情報論的学習理論入門 〜「機械学習 = データ圧縮」のアプローチ

Author: https://github.com/F-ridge

本記事は、物工/計数 Advent Calendar 2019の7日目の記事です。

はじめに

機械学習の理論へのアプローチとして、統計的学習理論・計算論的学習理論・情報論的学習理論、などといったものが挙げられます。それぞれの学習理論では、学習をどう定義し、学習を成功させるためのアルゴリズムをどう設計し、それをどう評価し、その限界はどれくらいであるか、といった考察を、異なる立場から行っています。

本記事では、「情報論的学習理論」という、情報理論の立場からの機械学習へのアプローチの紹介を行います。

ちなみに、東京大学計数工学科数理情報工学コース、及び、同大学大学院情報理工学研究科数理情報学専攻には、統計的学習理論を専門とされる教員[1]や情報論的学習理論を専門とされる教員[2]が在籍されています。私は今学期に情報論的学習理論の教員の講義を聴講しており、現時点での学習のまとめをしてみようかと思ったため、アドベントカレンダーの記事にすることにしました。

本記事は、「情報論的学習理論」の講義[3]内容 (の前半) や、講義で指定されている教科書[4][5][6]を参照しながら作成しました。

概要

情報理論で議論されることの中に、データ圧縮が挙げれられます。

情報論的学習理論では「学習」を、情報理論における「データ圧縮」とみなし、「データとモデル自身を合わせて、これらを最も小さく圧縮できるモデルが最良である」という立場から議論が行われます。「圧縮」は記述長を短くすることであるために、この考え方は**記述長最小原理** (Minimum Description Length 原理、**MDL原理**) と呼ばれています。

本記事の後半では、「学習=データ圧縮」とみなしたときの学習の基準 (情報量基準) は何か？ その基準にそって学習すると何が嬉しいのか？ 逆に欠点はあるのか？ といったことも紹介していきます。

なお、本記事では、対数の底は 2 としています。

情報理論の前提知識

まず、この節では、本記事で必要となる、情報理論の知識を紹介します。

語頭符号化

定義 (符号化 coding)

χ を、有限種類のシンボルからなるアルファベットとする。$\{0,1\}^*$を、0 と 1 からなる、任意の長さの系列の集合とする。写像$\pi : \chi^n \to \{0,1\}^*$ を符号化と呼ぶ。

定義 (語頭符号化 prefix coding)

符号化 π について、任意の $x_1, x_2 \in \chi^n$ に対して、$\pi(x_1)$ と $\pi(x_2)$ の一方が他方の先頭部分に一致することがないとき、π を語頭符号化と呼ぶ。

例

$\chi = \{a, b, c, d\}$ に対して、

$\pi(a) = 1, \pi(b) = 01, \pi(c) = 001, \pi(d) = 000$ とすると π は語頭符号化。

たとえば、符号化された文が 001101000101 だったとき、文の先頭から順に、上の4つのパターンに一致しないか見ていくことで、$001\ 1\ 01\ 000\ 1\ 01$ というパターンが連なっていたということが一意にわかります。そのため、符号化される前の、もとの文章は $cabdab$ だったということを一意に特定することができます。しかも、頭から順に文字を1度ずつ見るだけで復号ができます。

このように、語頭符号化によって符号化された文は、文中に区切り記号がなくても、先頭から順に文を見ていくことで一意に、かつ高速に復号することができます。

Kraftの不等式

定義

π を符号化関数とする。$x \in \chi$ に対して、符号長を $l(x) = |\pi(x)|$ (符号化された後の系列の文字数) と表す。

定理 (Kraftの不等式とその条件)

$$\text{語頭符号化 } \pi \text{ が存在} \Leftrightarrow \sum_{x^n \in \chi^n} 2^{-l(x^n)} \leq 1$$

