ベイズ推論

Chapter 1（前半）

この本の位置付け.
- ベイズ主義機械学習(ベイズ学習)に基づいた,実践的なデータ解析アルゴリズムの構築法に関する解説本.

⇝ 目的や状況に合わせて自由にアルゴリズムを作れるようになることが目標.

Notation

確率分布
- $\mathcal{N}$: ガウス分布
- Bern : ベルヌーイ分布
- Bin : 二項分布
- Cat : カテゴリ分布
- Mult : 多項分布
- Poi : ポアソン分布
- Beta : ベータ分布
- Dir : ディリクレ分布
- Gam : ガンマ分布
- $\mathcal{W}$: ウィシャート分布

1. 機械学習とベイズ学習.
1.1. 機械学習とは

郷本における機械学習のテイギ.

機械学習
…データに潜む規則や構造を抽出することにより、未知の現象に対する予測やそれに基づく判断を行うための計算技術の総称.

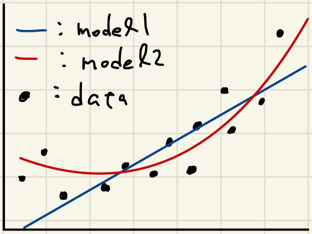
1.2. 機械学習の代表的なタスク.
◎各タスクはベイズ推論によるフレームワークで形式的にアプローチできる.

1.2.1. 回帰.
回帰 (regression)
- ある M 次元の入力 $x \in \mathbb{R}^M$ から連続値の出力 $y \in \mathbb{R}$ を予測するための関数 $y = f(x)$ をデータから求めるタスク.

線形回帰
$X = \{x_1, \dots, x_N\}$: 入力データ
$Y = \{y_1, \dots, y_N\}$: 入力データに対応する実数値出力データ.
$w \in \mathbb{R}^M$: パラメータ.

$$y_n = w^T x_n + \underbrace{\varepsilon_n}_{\text{確率的に生成されるノイズ.}} \qquad (1.1)$$

目標
手元にあるデータ X および Y を使い、各変数 x と y の関係をうまく捉えられるように w を求める.



・$M=1$ のとき、青線。

$y = wx$

⇝ 手元にあるデータの組から w を学習できれば、新規の入力点 x^* が与えられたときに、未知の出力 y^* を予測できる。

・$M=3$ のとき、

$(w_1, w_2, w_3)^T$: パラメータ

$(1, x, x^2)^T$: 入力変数

とする。

⇝ $y_n = w_1 + w_2 x_n + w_3 x_n^2 + \varepsilon_n$ (1.2)

上の2つの例は多項式回帰(polynomial regression)

… 元の観測データを解析しやすくするという意味で特徴量抽出(feature extraction)

1.2.2. 分類

出力値 y が有限個のシンボルに限定。

例: $y_n \in \{0, 1\}$

$\mu_n \in (0, 1)$: $y_n = 1$ となる確率。

$x_n \in \mathbb{R}^M$, $w \in \mathbb{R}^M$ から μ_n を表現するモデルは、

$\mu_n = f(w^T x_n)$ (1.3)

f にはシグモイド関数(sigmoid function)がよく使われる。

$$f(a) = Sig(a) := \frac{1}{1+e^{-a}} \quad (1.4)$$

・$w^T x$ の値によらず、$\mu_n \in (0, 1)$ となる。

◎ 実際に観測が $y_n = 0$ となるか、$y_n = 1$ となるかは μ_n のとる値に応じて、確率的に決定されるとする。

目標。

与えられたデータセット X, Y からパラメータ w を学習し、新規の入力 x^* に対する未知の出力値 y^* を確率的に予測する。

前述の考え方は、複数クラスの分類に対しても簡単に拡張できる。
K個のクラスに対する分類器
ソフトマックス関数 (softmax function)

