情報論的学習理論

山西健司（2010）　共立出版

2022/6/7　理論談話会#13

学部4年　金田侑大

概要

情報論的

情報科学で扱われる問題

- データを効率的に**圧縮**するとは？
- あるデータの**最短符号長**はどのくらいか？
- データを短く**符号化**するアルゴリズムは？

+

学習理論

確率分布推定や将来予測をする際に扱う問題

- 何を**評価基準**に学習の成功を判断するか？
- 二つの**モデル**のどちらがよく学習できているといえるか？

$$P(A) = \frac{\exp(V_A)}{\exp(V_A) + \exp(V_B)}$$

概要

第1章　符号化と学習

データの圧縮方法として**語頭符号化**を導入

語頭符号化の符号長の下限が**Shannonの情報量**であることを示す

Shannonの情報量の発展として**確率的コンプレキシティ**を定義

情報論的

第2章　一括学習とモデル選択

得られたデータすべてを用いて**一括で**学習を行う場合

第3章　逐次符号化と逐次的予測

毎時得られるデータを用いて**逐次的に**学習を行う場合

学習理論

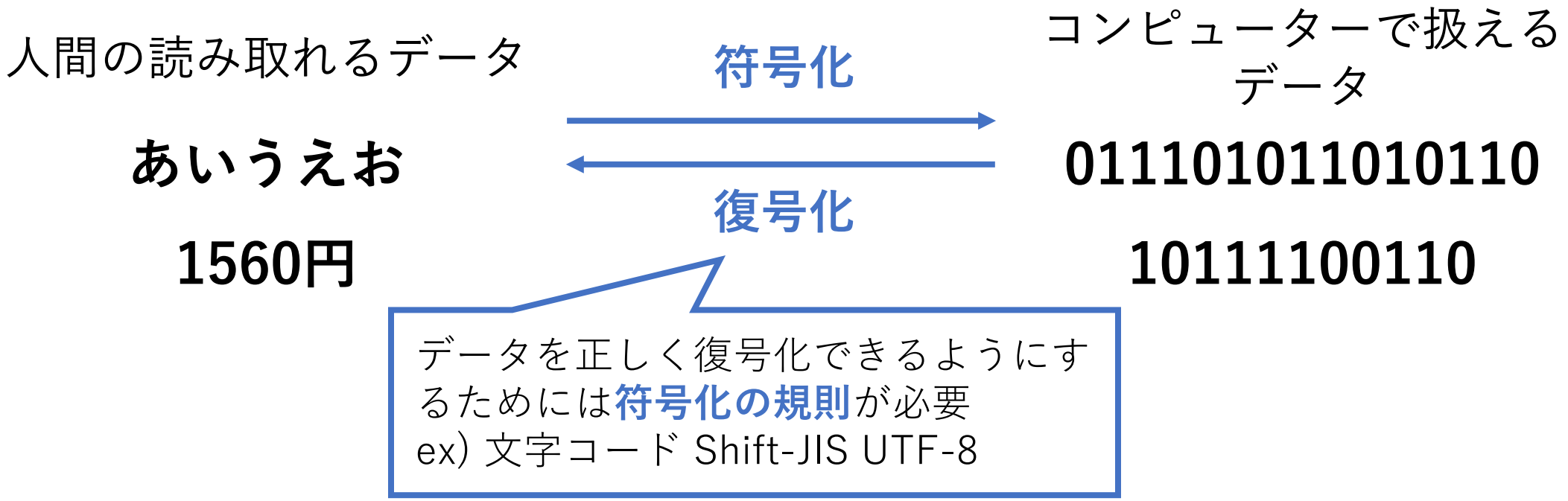
第1章

符号化と学習

符号化とは

情報科学の分野では…

データをコンピューター上で取り扱える2進数文字列に変換することを**符号化（encode）**という



語頭符号化

定義：語頭符号化

任意の文字 x_1, x_2 に対して、それを符号化した後の文字列 $\pi(x_1)$, $\pi(x_2)$ の一方が他方の先頭部分となっていない符号化を**語頭符号化**と呼ぶ。

例

語頭符号化でない

あ ⟶ 110

い ⟶ 1101

11010111011011…

「あ」と読めばいいか「い」と読めばいいか分からない！

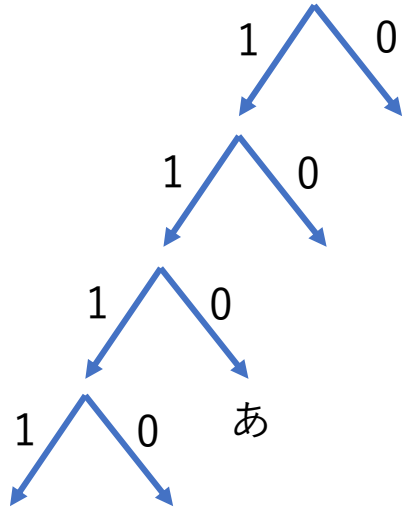
語頭符号化である

あ ⟶ 110

い ⟶ 101

110|101|110|110|11…

頭から順に読むことができる！



平均符号長

大量のデータを通信することを考えると、**符号長の期待値**が小さいほど効率的にデータを符号化できていると言える
（つまり、よく出る文字ほど短く符号化して、あまり出ない文字を長く符号化する）

