすると

$$p(x = v_1) + p(x = v_2) + \cdots + p(x = v_{N-1}) + p(x = v_N) = 1$$

- 世の中の event（事象）は、ある確率変数が特定の値を取ること（ex. $x_{21} = "apple"$）として表現する。

同時確率 joint probability

- $p(x = v, z = c)$ は、$x = v$ かつ $z = c$ となる確率。
- $\sum_x \sum_z p(x, z) = 1$ を満たす。
  - つまり、x がとりうる値を $\{v_1, \ldots, v_N\}$、z がとりうる値を $\{c_1, \ldots, c_K\}$ とすると

$$
\begin{aligned}
& p(x = v_1, z = c_1) + \cdots + p(x = v_1, z = c_K) \\
+\ & p(x = v_2, z = c_1) + \cdots + p(x = v_2, z = c_K) \\
+\ & \cdots + p(x = v_N, z = c_1) + \cdots + p(x = v_N, z = c_K) = 1
\end{aligned}
$$

周辺化 marginalization

$$p(x) = \sum_{z} p(x, z)$$

省略なしで書くと

$$p(x = v) = p(x = v, z = c_1) + \cdots + p(x = v, z = c_K)$$

- 特定の確率変数がとりうるすべての値にわたって確率を足し合わせることを、その変数の周辺化 (marginalization) と言う。
- 得られた確率を周辺確率 (marginal probability) と言う。

例：ハンバーガー

	標準体重未満	標準体重以上
毎日食べる	2	8
毎日は食べない	38	2

この例を確率変数を使って書くと…

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

- このように、確率変数を使って世の中の出来事を表現し直すことが、統計モデリングの第一歩。

問題 1-1

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

- $p(x = a)$ を求めよ。
- $p(z = t)$ を求めよ。
- $p(x = a, z = s)$ を求めよ。
- $p(x = a, z = t)$ を求めよ。

条件付き確率 conditional probability

$$p(x = v \mid z = c) \equiv \frac{p(x = v, z = c)}{p(z = c)}$$

- $z = c$ が所与のとき、$x = v$ である確率
- $\sum_x p(x \mid z = c) = 1$

問題1-2

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

- $p(x = a \mid z = s)$ を求めよ。
- $p(x = a \mid z = t)$ を求めよ。
- $p(z = s \mid x = a)$ を求めよ。
- $p(z = t \mid x = a)$ を求めよ。

周辺確率と条件付き確率の関係

- $p(z = s)$ は、x の値が未知のとき、$z = s$ である確率。
- $p(z = s \mid x = b)$ は、$x = b$ が観測されているとき、$z = s$ である確率。

- この 2 つの確率 $p(z)$ と $p(z \mid x)$ の関係は？
- これを表現するのが、ベイズ則。

ベイズ則 Bayes rule

- 事象 x の観測で z の確率が変わる、という式

$$p(z \mid x) \propto p(x \mid z)\, p(z)$$

- ある仮説が成り立つ確率 $p(z)$ は…
- その仮説が成り立っていると尤もらしさが増す（減る）事象 x が観測されると…
- 高い（低い）値 $p(z \mid x)$ になる

ベイズ則の証明

条件付き確率の定義より

$$p(z \mid x) = \frac{p(x, z)}{p(x)}, \qquad p(x \mid z) = \frac{p(x, z)}{p(z)}$$

二番目の式より $p(x, z) = p(x \mid z)\, p(z)$ なので、これを一番目の式に代入して

$$p(z \mid x) = \frac{p(x \mid z)\, p(z)}{p(x)}$$

$p(x)$ は z に依存せず、一定の値をとるので

$$p(z \mid x) \propto p(x \mid z)\, p(z)$$

ベイズ則の「比例する $\propto$」という記号

$$p(z \mid x) \propto p(x \mid z)\, p(z)$$

- $\propto$ は「左辺が右辺に比例する」という意味
- $p(z \mid x)$ を具体的な値として求めるには、まず、比例定数を求める必要がある。
- $\sum_z p(z \mid x) = 1$ が満たされるようにする比例定数を求める。

比例定数も含めてベイズ則を書くと・・・

$$p(z \mid x) = \frac{p(x \mid z)\, p(z)}{\sum_{z'} p(x \mid z')\, p(z')} = \frac{p(x \mid z)\, p(z)}{p(x)}$$