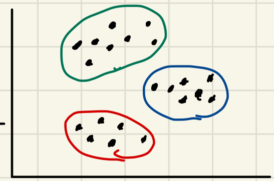
$$f_k(a) = SM_k(a) := \frac{e^{a_k}}{\sum_{k'=1}^{K} e^{a_{k'}}} \qquad (1.5)$$

を使うことで作成。

1.2.3. クラスタリング

クラスタリング (clustering)
与えられたN個のデータ $X = \{x_1, \dots, x_N\}$ を何かしらの基準に従って
K個の集合に分けるタスク。

例: ガウス混合モデル (Gaussian mixture model)
すべてのデータ X に対して、クラスタ所属の推定値 $S = \{s_1, \dots, s_N\}$ が
割り当てられる。
… クラスkの平均 μ_k, 共分散行列 Σ_k も学習。

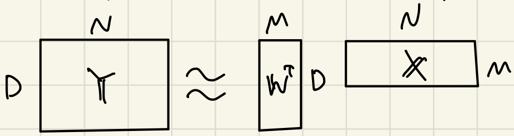


- 推定された S を使って単純にデータを分類、可視化
- 複数の異なる傾向をもつようなデータをモデリング。

1.2.4. 次元削減

線形次元削減 (linear dimensionality reduction)
通常 $M < D$ とする
データ $Y \in \mathbb{R}^{D\times N}$ を $W \in \mathbb{R}^{M\times D}$ および $X \in \mathbb{R}^{M\times N}$ を使って
近似する。

$$Y \approx W^T X \qquad (1.6)$$



⇝ 十分大きなNに対して、データ数が DN から $MD + MN$ に
圧縮できる。

例: $D = 100$, $N = 1000$, $M = 10$
→ $DN = 100000$, $DM + MN = 11000$ (9割削減)

一般に $W^T X$ は元のデータ Y を完全には復元できないが、
Y に存在する特徴的な情報をうまく保持するように学習アルゴリズム
が設計される。
… Y と $W^T X$ の誤差をある基準に従って最小になるように
W^T, X を決める。

@ 次元削減後の X はデータ Y に含まれるノイズを排除した本質的な
情報を保持していると考えることができる
→ X を他の分類アルゴリズムの入力として使うことも、実践では
よく行われる。

行列 Y のある n 番目の列に注目すると、
$$y_n \approx W^T x_n \qquad (1.7)$$
… 本質的には線形回帰 (1.1) と同じ。
… 各 x_n はデータから推定。

1.2.5. その他の代表的なタスク
- 与えられた信号の系列を統計モデルに基づき分離。
… 音声信号分離。
- 欠損値の補間
… ECサイトにおける商品のレコメンデーション
… 一般的なログデータの解析。

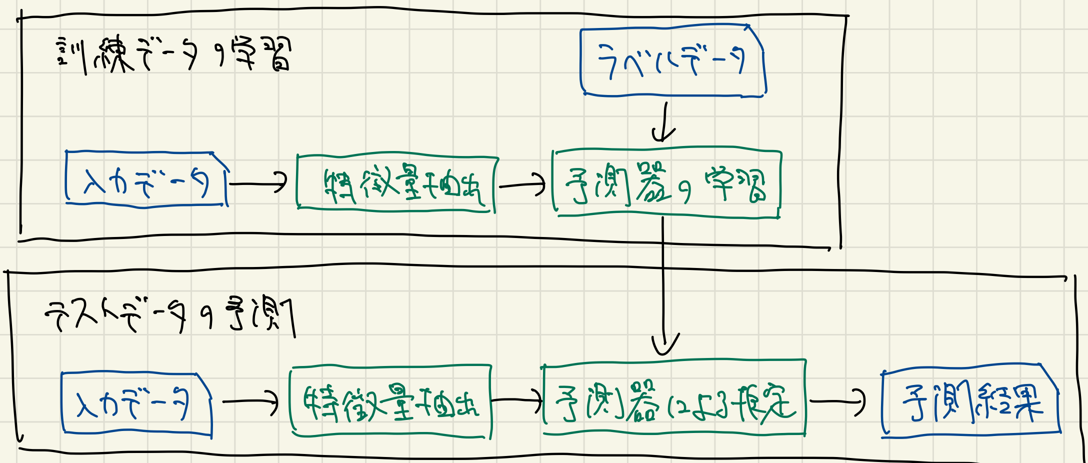
生成モデル (generative model)
… モデルが生のデータの生成過程を直接記述していれば、
人工的なデータのシミュレーションが可能。