定理1：語頭符号化の下限

ある確率分布 $P(X)$ に従って独立に文字 X が生成する文字列を仮定する。この時、どんな語頭符号化 l であっても

$$E[l(X)] = \sum_X P(X)l(X) \geq E[-\log_2 P(X)]$$

Shannonの情報量

が成立する

符号長がShannonの情報量となる仮想の語頭符号化を考えると、その符号化が平均符号長最小の符号化である

期待対数損失

定義：期待対数損失

データ X に対してその確率分布を $Q(X)$ と推定したとする。このとき $-\log_2 Q(X)$ を**対数損失**と呼ぶ。さらに真の分布 $P(X)$ 上での期待値 $E_P[-\log_2 Q(X)]$ を**期待対数損失**と呼ぶ。

解釈

データ x_1，x_2，x_3 がそれぞれ c_1，c_2，c_3 回出現していたとする。

尤度 $L = Q(x_1)^{c_1} Q(x_2)^{c_2} Q(x_3)^{c_3}$

対数尤度 $\log_2 L = c_1 \log_2 Q(x_1) + c_2 \log_2 Q(x_2) + c_3 \log_2 Q(x_3)$

データ数を増やし全データ数で割り負にする

$$-P(x_1)\log_2 Q(x_1) - P(x_2)\log_2 Q(x_2) - P(x_3)\log_2 Q(x_3)$$

期待対数損失

期待対数損失の下限

定理2：期待対数損失の下限

真の分布が$P(X)$であるデータXに対してその確率分布を任意に$Q(X)$と推定したとする。このとき期待対数損失について以下の式が成立する。

$$E_P[-\log_2 Q(X)] \geq E_P[-\log_2 P(X)]$$

Shannonの情報量

つまり…

平均符号長が最小となる符号化を求める
II
期待対数損失が最小となる確率分布を求める
（尤度最大化）

どちらも

$$-\log_2 P(X)$$

Shannonの情報量

が解になる

平均符号長最小化の学習問題における解釈

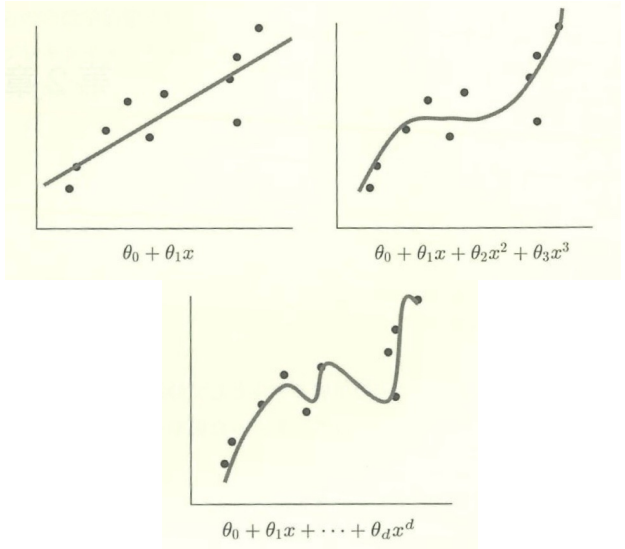
平均符号長最小化とは…

推定したい確率分布を真の分布に近づける行為である

期待対数損失（≒負の対数尤度）の最小化と等しい

現実をより簡潔に表現できるモデルを探す行為である

モデルの情報量が多いほど符号長は長くなる（過学習を防ぐ）



オッカムの剃刀

ある事柄を説明するには、より単純なものほど良い。

記述長最小原理

学習とはできるだけデータを圧縮できるような構造を見つけ出すことである

符号化の例

解きたい学習問題

いま n_1 文字の1と $n - n_1$ 文字の0からなる文字列 X^n（0100011010…など）がある。各文字 X は以下のベルヌイモデルに従って独立に生成されているとする。

$$P(X|\theta) = \begin{cases} \theta & : X = 1, \\ 1-\theta & : X = 0 \end{cases} \qquad (\text{ただし、} 0 \leq \theta \leq 1)$$

このとき、ベルヌイモデルを用いて X^n をできるだけ短く符号化することを考える。

二段階符号化、ベイズ符号化、正規化最尤符号化、数え上げ符号化

二段階符号化

Step1：Shannonの情報量（平均符号長の最小値）を θ で最小化する

$$\min_{\theta}(-\log_2 P(X^n|\theta)) = \min_{\theta}(-n_1 \log_2 \theta - (n-n_1)\log_2(1-\theta))$$

$$= n\left(-\frac{n_1}{n}\log_2\frac{n_1}{n} - \left(1-\frac{n_1}{n}\right)\log_2\left(1-\frac{n_1}{n}\right)\right) = nH\left(\frac{n_1}{n}\right)$$

Step2：Step1の θ を量子化する

任意の実数をそのまま通信することはできないので $0 \leq \theta \leq 1$ 上の有限個の代表点 $\bar{\theta}$ を代わりに利用する（＝**量子化**）