証明は文献[7]を参照ください。

この不等式は再登場します。情報理論で考えている符号長の最小化が、機械学習においてはどのような意味を持つのかを次章で解説します。

学習理論への繋がり

Kraftの不等式を頭に入れておいて、機械学習の話に入っていきましょう。前章で紹介した情報理論が、機械学習にどう生かされるかを説明していきます。

ここでの前提知識として、統計学で出てくる最尤推定量について説明します。

最尤推定量

モデルのパラメータ推定をする際に、何かしらの尺度において「良い」推定量を選びたい、という気持ちになります。そこで、データが観測されたときに、「尤もらしさ」、すなわち、そのデータが観測される確率を、モデルのパラメータの関数として表し、この関数が最大値をとるときのパラメータを推定量として採用する、という手法が考えられます。この関数を尤度と呼び、このようにして採用されたパラメータを最尤推定量と呼びます。

確率分布やパラメータの次元は決めており、あとはそのパラメータを推定したい、という状況を想定し、式で表すと以下のようになります。

データ列 $x^n = x_1 x_2 \dots x_n \in \chi^n$: 既知とする。

モデルクラス (推定の候補となるモデル) は $\mathcal{P} = \{p(x^n;\theta) | \theta \in \Theta \subset \mathbb{R}^k\}$ となる。

ただし、p は確率関数、Θ はパラメータ θ がとりうる値の集合、k はパラメータの次元とする。

尤度関数は $\mathcal{L}(\theta) = p(x^n;\theta)$ であり、

最尤推定量は

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \mathcal{L}(\theta) \\
&= \arg\max_{\theta} p(x^n;\theta) \\
&= \arg\min_{\theta} \{-\log p(x^n;\theta)\}
\end{aligned}
$$

と表される。

最尤推定量の嬉しい性質として、一致性 (推定値が真の値に確率収束)、漸近正規性 (推定値の分布が正規分布に確率収束)、漸近有効性 (推定値の分散はCramér–Raoの不等式の下限を達成) が挙げられます。

なお、$-\log p(x^n;\theta)$ は負の対数尤度と呼ばれ、$p(x^n;\theta)$ の形のままで扱うよりも計算がやりやすくなることが往々にしてあります。

また、この $-\log p(x^n;\theta)$ の形で尤度を表すことで、情報理論における記述長の最小化と、尤度の最大化の対応をはっきりと確認できるようになります。このことを以下で説明します。

符号長最小化と尤度最大化の対応

もう1つ、新たに言葉が必要なので定義を述べます。

定義 (劣確率分布 subprobability distribution)

写像 $p : \chi^n \to [0,1]$ が、任意の$x^n \in \chi^n$ に対して $p(x^n) \geq 0$ を満たし、$\sum_{x^n \in \chi} p(x^n) \leq 1$ を満たすとき、p を χ^n 上の劣確率分布と呼ぶ。

※不等式が等号で成り立つとき、すなわち $\sum_{x^n \in \chi} p(x^n) = 1$ のときには p は確率関数になります。

※ダミーの x を設定して、$x = 1 - \sum_{x' \in \chi^n} x'$ とすることで確率分布を構成できるため、以下では通常の確率分布として扱います。

さて、この「全部足して1以下」という形の不等式、他にも登場していました。Kraftの不等式 $\sum_{x^n \in \chi^n} 2^{-l(x^n)} \leq 1$ です。そこで、シグマで足しているものに注目して、$P(x^n) = 2^{-l(x^n)}$ と対応させることで、語頭符号化と劣確率分布は1対1に対応していることがわかります。また、$P(x^n) = 2^{-l(x^n)} \Leftrightarrow l(x^n) = -\log P(x^n)$ となります。左辺は符号長、右辺は負の対数尤度となっています。したがって、このような劣確率分布 P を考えれば、**符号長の最小化と尤度の最大化は等価になる**ことがわかります。