1.3. 機械学習の2つのアプローチ
1.3.1. ツールボックスとしての機械学習
ツールボックスによるアプローチ
…既存のさまざまな予測アルゴリズムに対してデータを与え、その中から何らかの基準に従って性能の良いアルゴリズムを選ぶことで最終的な予測や判断を行う。

e.g. 最近傍法 (nearest neighbor)
…過去のデータに最も近い入力データを検索し、そのデータに対応するラベル(出力値)を予測結果とする。
サポートベクターマシン (support vector machine)
ブースティング (boosting)
ランダムフォレスト (random forest)

⤳ 教師あり学習 (supervised learning)
入力データと出力データ(正解ラベル)のペアを訓練データとして予測のための関数を学習する。

① アルゴリズム自体が特定の分野やデータに対してデザインされていることはあまりない。
⤳ 特徴量抽出 (feature extraction) と呼ばれるプロセスを前段階に置くことで個々のタスクに対する性能向上を目指す。



特徴量抽出の例
変換(多項式内積、フーリエ変換)、次元削減

アプローチの利点
・高度な数学の知識がなくとも、プログラミング技術があれば機械学習システムを構築できる。
　… 最適化はライブラリで提供。
・比較的簡単なチューニングを行うことで、所望の結果が得られることがある

アプローチの欠点
・使いこなすためには異なる構築思想を持った多くのアルゴリズムの動作原理を逐一理解しておかなければならない。
・精度や適用できる問題の範囲が限られる。

1.3.2. モデリングとしての機械学習
モデリングとしての機械学習
… データに関するモデル（仮説）を事前に構築。
　⤳ モデルの含むパラメータや構造をデータから学習することで、
　　何かしらの有益な予測や判断を行う。

① 対象となっている課題に対して仮定・制約できることを数理的に記述。
→ 推論や最適化などの計算的な手法に基づいてデータの背後に潜む特徴を抽出する。

トピックモデル（topic model）… 自然言語処理
・文章が複数の潜在的な（観測できない）トピックやテーマを持っていると仮定。
　→ トピックに応じて文章中の個々の語が出現しているとしてモデル化。

時系列モデル（time series model）
状態空間モデル（state space model）
e.g. 隠れマルコフモデル（hidden Markov model, HMM）
　… データの離散的な変化をモデル化。
　線形動的システム（linear dynamical system）
　… 状態の連続的な変化をモデル化。
　⤳ カルマンフィルタ（Kalman filter）… 推論アルゴリズム

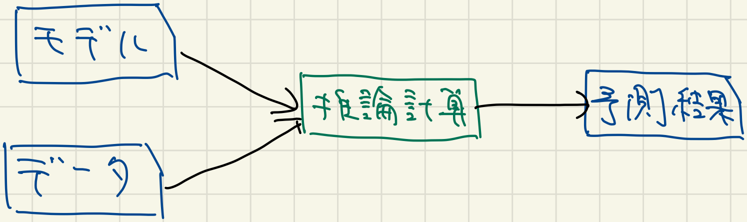
確率的ブロックモデル (stochastic block model)

…「つながり」のデータを記述するためのモデル.

深層学習 (deep learning)

… 一部は人間の脳における階層的な情報処理を部分的にモデル化.
⇝ 画像認識や音声認識のタスクで高い精度.

◎ モデリングアプローチのもっとも肝心な点は,
解きたい問題や解析したい対象に応じて
数理的にモデルを拡張したり組み合わせたりする「デザイン」
にフォーカスされていること.