最尤値 $\hat{\theta}$ とすると代表点は $\bar{\theta} \approx \hat{\theta} + \delta$ と書き直せて符号長は $-\log_2 P(X^n|\hat{\theta}+\delta)$

Step3：θ 自体も符号化する

量子化幅 δ ごとに代表点が生成されるとするとその存在確率は δ で符号長は $-\log_2 \delta + O(1)$

データの最小符号化

モデルの最小符号化

最終的な符号長 $\min_{\delta}(-\log_2 P(X^n|\hat{\theta}+\delta) - \log_2\delta + O(1)) = nH\left(\frac{n_1}{n}\right) + \frac{1}{2}\log_2 n + \frac{1}{2}\log_2\frac{n^2}{n_1(n-n_1)} + O(1)$

ベイズ符号化

ベイズの定理を用いて

$$P(X^n) = \int \pi(\theta) P(X^n|\theta) d\theta \qquad \left(\text{ここで事前分布 } \pi(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)} d\theta} \right)$$

Fisherの情報量

最終的な符号長は

$$-\log_2 P(X^n) = nH\left(\frac{n_1}{n}\right) + \frac{1}{2}\log_2 \frac{n}{2\pi} + \log_2 \pi + O(1)$$

正規化最尤符号化

これまで一つの文字列 X^n に対する符号化のみを考えてきたが複数データが存在する中で目標とする文字列 X^n の最短符号化を考える。

$$P(X^n) = \frac{\min_\theta P(X^n|\theta)}{\sum_{X^n} \min_\theta P(X^n|\theta)}$$

各データについて最尤化すると確率の和が1を超えるので正規化する

最終的な符号長は

$$-\log_2 P(X^n) = -\log_2 \left(\min_\theta P(X^n|\theta) \right) + \log_2 \left(\sum_{X^n} \min_\theta P(X^n|\theta) \right)$$

$$= nH\left(\frac{n_1}{n}\right) + \frac{1}{2}\log_2 \frac{n\pi}{2} + o(1)$$

数え上げ符号化

問題の条件に合うような組み合わせの数は $\binom{n}{n_1} = \frac{n!}{n_1!(n-n_1)!}$

これがすべて等確率で生成されるモデルを考える。モデル自体の符号化は n_1 が $0 \leq n_1 \leq n$ の範囲の整数から等確率で生成されると考えると $\log_2(n+1)$ の符号長が必要である。

最終的な符号長はStirlingの公式を用いて

$$
\begin{aligned}
&-\log_2 \frac{n!}{n_1!(n-n_1)!} + \log_2(n+1) \\
&\approx nH\left(\frac{n_1}{n}\right) + \frac{1}{2}\log_2 \frac{n}{n_1(n-n_1)} + \log_2(n+1) - \log_2\sqrt{2\pi}
\end{aligned}
$$

符号長の比較

二段階符号化 $nH\left(\frac{n_1}{n}\right)+\frac{1}{2}\log_2 n+\frac{1}{2}\log_2\frac{n^2}{n_1(n-n_1)}+O(1)$

ベイズ符号化 $nH\left(\frac{n_1}{n}\right)+\frac{1}{2}\log_2\frac{n}{2\pi}+\log_2\pi+O(1)$

正規化最尤符号化 $nH\left(\frac{n_1}{n}\right)+\frac{1}{2}\log_2\frac{n\pi}{2}+o(1)$ ←実は一番良い符号化

log nの項までは符号長が同じ（＝nが大きくなればほぼ一致する）

数え上げ符号化 $nH\left(\frac{n_1}{n}\right)+\frac{1}{2}\log_2\frac{n}{n_1(n-n_1)}+\log_2(n+1)-\log_2\sqrt{2\pi}$

他の符号化より1/2 log nだけ符号長が長い
（ベルヌイモデルの仮定をしていない分）

確率的コンプレキシティ

今考えている問題を整理（パラメータ推定問題）

n : 学習データのデータ数に相当

データ列 X^n に対して、その生成確率分布をパラメータ θ を用いた確率モデル $P(X^n|\theta)$ で推定したい。（これら確率モデルの集合 $\mathcal{P}_k = \{P(X^n|\theta) \mid \theta \in \Theta_k\}$ を**確率モデルクラス**と呼ぶ）

k次元のパラメータ空間

定義：確率的コンプレキシティ

データ列 X^n を $\mathcal{P}_k$ 上で正規化最尤符号化を行ったときの符号長を**確率的コンプレキシティ**と呼ぶ。この量は確率モデルクラスのFisherの情報行列が良い性質を満たすとき

$$SC(X^n : \mathcal{X}_k) = \min_{\theta \in \Theta_k} (-\log_2 P(X^n|\theta)) + \frac{k}{2} \log_2 \frac{n}{2\pi} + \log_2 \int \sqrt{|I(\theta|} d\theta + o(1)$$

となる。

確率的コンプレキシティの正当化

定理3：真の分布が未知の場合の語頭符号長の下限

データ列 X^n が確率モデルクラス $\mathcal{P}_k$ 上の $P^*(X^n) = P(X^n|\bar{\theta})$ に従って生成されているとする。また θ の最尤推定値が $\bar{\theta}$ を平均とした正規分布に従うように現れると仮定する。n が大きくなるにつれて測度が0になる θ を除いて、任意の語頭符号化 l と任意の $\varepsilon > 0$ について以下が成立。

$$E_{P^*}[l(X^n)] \geq E_{P^*}[-\log_2 P^*(X^n)] + \frac{k-\varepsilon}{2}\log_2 n$$