そして情報量基準へ

次に、パラメータ推定に限らず、モデル選択も行う状況を考えてみましょう。一般に、モデルはパラメータの次元よりも広い概念ですが、本記事では簡単のために、モデルのパラメータの次元を決定する問題を考えます。すなわち、パラメータの次元を k として、モデルクラスが

$$\mathcal{P} = \bigcup_k \mathcal{P}_k, \quad \mathcal{P}_k = \{p(x^n; \theta, k) | \theta \in \Theta_k \subset \mathbb{R}^k\}$$

と書ける場合を考えます。

このとき、パラメータの次元 k を大きくしていくとモデルの表現力は高くなり、与えられたデータへの当てはまりは良くなっていきますが、k が大きすぎると未知のデータに対してはかえって当てはまりが悪くなる（過学習）というトレードオフの関係があります。そこで、適切な大きさの k を選ぶ必要が出てきます。

モデル選択のための道具として、情報量基準やクロスバリデーション[8]が挙げられますが、本記事では情報量基準に絞って説明していきます。

情報量基準とは、モデルに関する最適化の基準のことで、これを最大化、または最小化するモデルを採用する、という方針でモデルを選びます。情報量基準にはAIC、BIC[9]などさまざまなものがあります。それぞれの情報量基準は異なる性質を持ち、異なる情報量基準を用いれば異なるモデルが選択されます。本記事では典型的な情報量基準としてAICを紹介した後、「データとモデル自身を合わせて、これらを最も小さく圧縮できるモデルが最良である」という立場、すなわち**MDL** (minimum description length) **原理**の立場から導かれる情報量基準である、MDL基準という情報量基準を紹介します。

AIC

典型的な情報量基準の例として、AIC (Akaike's information criterion) を紹介します。AICは次の式で与えられる情報量基準であり、これを最小化するモデルが選択されます。

定義 (AIC)

$x^n = x_1 \dots x_n$: データ列、k : パラメータの次元、$\hat{\theta}(x^n)$: パラメータの次元が k のときの、パラメータの x^n からの最尤推定量とすると、

$$\mathrm{AIC}_k(x^n) = -\log p(x^n; \hat{\theta}(x^n), k) + k$$

$-\log p(x^n;\hat{\theta}(x^n),k)$ は負の対数尤度なので、この項はモデルの当てはまりの良さであると解釈でき、k はモデルの複雑さに対するペナルティと解釈できます。

AICは、期待平均対数尤度 $nE_{X^n}E_{x^n}[-\log p(X;\hat{\theta}(x^n),k)]$ の不偏推定量として導出されたものです。ただし、X^n : 未知のデータとしています。そのため、AICを最小にするということは期待平均対数尤度の不偏推定量を最小にするということと等価であり、AICの持つ性質の1つとしてこのことが挙げられます。一方で、AICは一致性を持たない (k の候補に真の値 k^* が含まれていたとしても、AICで選ばれた $\hat{k}$ はデータ数 n を大きくしたときに $\hat{k}$ に一致 (確率収束) しない) ことが知られています。そのため、機械学習で達成したい目標によっては、情報量基準にAICを用いるのが最良とは限らないことがわかります。

MDL基準

「データとモデル自身を合わせて、これらを最も小さく圧縮できるモデルが最良である」という立場、すなわち **MDL** (minimum description length) **原理**の立場から導出される情報量基準として、MDL基準が存在します。これは、次の表式で与えられます。

$$\mathrm{MDL}_k(x^n) = -\log p(x^n;\hat{\theta}(x^n),k) + \log\sum_{X^n} p(X^n;\hat{\theta}(X^n),k) + l(k)$$

この量を最小化する k を選ぶ、という基準です。ここで、X^n は未知のデータ、$\hat{\theta}(x^n)$ は最尤推定量、$l(k)$ は k の符号長としています。