特徴

主要なアルゴリズムの開発プロセスが

- ○ モデル構築
×
- ○ 推論計算

として明確に切り分けられる.

アプローチの利点

- ○ ツールボックスによるアプローチと比べて精度と柔軟性の点で優れている.
  - ・タスクに合ったモデルのデザインに焦点.
  - ・データに関する仮定を数理的に明記することで, 付随するタスクが自然に解けたり, 他のモデルや種類の異なるデータを統合できる.
- ○ 解析の対象となるシステムに対する不確実性をうまく表現できる.
- ○ 特性が良く知られている確率分布をブロックのように組み合わせることで, 各々の課題に適した解法を一貫性をもって提供できる.

アプローチの欠点.
- ある程度の数学の知識を要する.
- 計算時間やメモリコストがかかる.

1.4. 確率の基本計算

1.4.1. 確率分布

Def (確率密度関数)

Let $x = (x_1, \dots, x_M)^T \in \mathbb{R}^M$: 連続値

$p(x)$ is a p.d.f.

$$\overset{\text{def}}{\iff} \quad p(x) \geq 0 \qquad (1.8)$$

$$\int p(x)\, dx = \int \cdots \int p(x_1, \dots, x_M)\, dx_1 \cdots dx_M = 1 \qquad (1.9)$$

例: $M=1$ の Gaussian distribution に対する p.d.f.

$x \in \mathbb{R}$,

$\mu \in \mathbb{R}$: 平均, $\sigma^2 \in \mathbb{R}^+$: 分散.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (1.10)$$

Def (確率(質量)関数)

Let $x = (x_1, \dots, x_M)^T$: 離散値

$p(x)$ is a p.m.f.

$$\overset{\text{def}}{\iff} \quad p(x) \geq 0 \qquad (1.11)$$

$$\sum_{x} p(x) = \sum_{x_1} \cdots \sum_{x_M} p(x_1, \dots, x_M) = 1. \qquad (1.12)$$

例: ベルヌーイ分布

$x \in \{0, 1\}$, $\mu \in (0, 1)$

$$p(x) = \mu^x (1-\mu)^{1-x} \qquad (1.13)$$

$\begin{cases} x=1 : \text{コインの表が出る} \\ x=0 : \text{コインの裏が出る} \end{cases}$ 事象とする。

$$p(x=1) = \mu \quad (1.14)$$
$$p(x=0) = 1-\mu \quad (1.15)$$

μ：コインの表が出る確率。

この本で扱う範囲においては、p.d.f. や p.m.f. で定義される分布を確率分布と呼ぶことにする。

以降、一般的な議論を行う際には主に積分表記を用いる。

1.4.2. 確率分布の推論。
同時分布 (joint distribution) $p(x,y)$
ある2つの変数 x と y に対する確率分布

周辺化 (marginalization)

$$p(y) = \int p(x,y)\,dx \quad (1.16)$$

一方の変数 x を積分により除去する操作。
$p(y)$ を周辺分布 (marginal distribution) という。

条件付き分布 (conditional distribution)

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad (1.17)$$

y に対して特定の値が決められたときの x の確率分布。
$p(x|y)$ は x の確率分布
∴ y はこの分布の特性を決めるパラメータのようなもの。

(1.16)、(1.17) より、

$$\int p(x|y)\,dx = \frac{\int p(x,y)\,dx}{p(y)} = \frac{p(y)}{p(y)} = 1. \quad (1.18)$$

$p(x,y) \geq 0$, $p(y) \geq 0$ より、条件付分布 $p(x|y)$ は (1.8), (1.9) をみたす。

x が与えられたときの y の条件付分布も同様に、

$$p(y|x) = \frac{p(x,y)}{p(x)} \qquad (1.19)$$

Th (ベイズの定理)

$$\boxed{p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} = \frac{p(y|x)\,p(x)}{\int p(x,y)\,dx} \qquad (1.20)}$$