定理1と比べると語頭符号化の際に確率分布の推定が必要になるため、その分符号長が長くなる

補題

上と同じ条件で $\left|E_{P^*}[-\log_2 P(X^n|\theta)] - \left(E_{P^*}[-\log_2 P(X^n|\hat{\theta})] + \frac{k}{2}\right)\right| = o(1) \xrightarrow{n\to\infty} 0$

確率的コンプレキシティの正当化

確率的コンプレキシティの真の分布における期待値

$$E_{P^*}[SC(X^n:\mathcal{P}_k)] = E_{P^*}[-\log_2 P(X^n|\hat{\theta})] + \frac{k}{2} + \frac{k}{2}\log_2 n - \frac{k}{2}\log_2 \pi + \log_2 \int \sqrt{|I(\theta|}d\theta + o(1)$$

この部分はnが十分大きいとき定理3の下限に一致　この部分も正規化最尤符号化が最小をとる

定理4：ミニマックスリグレットを達成する符号化

データ列 X^n の集合 $\mathcal{X}^n$ と確率モデルクラス $\mathcal{P}_k$ があるとき、語頭符号化 l について

$$\mathcal{R}_n(\mathcal{P}_k) = \inf_{l} \sup_{X^n \in \mathcal{X}^n} (l(x^n) - \min_{\theta \in \Theta_k}(-\log_2 P(X^n|\theta)))$$

を**ミニマックスリグレット**と呼ぶ。このミニマックスリグレットは**正規化最尤符号化**によって実現される。

第1章のまとめ

- 平均符号長最小化と尤度最大化はどちらも同じ計算である。

  さらに符号長を最小化することでモデルの複雑さを抑えることができるので、学習問題においては平均符号長が最小となるようなモデルを選ぶとよい。

- 真の分布が特定できている時、**Shannonの情報量**が最短符号長となる

- 真の分布が特定できていない時、**確率的コンプレキシティ**が最短符号長となる

  正規化最尤符号化が効率的に圧縮できる符号化である。

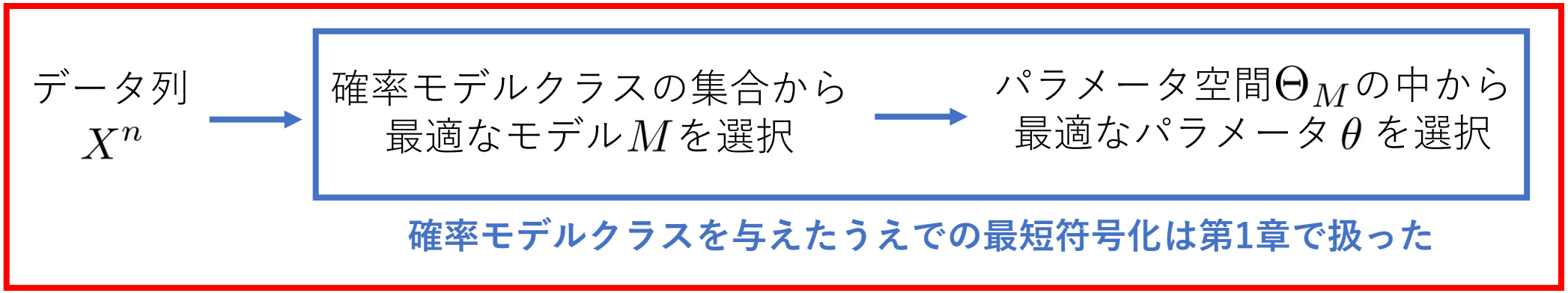
  厳密には語頭符号化符号長の下限とO(1)の誤差を持っているが、**ミニマックスリグレット**を実現する符号化であることから正当化できる。

第2章

一括学習とモデル選択

一括学習とMDL規準

一括学習



確率モデルの選択も含めた最短符号化を第2章で考える

MDL規準

データ列 X^n が与えられている状態で確率モデルMのモデル評価規準

$$\mathcal{L}(X^n : M) = \underline{\mathcal{L}(X^n|M)} + \underline{\mathcal{L}(M)}$$