MDL基準の表式における $\log\sum_{X^n} p(X^n;\hat{\theta}(X^n),k)$ はパラメトリックコンプレキシティと呼ばれる量で、モデルクラス $\mathcal{P}_k = \{p(x^n;\theta,k)|\theta\in\Theta_k\subset\mathbb{R}^k\}$ の情報論的な複雑さを表しています。一般にパラメトリックコンプレキシティの計算は困難ですが、漸近的な近似式を用いて計算することができます[10]。結論だけ示すと、いくつかの前提のもとで、

$$\log\sum_{X^n} p(X^n;\hat{\theta}(X^n),k) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int\sqrt{|I(\theta)|}d\theta + o(1)$$

が成り立ちます。ここで、$I(\theta)$ はFisher情報行列で、その (i,j) 成分は $\lim_{n\to\infty}\frac{1}{n}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x^n;\theta,k)\right]$ で与えられます。$|I(\theta)|$ は、Fisher情報行列の行列式です。また、$o(1)$ は、$\lim_{n\to\infty} o(1) = 0$ となる量です。上の式が成り立つ前提は以下のとおりです。

- θ に関して中心極限定理が成り立つ。つまり、x^n からの最尤推定量 $\hat{\theta}$ に対して、$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{p} N(0,I^{-1}(\theta))$ (正規分布)
- $I(\theta)$ が θ に関して連続で、$\int\sqrt{|I(\theta)|}d\theta < \infty$

この漸近近似式の証明は文献[6]を参照ください。

MDL基準の性質

MDL基準がもつ性質のうち、いくつかについて説明します。

ミニマックスリグレットに関する最適性

$\mathrm{MDL}_k(x^n)$ は、データの従う確率分布が未知のときに、ミニマックスリグレットという量の下限を達成する符号長になっています。このことをこの節で説明します。

定義 (Minimax regret)

モデルクラス $\mathcal{P}_k = \{p(x^n;\theta,k)|\theta \in \Theta_k \subset \mathbb{R}^k\}$ に対するミニマックスリグレットは以下の式で与えられる。

$$\min_q \max_{x^n} \left\{ -\log q(x^n) - \min_\theta \left(-\log p(x^n;\theta,k)\right) \right\}$$

ただし q は確率分布とする。

先の「符号長最小化と尤度最大化の対応」で述べたように、$-\log q(x^n)$ は語頭符号長に対応します。この、ミニマックスリグレットという量は、確率分布 q を (任意に) もってきたときに、q に対応する符号化による符号長が、モデルクラス $\mathcal{P}_k$ を用いた時の最小符号長に対して最悪の場合でどれだけ大きくなるかを評価し、q に関して最小値をとったものです。

ミニマックスリグレットの下限は、正規化最尤分布という確率分布で達成されます。まずは、正規化最尤分布の定義を記します。

定義 (正規化最尤分布、normalized maximum likelihood 分布、NML分布)

$$p_{\mathrm{NML}}(x^n) = \frac{p(x^n;\hat{\theta},k)}{\sum_{X^n} p(X^n;\hat{\theta},k)}$$

定理 (MDL基準のミニマックスリグレットに関する最適性)

ミニマックスリグレットの下限は、NML分布により達成される。

(証明) 背理法で示す。

NML分布 p_{NML} 以外の確率分布 $\tilde{p}$ により、ミニマックスリグレットの下限が達成されたとする。このとき、$\exists x_*^n \in \chi^n,\ \ \tilde{p}(x_*^n) < p_{\mathrm{NML}}(x_*^n)$ であり、この x_*^n について、

$$\begin{aligned}
&\min_q \max_{x^n} \left\{ -\log q(x^n) - \min_\theta \left(-\log p(x^n;\theta,k)\right) \right\} \\
&\geq \max_{x^n} \left\{ -\log \tilde{p}(x^n) - \min_\theta \left(-\log p(x^n;\theta,k)\right) \right\} \\
&\geq -\log \tilde{p}(x_*^n) - \min_\theta (-\log p(x_*^n;\theta,k)) \\
&> -\log p_{\mathrm{NML}}(x_*^n) - \min_\theta (-\log p(x_*^n;\theta,k)) \\
&= \left\{ -\log p(x^n;\hat{\theta}(x^n),k) + \log \sum_{X^n} p(X^n;\hat{\theta}(X^n),k) \right\} + \log p(x_*^n;\hat{\theta}(x_*^n),k)) \\
&= \log \sum_{X^n} p(X^n;\hat{\theta}(X^n),k)
\end{aligned}$$