結果 y が得られたときの原因 x の確率

原因 x から結果 y が得られる確率。

◎ ベイズの定理は、時間順行の確率から時間遡行の確率を求められるというもの。

独立性 (independence)

$$p(x,y) = p(x)\,p(y) \qquad (1.21)$$

をみたすとき、x と y は独立であるという。

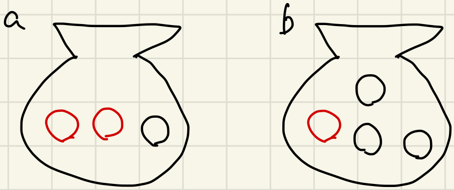
$$\frac{p(x,y)}{p(y)} = p(x) \quad (\because (1.21)) \qquad \therefore\ p(x|y) = p(x) \qquad (1.22)$$

y が条件で与えられても x の確率分布は変わらない。

◎ 確率推論に基づく機械学習の方法論は、周辺化、条件付分布の計算ルールがベース

ベイズ推論 (Bayesian inference / reasoning)
(確率)推論 (inference)
… ある同時分布が与えられたとき、そこから興味の対象となる条件付分布や周辺分布を算出すること。

1.4.3. 赤玉白玉問題.



◎ 完全にランダムな確率（$1/2$ ずつ）で袋 a または b のいずれかを選び、選んだ袋から玉を1つずつ取り出す試行を考える.

$$\begin{cases} x=a : 袋\ a\ が選ばれる \\ x=b : 袋\ b\ 〃 \end{cases}$$ 事象.

$$p(x=a) = \frac{1}{2} \qquad (1.23)$$
$$p(x=b) = \frac{1}{2} \qquad (1.24)$$

$$\begin{cases} y=r : 取り出された玉が赤 \\ y=w\ 〃\ 白 \end{cases}$$

$$p(y=r \mid x=a) = \frac{2}{3} \qquad (1.25)$$
$$p(y=w \mid x=a) = \frac{1}{3} \qquad (1.26)$$
$$p(y=r \mid x=b) = \frac{1}{4} \qquad (1.27)$$
$$p(y=w \mid x=b) = \frac{3}{4} \qquad (1.28)$$

(1.23)〜(1.28)を用いると、
同時確率
$p(x=a, y=r)$: 選ばれた袋が a で、赤玉が出る確率
$p(x=b, y=r)$: 〃 b 〃
は.

$$p(x=a, y=r) = p(y=r \mid x=a)\, p(x=a)$$
$$= \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3} \qquad (1.29)$$

$$p(x=b, y=r) = p(y=r \mid x=b)\, p(x=b)$$
$$= \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} \qquad (1.30)$$

「選ばれた袋に関わらず、取り出された玉が赤玉である」確率
⇝ 周辺確率 $p(x=r)$ に対応。

$$p(y=r) = \sum_x p(x, y=r)$$

$$= p(x=a, y=r) + p(x=b, y=r) \quad (1.31)$$

$$= \frac{1}{3} + \frac{1}{8} = \frac{11}{24} \quad (1.32)$$

「取り出された玉が赤であることが分かった場合に、選ばれた袋がaである」確率
⇝ $p(x=a \mid y=r)$

$$p(x=a \mid y=r) = \frac{p(x=a, y=r)}{p(y=r)} \quad (1.33)$$

$$= \frac{\frac{1}{3}}{\frac{11}{24}} = \frac{8}{11} \quad (1.34)$$

同じ条件下で、選ばれた袋がbであった場合を考える。

$$\sum_x p(x \mid y=r) = p(x=a \mid y=r) + p(x=b \mid y=r)$$

$$= 1 \quad (1.35)$$

より、

$$p(x=b \mid y=r) = 1 - p(x=a \mid y=r) = \frac{3}{11} \quad (1.36)$$

事後分布 (posterior)

データが観測されたあとのxの分布
(1.34), (1.36)

事前分布 (prior)

データを観測する前の分布
(1.23), (1.24)

1.4.4. 観測データが複数個ある場合.
1.4.3.の問題を発展させた問題を考える.