確率的コンプレキシティに等しい　　**モデル自体の語頭符号化符号長**

ヒストグラムの学習

ヒストグラムを用いて母集団の確率分布を求めたい場合を考える

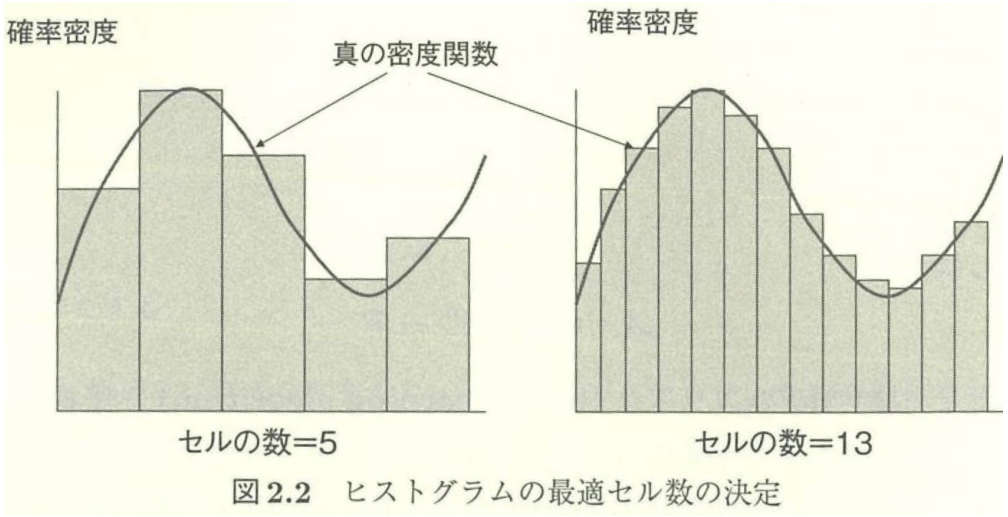


図 2.2　ヒストグラムの最適セル数の決定

セル数 k として各セルの確率密度関数の値

$$\theta = (\theta_1, \ldots, \theta_k)$$

を求める問題

モデル選択問題　　セル数 k の決定問題

パラメータ推定問題　θ の決定問題

セル数 k の符号化には $\log_2 2.865 + \log_2 k + \log_2(\log_2 k) + \cdots$ の符号長が必要なので

$$\min_{\theta}(-\log_2 P(x^n|\theta)) + \frac{k}{2}\log_2 \frac{n}{2\pi} + \log_2 \int \sqrt{|I(\theta)|} + \log_2 2.865 + \log_2 k + \log_2(\log_2 k) + \cdots$$

を最小化する k が最も良いセル数であることがわかる

有限分割型確率的規則の学習

定義域を有限個の領域にわけ、各領域ごとに確率を推定する問題

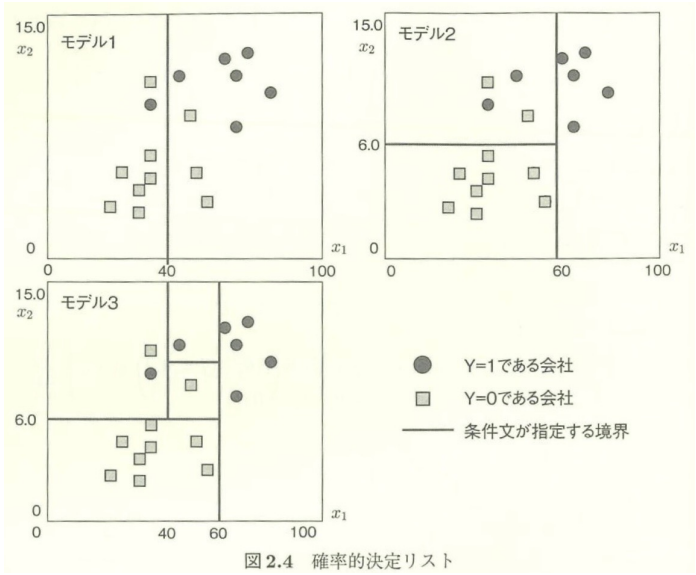


図 2.4　確率的決定リスト

表 2.2　モデル 1,2,3 の記述長の比較

	データ符号長	モデル符号長	総符号長
モデル 1	16.3498	4	20.3498
モデル 2	7.9867	8	15.9867
モデル 3	5.4867	16	21.4867

x 個の説明変数がある定義域 $\mathcal{X}(\subset \mathbb{R}^x)$ と値域 $\mathcal{Y} = \{0, 1 \cdots, m\}$ からなるデータ $(X, Y)^n$

X が領域 C_i に入ったとき $Y = j$ である確率 $\theta_j^{(i)}$ を求める問題

モデル選択問題　　領域 C_i の決定問題

パラメータ推定問題　$\theta_j^{(i)}$ の決定問題

たとえば $X_1 \geq 40$ という領域であれば

- X_1 を X_1, X_2 から選ぶのに1bit
- $\geq$ を $\geq, \leq$ から選ぶのに1bit
- 40を20,40,60,80から選ぶのに2bit

MDL学習の収束速度

データ列 x^n から最尤推定された確率分布 $\hat{P_{[x^n]}}(X)$ と真の分布 $P^*(X)$ の距離を次のように決める

$$d_H(P^*, \hat{P}_{[x^n]}) = \sum_X \left(\sqrt{P^*(X)} - \sqrt{\hat{P}_{[x^n]}(X)} \right)^2$$

定理5：MDL学習の収束速度

i) 真の分布が確率モデルクラス $\mathcal{P}_k$ に含まれている場合

$$E_{P^*}[d_H(P^*, \hat{P}_{[x^n]})] = O\left(\frac{k \log_2 n}{n}\right) \xrightarrow{n\to\infty} 0$$

ii) 真の分布と確率モデルクラス $\mathcal{P}_k$ の誤差が $\inf_{Q\in\mathcal{P}_k} E_{P^*}\left[\log_2 \frac{P^*(X)}{Q(X)}\right] = O\left(\frac{1}{k^\alpha}\right)$ である場合

$$E_{P^*}[d_H(P^*, \hat{P}_{[x^n]})] = O\left(\left(\frac{\log_2 n}{n}\right)^{\frac{\alpha}{\alpha+1}}\right) \xrightarrow{n\to\infty} 0$$