が成り立つ。一方で、

$$
\begin{aligned}
&\min_{q}\max_{x^n}\left\{-\log q(x^n)-\min_{\theta}\left(-\log p(x^n;\theta,k)\right)\right\}\\
&\leq \max_{x^n}\left\{-\log p_{\mathrm{NML}}(x^n)-\min_{\theta}(-\log p(x^n;\theta,k))\right\}\\
&=\max_{x^n}\left[\left\{-\log p(x^n;\hat{\theta}(x^n),k)+\log\sum_{X^n}p(X^n;\hat{\theta}(X^n),k)\right\}+\log p(x_*^n;\hat{\theta}(x_*^n),k))\right]\\
&=\log\sum_{X^n}p(X^n;\hat{\theta}(X^n),k)
\end{aligned}
$$

でもあるため、ふたつの不等式は矛盾する。よって、ミニマックスリグレットの下限を達成できる確率分布はNML分布 p_{NML} に限る。

NML分布は χ^n 上で確率分布をなしています。NML分布を構成する際のお気持ちを簡単に説明します。

先の「符号長最小化と尤度最大化の対応」で述べたように、尤度最大化と語頭符号長の最小化は表裏一体の関係にあります。そこで、最尤推定量 $\hat{\theta}(x^n)$ を使って $l(x^n)-\log p(x^n;\hat{\theta}(x^n),k)$ の符号長での符号化ができないだろうか、という気持ちになります。しかし、この符号長での語頭符号化は実現できません。なぜならば、$\hat{\theta}(x^n)$ は各 x^n に関して尤度が最大になるように選ばれているために、
$\sum_{x^n}2^{l(x^n)}=\sum_{x^n}p(x^n;\hat{\theta}(x^n),k)>1$ となり、Kraftの不等式を満たさない (あるいは、$p(x^n;\hat{\theta}(x^n),k)$ は χ^n 上での確率分布にならない) ためです。そこで、$\sum_{X^n}\max_{\theta}p(X^n;\theta,k)$ で割って正規化することで確率分布として使えるようにしています。

NML分布の符号長は

$$
\begin{aligned}
l_{\mathrm{NML}}(x^n;k)&=-\log\frac{p(x^n;\hat{\theta},k)}{\sum_{X^n}p(X^n;\hat{\theta},k)}+l(k)\\
&=-\log p(x^n;\hat{\theta}(x^n),k)+\log\sum_{X^n}p(X^n;\hat{\theta}(X^n),k)+l(k)
\end{aligned}
$$

となります。つまり、NML分布の符号長、すなわちミニマックスリグレットの下限を達成する符号長を最小化しなさい、というのがMDL基準だったというわけです。

ちなみに、NML分布の符号長は、「データ列 x^n の、モデルクラス $\mathcal{P}_k$ に対する**確率的コンプレキシティ**」と呼ばれています。

一致性

AICは一致性をもちませんでしたが、MDLは一致性をもちます。このことを、平均符号長の下限式から説明していきます。

定理 (平均符号長の下限: 真の分布が未知の場合)