ある町内会で、袋 a, b を使ったくじ引き大会が開かれるとする.

Setting
- 主催者はどちらの袋を今回のくじ引き大会に用いるかを等確率で事前に選ぶ. どちらを選んだかは参加者に知らせない.
- 参加者は選ばれた袋 x から玉を取り出し、色を確認したら元の袋に戻す(復元抽出)

3人の参加者がくじ引きをし、次の結果を得た.

$$\{y_1, y_2, y_3\} = \{r, r, w\} \qquad (1.37)$$

このとき、主催者によって選ばれた袋 x が a, b のどちらであったかを知るためにはどうすればよいか.

⇝ 次の確率を推論する問題になる.

$$p(x \mid y_1 = r, y_2 = r, y_3 = w) \qquad (1.38)$$

ここで、復元抽出の仮定から、

$$p(y_1 = r, y_2 = r, y_3 = w \mid x) = p(y_1 = r \mid x)\, p(y_2 = r \mid x)\, p(y_3 = w \mid x) \qquad (1.39)$$

∵ 袋がすでに選択された状態において、ほかの参加者との順番、すでに取り出された結果とは無関係.

ベイズの定理と(1.39)より、

$$p(x \mid y_1 = r, y_2 = r, y_3 = w)$$
$$= \frac{p(y_1 = r, y_2 = r, y_3 = w \mid x)\, p(x)}{p(y_1 = r, y_2 = r, y_3 = w)}$$
$$= \frac{p(y_1 = r \mid x)\, p(y_2 = r \mid x)\, p(y_3 = w \mid x)\, p(x)}{p(y_1 = r, y_2 = r, y_3 = w)} \qquad (1.40)$$

(1.40) より、

$$p(x \mid y_1 = r, y_2 = r, y_3 = w)$$
$$\propto p(y_1 = r \mid x)\, p(y_2 = r \mid x)\, p(y_3 = w \mid x)\, p(x) \qquad (1.41)$$

これより、

$$p(x = a \mid y_1 = r, y_2 = r, y_3 = w)$$
$$\propto p(y_1 = r \mid x = a)\, p(y_2 = r \mid x = a)\, p(y_3 = w \mid x = a)\, p(x = a)$$
$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{27} \qquad (1.42)$$

および、

$$p(x = b \mid y_1 = r, y_2 = r, y_3 = w)$$
$$\propto p(y_1 = r \mid x = b)\, p(y_2 = r \mid x = b)\, p(y_3 = w \mid x = b)\, p(x = b)$$
$$= \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{128} \qquad (1.43)$$

これと、$\sum_x p(x \mid y_1 = r, y_2 = r, y_3 = w) = 1$ より、

$$p(x = a \mid y_1 = r, y_2 = r, y_3 = w) = \frac{2/27}{2/27 + 3/128} = \frac{256}{337} \qquad (1.44)$$

$$p(x = b \mid y_1 = r, y_2 = r, y_3 = w) = \frac{3/128}{2/27 + 3/128} = \frac{81}{337} \qquad (1.45)$$

上の計算で用いた
「分子だけ計算してあとで正規化」
というテクニックはよく用いられる。

ここからさらに観測値が得られたとする。

(例1) $\{y_1, \ldots, y_8\} = \{r, r, w, w, r, r, w, r\}$ (1.46)

$\rightsquigarrow p(x = a \mid y_1, \ldots, y_8) = 0.9221\ldots$ (1.47)

(例2) 50回の試行で赤玉29個、白玉21個

$\rightsquigarrow p(x = a \mid y_1, \ldots, y_{50}) = 0.9999\ldots$ (1.48)

◎ なるべく多くのデータを活用して、観測されていない未知の値を明らかにする

1.4.5. 逐次推論

逐次推論 (sequential inference)