MDL学習の収束速度

有限分割型確率的規則のMDL学習の収束速度についてはより厳密に計算できる

定理6：有限分割型確率的規則の収束速度

データ列$D_i = (X_i, Y_i)\ (i = 1, \ldots, n)$を$k^*$次の有限分割型の確率的規則$M^*$で推定する。このとき$\lambda \geq 2$ならば、$M^*$の符号化を$l_n(M^*)$とすると任意の$\varepsilon > 0$で以下が成立。

$$Prob[d_H(P^*, \hat{P}_{[D^n]}) > \varepsilon] < \exp\left[-\frac{n\varepsilon}{2} + \left(\frac{\lambda l_n(M^*)}{2} + \frac{k^*}{2}\right)\ln 2\right]$$

また、$Prob[d_H(P^*, \hat{P}_{[D^n]}) > \varepsilon] < \delta$となるサンプル数$n_{MDL}$については以下が成立する。

$$n_{MDL} = O\left(\frac{k^*}{\varepsilon}\log_2\frac{k^*}{\varepsilon} + \frac{\mathcal{L}(M^*)}{\varepsilon} + \frac{1}{\varepsilon}\log_2\frac{1}{\varepsilon}\right)$$

MDL学習で推定した分布はnが大きくなるにつれて真の分布に収束する

AICとの比較

データX^nから最尤推定されたパラメータ$\hat{\theta}(X^n)$を考える。未知のデータZに対して対数尤度の期待値$E_Z[-\log_2 P(Z|\hat{\theta})]$は**モデルが未知のデータをどれだけ予測できるか**を表す。

定義：赤池情報量規準(AIC)

期待対数尤度のデータ数倍 $nE_Z[-\log_2 P(Z|\hat{\theta})]$ のデータ列 X^n に対する不偏推定量は

小さいほど予測が現実に近い

$$AIC(X^n) = -\log_2 P(X^n|\hat{\theta}) + k$$

パラメータ数に対する罰則項の役割

と表せ**赤池情報量規準**と呼ぶ。すなわち以下が成立。

$$E_{X^n}[nE_Z[-\log_2 P(Z|\hat{\theta}(X^n))]] = E_{X^n}[AIC(X^n)] = E_{X^n}[-\log_2 P(X^n|\hat{\theta}) + k]$$

MDL規準も赤池情報量規準もモデル選択を評価するための規準である

AICとの比較

何を目的にモデルを選択するか（**メタ基準**）をふまえて使い分ける必要がある

赤池情報量規準(AIC)

次に得られるデータに対して近い予測ができるモデルを良く評価している

データ数nが無限になっても**推定分布が真の分布に近づくとは限らない**

（データ数が少ない時はパラメータ数の少ないモデルを選びデータ数が増えるとパラメータ数の多いモデルを選ぶ傾向にある）

MDL規準

データ数nが無限になると推定分布が真の分布に近づく

収束するまでのサンプル数を概算できる点で有効

たとえば…

地域Aの行動モデルを地域Bに転用したいときはAICよりもMDL規準のほうが有効
（真の分布が必要）

第2章のまとめ

- 確率的コンプレキシティにモデル自身の符号長を足した**MDL規準**を用いてモデル選択を行うことができる

  モデルの符号長は利用するモデルに合わせて個別に計算する

- MDL規準をもとにした学習はデータ数が増えるにつれて真の分布に近づく

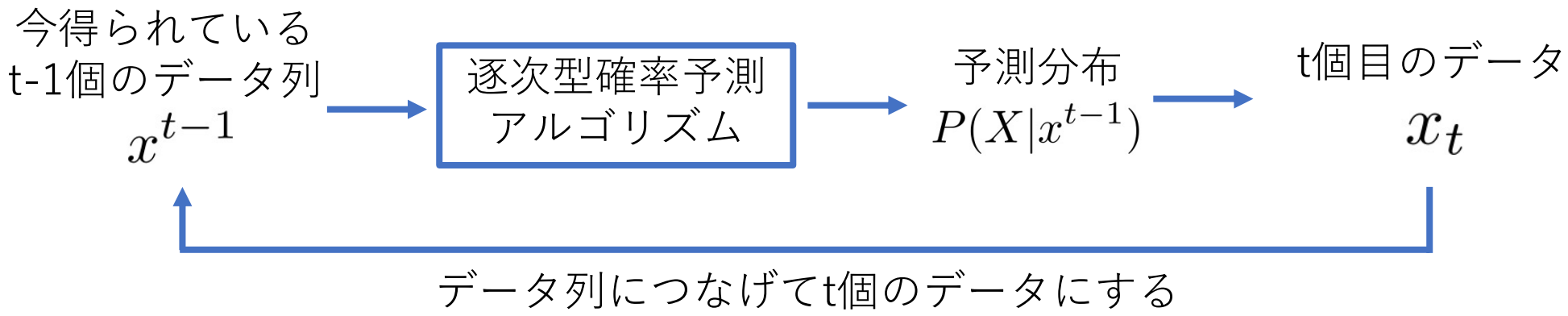
  未知データの予測精度が高いモデルを選ぶ一方で真の分布を選べるとは限らない
  赤池情報量規準(AIC)とうまく使い分けることが求められる