真の分布は未知、すなわち θ, k の真の値は未知とする。また、θ に関しての中心極限定理、すなわち、x^n からの最尤推定量 $\hat{\theta}$ に対して、$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{p}N(0,I^{-1}(\theta))$ の成立を仮定する。このとき、$\forall l$: 語頭符号長関数、$\forall\epsilon>0$ に対して、データ数 $n\to\infty$ でLebesgue測度が 0 になるような集合を除いて次式が成立する。

$$
E_{\theta}[l(x^n)]\geq H_n(p)+\frac{k-\epsilon}{2}\log n
$$

ここで、E_{θ} は $p(X^n;\theta,k)$ (X^n: 未知データ) に関する期待値を表し、$H_n(p)$ はエントロピー、すなわち、$H_n(p)=E_{\theta}[-\log p(X^n;\theta,k)]$ である。

証明は文献[6]を参照ください。

定理 (MDL推定モデルの一致性)

以下を仮定する。

- θ に関して中心極限定理が成り立つ。つまり、x^n からの最尤推定量 $\hat{\theta}$ に対して、
  $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{p} N(0, I^{-1}(\theta))$ (正規分布)
- $I(\theta)$ が θ に関して連続で、$\int \sqrt{|I(\theta)|}d\theta < \infty$

このとき、真の k を k^* とすると、$\lim_{n\to\infty} P(\hat{k}(x^n) = k^*) = 1$ が成り立つ。

ここで、$\hat{k}(x^n)$ は、k の、x^n からのMDL基準による推定値である。

(なんとなくの証明)

この仮定のもとでは、$\mathrm{MDL}_k = -\log p(x^n; \hat{\theta}(x^n), k) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int\sqrt{|I(\theta)|}d\theta + o(1)$ と計算できる。そのため、データ数 n が十分大きいとき、MDL基準による推定量 $\hat{k}(x^n)$ は、
$E_\theta[-\log p(X^n; \hat{\theta}(X^n), k)] + \frac{k}{2}\log n$ を最小化したものに近づく。一方で、平均符号長の下限式より、平均符号長は $E_\theta[-\log p(X^n; \theta, k)] + \frac{k}{2}\log n$ により下から抑えられる。よって、$\hat{k}(x^n)$ は、真の値 k^* と漸近的に等しくなる。

注 (「平均符号長の下限: 真の分布が未知の場合」の補足)

「平均符号長の下限: 真の分布が未知の場合」の下限式ですが、θ が偶然 $l(x^n) = -\log p(x^n; \theta, k)$ となるように選ばれると下限はより小さくなります。しかし、このような「偶然」は確率 0 でしか起きません。定理で「Lebesgue測度が 0 になるような集合を除いて」とことわっているのはこのためです。

ちなみに、真の分布が既知の場合の下限式は、次の定理から得られます。

定理 (平均符号長の下限: 真の分布が既知の場合)

p : 真の分布とし、x^n はこの分布から生成されているとする。このとき、任意の語頭符号長関数 l に関して次が成り立つ。

$$E_p[l(x^n)] \geq H_n(p)$$

ここで、E_p は x^n の発生についての p に関する期待値を表す。

(証明)

確率分布 q を任意にひとつもってきて、$l(x^n) = -\log q(x^n)$ とできる。このとき、

$$\begin{aligned}
E_p[l(x^n)] &= -\sum_{x^n} p(x^n)\log q(x^n) \\
&= \sum_{x^n} p(x^n)\log\frac{p(x^n)}{q(x^n)} - \sum_{x^n} p(x^n)\log p(x^n) \\
&= \sum_{x^n} p(x^n)\log\frac{p(x^n)}{q(x^n)} + H_n(p) \\
&\geq \sum_{x^n} p(x^n)\left(1 - \frac{p(x^n)}{q(x^n)}\right) + H_n(p) \\
&= \sum_{x^n} p(x^n) - \sum_{x^n} q(x^n) + H_n(p) \\
&= H_n(p)
\end{aligned}$$