- 足して 1 になるようにする定数のことを、規格化定数 (normalizing constant) と呼ぶ。

先の例

	$x = a$	$x = b$
$z = s$	2	8
$z = t$	38	2

$$p(z = s \mid x = b) \propto p(x = b \mid z = s)\, p(z = s) \propto \frac{8}{10} \times \frac{1}{5} = \frac{4}{25}$$
$$p(z = t \mid x = b) \propto p(x = b \mid z = t)\, p(z = t) \propto \frac{2}{40} \times \frac{4}{5} = \frac{1}{25}$$

$\sum_z p(z \mid x = b) = 1$ が満たされないといけないので…（次スライド）

比例関係から実際の確率へ

規格化定数は $\frac{4}{25} + \frac{1}{25}$

よって

$$p(z = s \mid x = b) \propto \frac{4}{25},$$

$$p(z = t \mid x = b) \propto \frac{1}{25}$$

- これらの比例関係を規格化することで、以下を得る。

$$p(z = s \mid x = b) = \frac{4}{5}$$

$$p(z = t \mid x = b) = \frac{1}{5}$$

Contents

確率分布

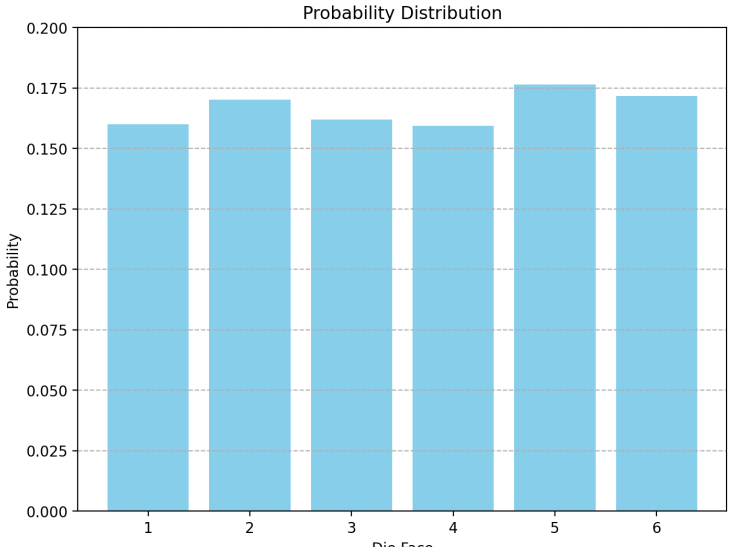
- 確率変数がとりうる値のすべてについて、その値をとる確率を示したものを、確率分布と呼ぶ。
- 確率変数がとりうる値すべてについて、その値をとる確率を足し合わせると、1になる。
- 連続値をとる確率変数の場合は、和ではなく積分で考える。