第3章

逐次符号化と逐次的予測

逐次型確率予測アルゴリズム

逐次型確率予測アルゴリズム



累積対数尤度

$$\sum_{t=1}^{n}\left(-\log_2 P_A(x_t|x^{t-1})\right)$$

冗長度

$$R_n(A:\mathcal{P}_k,P^*)$$

$$= E_{P^*}\left[\sum_{t=1}^{n}-\log_2 P_A(X_t|X^{t-1})\right] - \inf_{\theta\in\Theta_k} E_{P^*}\left[\sum_{t=1}^{n}-\log_2 P(X_t|X^{t-1}:\theta)\right]$$

逐次符号化の符号長

一括学習したときの符号長

最尤予測アルゴリズム

データ列 x^{t-1} が与えられたときの最尤推定値を $\hat{\theta}_{t-1}$ とする。各時点の予測分布を

$$P_{ML}(X|x^{t-1}) = P(X|\hat{\theta}_{t-1})$$

とするアルゴリズムを**最尤予測アルゴリズム**という。

定理7：最尤予測アルゴリズムの冗長度

真の分布 $P^*(X^n)$ に従って生成されるデータ列について、確率モデルクラス $\mathcal{P}_k$ の最尤推定値が真のパラメータを平均とした正規分布に従うように現れると仮定すると以下が成立。

$$E_{P^*}\left[\sum_{t=1}^{n} -\log_2 P(X_t|\hat{\theta}_{t-1})\right] = E_{P^*}\left[-\log_2 P^*(X^n)\right] + \frac{k}{2}\log_2 n + o(\log_2 n)$$

最尤予測アルゴリズムの累積対数損失を**予測的確率的コンプレキシティ**と呼び、同じデータの確率的コンプレキシティとはo(log n)の誤差で一致する。

ベイズ予測アルゴリズム

データ列 x^{t-1} が与えられたとき、予測分布をベイズの定理より

$$P_{Bayes}(X|x^{t-1}) = \int P(\theta|x^{t-1})P(X|x^{t-1}:\theta)d\theta$$

$$P(\theta|x^{t-1}) = \frac{\pi(\theta)P(x^{t-1}|\theta)}{\int \pi(\theta')P(x^{t-1}|\theta')d\theta'}$$

とするアルゴリズムを**ベイズ予測アルゴリズム**という。

定理8：ベイズ予測アルゴリズムの冗長度

定理7と同じ条件でベイズ予測アルゴリズムの冗長度は

$$R_n(A_{Bayes}, \mathcal{P}_k, P^*) = \frac{k}{2}\log_2\frac{n}{2\pi e} + \log_2\frac{\sqrt{|I(\theta)|}}{\pi(\theta)} + o(1)$$

また、冗長度の最悪値 $\sup_{P^*\in\mathcal{P}_k} R_n(A_{Bayes}, \mathcal{P}_k, P^*)$ の下限は **Jeffreysの事前分布** $\pi(\theta) = \frac{\sqrt{|I(\theta)|}}{\int\sqrt{|I(\theta)|}d\theta}$ を事前分布としたベイズ予測アルゴリズムによって実現される。**←一番良いアルゴリズム！**

逐次型正規化最尤予測アルゴリズム

このアルゴリズムでは真の分布が確率モデルクラス$\mathcal{P}_k$上にあることを仮定しない。

各時点での予測分布は

$$P_{SNML}(x|x^{t-1}) = \frac{P(x \cdot x^{t-1}|\hat{\theta}(x \cdot x^{t-1}))}{\int P(x \cdot x^{t-1}|\hat{\theta}(x \cdot x^{t-1}))dx}$$

（t個目のデータについて正規化最尤分布を求める）

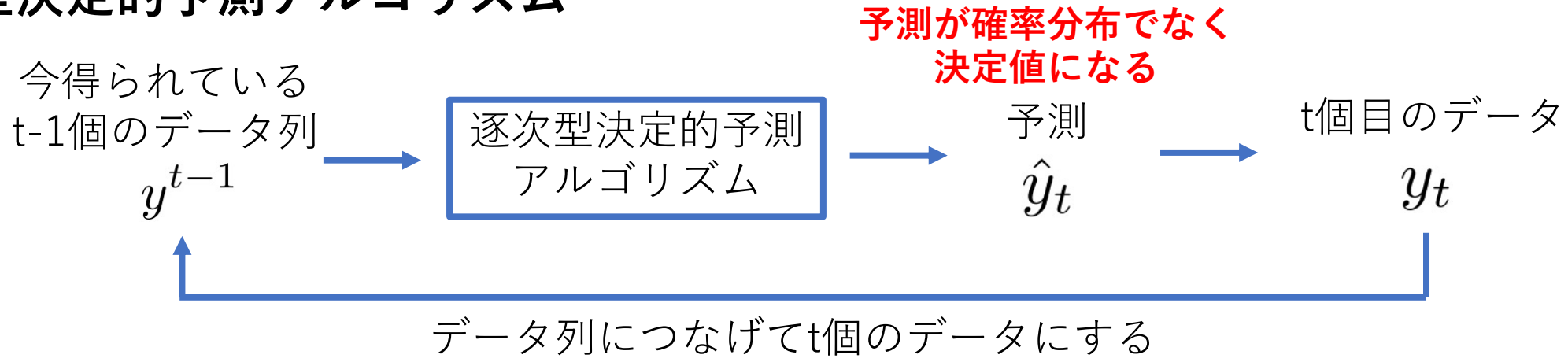
真の分布が確率モデルクラス$\mathcal{P}_k$にあるとは限らないため冗長度を求めることができず、**条件付きミニマックスリスク**によって正当化される。