・複数の独立なデータに関して成り立つ重要な性質。

1個の観測値 y_1 が得られたとする。
事後分布は、

$$p(x|y_1) \propto p(y_1|x)\,p(x) \qquad (1.49)$$

… 事前分布 $p(x)$ がデータ y_1 を観測することで $p(x|y_1)$ に更新された。

データが2個 $\{y_1, y_2\}$ 得られた場合、

$$\begin{aligned} p(x|y_1, y_2) &\propto p(y_1, y_2|x)\,p(x) \\ &\propto p(y_2|x)\,p(y_1|x)\,p(x) \\ &\propto p(y_2|x)\,p(x|y_1) \qquad (1.50) \end{aligned}$$

… データを1個観測した後の分布 $p(x|y_1)$ が、データ y_2 を観測することにより、事後分布 $p(x|y_1, y_2)$ に更新された。

一般に、観測値 $Y = \{y_1, \ldots, y_N\}$ のそれぞれが (x が与えられた条件で) 独立である場合、同時分布は、

$$p(x, Y) = p(x) \prod_{n=1}^{N} p(y_n|x) \qquad (1.51)$$

と書ける。

⇝ N番目のデータを受け取ったあとの事後分布は、

$$\begin{aligned} p(x|Y) &\propto p(x) \prod_{n=1}^{N} p(y_n|x) \\ &\propto p(y_N|x)\,\underbrace{p(x|y_1, \ldots, y_{N-1})}_{} \qquad (1.52) \end{aligned}$$

N−1個のデータを使って学習された事後分布。

逐次学習 (sequential learning)、追加学習 (incremental learning)、
オンライン学習 (online learning)

・以前に得られた事後分布を次の推論のための事前分布として使うような学習方法
(逐次的な更新手続き)

1.4.6. パラメータが未知である場合

1.4.3, 1.4.4 の問題は現実問題としてはシンプルすぎる。

例：・くじの主催者が本当に 1/2 の等確率で袋を選ぶかが信頼できない場合
・赤玉、白玉の出る比率を事前に正確に知ることができない場合
・袋の種類
・玉の色
⋮

⇝ 特に、「玉が取り出される比率のパラメータが未知」であり、いくつか玉を取り出すことでそのパラメータに関する推測を行いたいとする。

Setting

赤玉と白玉が複数個入った1つの袋があり、そこから玉を取り出して色を確認するという試行を N 回繰り返す。

目標

赤玉と白玉の未知の比率 $\theta \in (0, 1)$ に関して確率分布 $p(\theta)$ を考えることにより、確率的な推論だけを用いてデータから θ に関する知見を得ること。

$Y = \{y_1, \dots, y_N\}$: 観測データ
とする。

この問題を具体的に同時分布として記述。

$$p(\mathbf{Y}, \theta) = p(\mathbf{Y}|\theta)\, p(\theta) = \left(\prod_{n=1}^{N} p(y_n|\theta)\right) p(\theta) \quad (1.53)$$

各データ点 y_n が θの値で決定される確率分布によって生じる。

パラメータに関する事前分布。θのとりうる値に対する何らかの事前知識を反映。

◎ パラメータに関する事前知識を具体的に計算する方法として、
- 1. ベータ分布 (Beta distribution)
- 2. ディリクレ分布 (Dirichlet distribution)

を利用することになる。

目的

θのとりそうな値を、データ $\mathbf{Y}$ を観測することによって推論する。

⇝ $p(\theta|\mathbf{Y}) \propto p(\mathbf{Y}|\theta)\, p(\theta) \quad (1.54)$

を計算する。

- $p(\theta|\mathbf{Y})$: パラメータθの事後分布。データ $\mathbf{Y}$ の観測を通して事前知識 $p(\theta)$ が更新。

ベイズ推論を用いる機械学習では、
未知の値θに対して、何かしらの確率的な仮説 $p(\theta)$ を事前におき、
「データを観測することでその仮説を更新する」
という手続きをふむことで、直接観測できない現象を明らかにする。