離散確率分布（例：サイコロ）

- 事象：「1の目が出る」,「2の目が出る」等
- 確率分布：各事象に確率を定めたもの

- 離散的なので、事象を一個一個と数えられる。
- すべての事象の確率を足すと1になる。

例）サイコロの目の確率分布



連続確率分布（例：体重測定）

- 事象：「体重が 57.5kg」，「体重が 70.1kg」等
- 確率分布：事象の集合に確率を定めたもの

- 連続的なので、事象を個々別々に扱えない。
- すべての事象を含む集合の確率は 1。

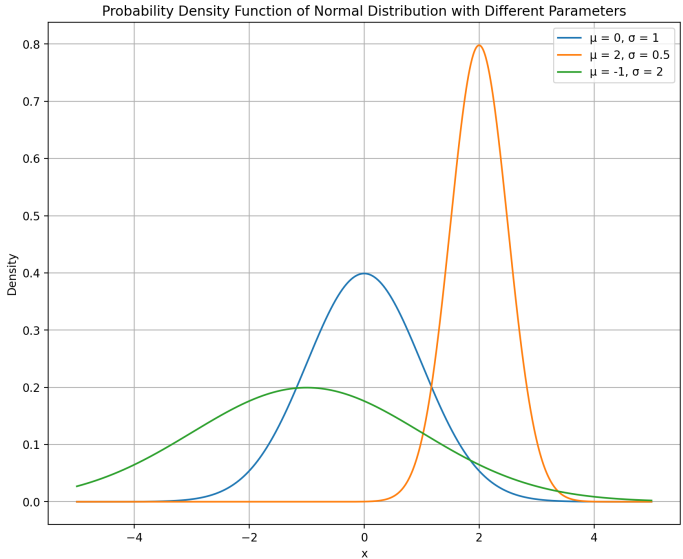
正規分布

- 平均 μ と標準偏差 σ で決まる分布
- 正規分布を表す関数

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

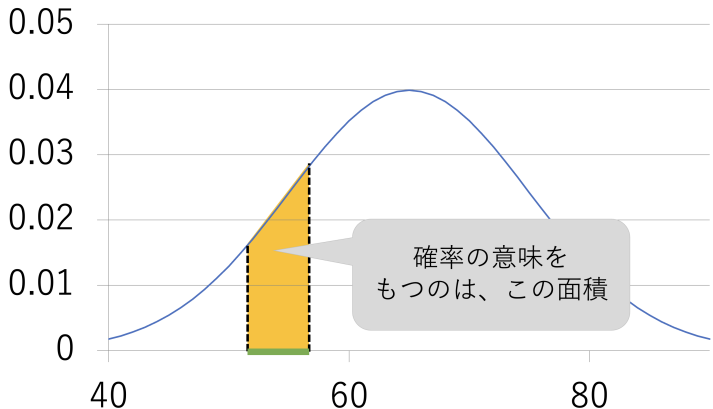
- この関数を、ある範囲で積分すると、その範囲の値をとる確率が求まる。
  - 例: $\int_{-3}^{2} f(x)dx$

正規分布 normal distribution

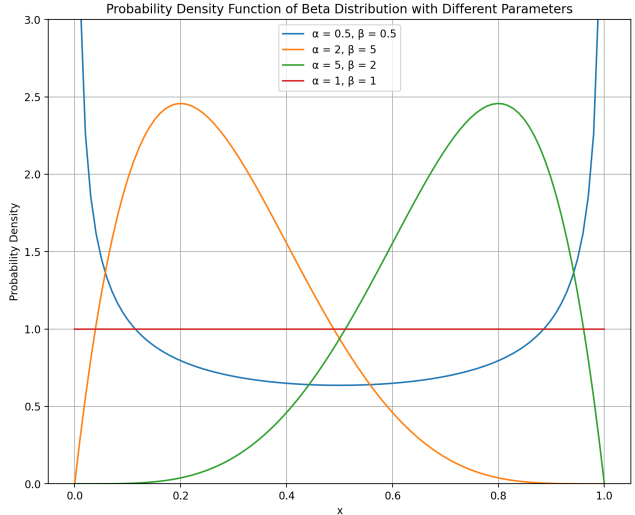


例：体重の確率分布

- 確率の意味をもつのは、面積。
- 曲線の下の面積（区間での積分）が確率を表す。



ベータ分布 Beta distribution



期待値 (1/2)

- 確率変数 x の期待値とは、その変数がとりうる値とその値をとる確率との積を、とりうる値すべてにわたって加算したもの。
- 連続値をとる確率変数の場合は、加算するのではなく積分する。

- 離散確率変数の場合：$\sum_x x\,p(x)$
- 連続確率変数の場合：$\int x\,p(x)\,dx$

期待値 (2/2)

- 変数 x の関数 $f(x)$ についても、各々の値をとる確率を掛けて和や積分をとることで、期待値が得られる。

- 離散確率変数の場合：$\sum_x f(x)\,p(x)$
- 連続確率変数の場合：$\int f(x)\,p(x)\,dx$

- 前のスライドは、$f(x)$ が恒等関数 $f(x) \equiv x$ の場合を説明していたとも言える。

問題 1-3

- 4 枚のコインを投げたとき、表が出た枚数と裏が出た枚数の差の絶対値の期待値はいくらになるか？
- ただし、これら 4 枚のコインは、表が出る確率も裏が出る確率も、ぴったり 0.5 だとする。

問題1-3の答え

$$\sum_{i=0}^{4} |\, i-(4-i)\,| \times \frac{4!}{i!(4-i)!} \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}\,(4\times 1+2\times 4+0\times 6+2\times 4+4\times 1) = \frac{24}{16} = \frac{3}{2}$$

課題 1

- わんこが 20 匹、にゃんこが 15 匹、バスに乗っている。
- わんこ 20 匹中、12 匹が白い毛である。
- にゃんこ 15 匹中、4 匹が白い毛である。
- このバスの中から無作為に 1 匹、乗客を選ぶ。

1. 選んだ乗客が白い毛である確率を求めよ。
2. 選んだ乗客がわんこであったときに、その乗客が白い毛である確率を求めよ。