$$\min_{Q(x|x^{t-1})} \max_{x} \{-\log_2 Q(x|x^{t-1}) - (-\log_2 P(x|x^{t-1} : \hat{\theta}(x \cdot x^{t-1})))\}$$

の最小値を実現するアルゴリズムになっている。

逐次型決定的予測アルゴリズム

逐次型決定的予測アルゴリズム



累積予測損失

$$\sum_{t=1}^{n} L(y_t, \hat{y}_t)$$

誤差関数が一般化される

冗長度

$$R_n(A, \mathcal{F}, P^*)$$

$$= E_{P^*}\left[\sum_{t=1}^{n} L(Y_t, \hat{Y}_t)\right] - \inf_{f \in \mathcal{F}} E_{P^*}\left[\sum_{t=1}^{n} L(Y_t, f(Y^{t-1}))\right]$$

逐次的学習の累積損失

一括学習したときの累積損失

一般化最尤予測アルゴリズム

Step1：最尤推定を用いてデータの確率分布を求める

データ列y^{t-1}に合うような確率モデルクラス$\mathcal{P}_k$をうまく見つけて

$$P(y^{t-1}|\hat{\theta}_{t-1}) = \max_{\theta} P(y^{t-1}|\theta)$$

Step2：以下で定めるL-変換により予測値を出力する

データy_tの確率分布が$P(y_t|\hat{\theta}_{t-1})$と求まったので

$$\hat{y} = \arg\min_{y} E_{P(y_t|\hat{\theta}_{t-1})}[L(Y, y)]$$

$L(y, P) = -\log_2 P(y)$と誤差関数を置くと最尤予測アルゴリズムと一致する。

最尤予測アルゴリズムの誤差関数を一般化したアルゴリズムである

集合型予測アルゴリズム

$\mathcal{F} = \{A_i : i = 1, \ldots, N\}$：有限個の逐次型予測アルゴリズムの集合

Step1：重みを初期化する

各アルゴリズムに対して$w_{0,i}(i = 1, \ldots, N)$を$w_{0,i} > 0$で任意に初期化する

Step2：時間ごとに繰り返し計算をする

各時間ごとに

$$W_t = \sum_{i=1}^{N} w_{t,i}$$

$$v_{t,i} = \frac{w_{t,i}}{W_t}$$

$$\Delta_t(y) = -\frac{1}{\lambda} \log \left(\sum_{i=1}^{N} v_{t,i} \exp(-\lambda L(y, \hat{y}_{t,i})) \right)$$

$$\hat{y}_{t,i} = \frac{1}{2} \left(L_0^{-1}(\Delta_t(0)) + L_1^{-1}(\Delta_t(1)) \right)$$

$$w_{t+1,i} = w_{t,i} \exp(-\lambda L(y, \hat{y}_{t,i}))$$

$$L(y, P) = -\log_2 P(y)$$

と誤差関数を置き$w_{0,i}$を事前分布とみなすと**ベイズ予測アルゴリズム**と一致する

第3章のまとめ

- 逐次的な確率予測で使われるアルゴリズムは3つ

  最尤予測アルゴリズム

  ベイズ予測アルゴリズム　← **確率モデルクラスの中に真の分布がある場合は一番良い**

  逐次型正規化最尤予測アルゴリズム　← **確率モデルクラスの中に真の分布がない時に使う**

- 逐次的な決定値予測で使われるアルゴリズムは2つ

  一般化最尤予測アルゴリズム　← **最尤予測アルゴリズム**の一般化

  集合型予測アルゴリズム　← **ベイズ予測アルゴリズム**の一般化

参考文献

- 山西健司. (2010). 情報論的学習理論. 共立出版.
- 山西健司. (1996). 確率的コンプレキシティと学習理論. オペレーションズリサーチ, 41(7), 379–386.
- 山西 健司. (2016). 記述長最小原理の進化：基礎から最新の展開, 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, 10 巻, 3 号, p. 186-194
- 粕谷 英一. (2015). 生態学におけるAICの誤用 : AICは正しいモデルを選ぶためのものではないので正しいモデルを選ばない(<特集2>生態学におけるモデル選択), 日本生態学会誌, 65 巻, 2 号, p. 179-185