ここで、不等号の部分の式変形では、$x > 0$ で $\log x \geq 1 - \frac{1}{x}$ が成り立つことを用いた。

なお、$\log x \geq 1 - \frac{1}{x}$が等号で成り立つのは $x = 1$ のときに限るため、$l(x^n) = -\log p(x^n)$ とすれば下限を達成できます。これは、x^n の p に関するShannon情報量と呼ばれます。

AIC vs. MDL

さて、本記事ではAICとMDL基準の、2つの情報量基準について説明してきました。異なる情報量基準で推定を行えば異なる推定モデルを得るため、実際にモデル推定を行うときには、どの情報量基準を使うべきか、ということも考えなくてはいけません。そこで、「情報量基準を選ぶ基準」というメタな基準から、AICとMDL基準を比較してみることにします。ここでは、以下の性質での比較を行います。

- 期待平均対数尤度 $nE_{X^n}E_{x^n}\left[-\log p(X;\hat{\theta}(x^n),k)\right]$ の不偏推定量を最小にするかどうか。
- 情報量基準による推定量は漸近有効性をもつかどうか。すなわち、情報量基準による推定量は、一般の推定量を用いたときの分散の下限を達成するかどうか。
- 一致性をもつかどうか。すなわち、データ数 $n \to \infty$ のとき、情報量基準による推定量が真の値に確率収束するかどうか。
- ミニマックスリグレットの下限を達成するかどうか。
- 情報量基準による推定量はoptimalityをもつかどうか。すなわち、推定量を $\bar{\theta}, \bar{M}$ とおき、$\bar{p}(X^n) = \arg\min_q \max_{\theta,M} \left\{-\log q(x^n) - (-\log p(x^n;\bar{\theta},\bar{M}))\right\}$ としたときに、$\bar{\theta}, \bar{M}$ は $\min_{\bar{\theta},\bar{M}} \max_{\theta,M} D(p_{\theta,M}||\bar{p})$ の下限を達成するかどうか。ただし、$D(f||g) = E_f\left[\log \frac{f(x)}{g(x)}\right]$ (Kullback-Leiber ダイバージェンス) 。

これらの性質が成り立つことを「○」、成り立たないことを「×」と書くと、AICとMDLに関しては次の表に示すとおりになります。

↓メタ基準＼情報量基準→	**AIC**	**MDL**	**「○」の証明**
期待平均対数尤度	○	×	Akaike, 1973
漸近有効性	○	×	Shibata, 1976
一致性	×	○	Rissanen, 1978
ミニマックスリグレット	×	○	Shtarkov, 1987
optimality	×	○	Rissanen, 2012

上に挙げた性質に関しては、AICとMDLは相補的な関係にあることがわかります。

おわりに

本記事では、情報理論から導かれる機械学習の理論の初歩的な部分を扱いました。情報理論と統計学のつながりや、そこからMDL基準という情報量基準が導かれること、そして、MDL基準はAICのような他の情報量基準とは異なった性質を持つ、という流れで説明してみました。

本記事で取り扱った内容は、教科書のはじめの1章ぶんくらいに過ぎず、情報論的学習理論はここからさらに拡がっているようなので、今後に学習した際には、また何かしら記事を書いてみようかと思います。

脚注・参考文献

人物の肩書などの情報は、2019年12月12日時点でのものになります。

---

1. 鈴木大慈 准教授 ホームページ: http://ibis.t.u-tokyo.ac.jp/suzuki/↩

2. 山西 健司 教授 ホームページ: http://www.ibis.t.u-tokyo.ac.jp/yamanishi/↩

3. https://catalog.he.u-tokyo.ac.jp/detail?code=4820-1026&year=2019↩

4. 青山和浩, 山西健司 著, 東京大学工学教程編纂委員会 編: システム工学 知識システムI 東京大学工学教程 知識の表現と学習, 丸善出版 (2017), pp.7-39.↩↩

5. 山西健司: 情報論的学習とデータマイニング, 朝倉書店 (2014), pp.110-121.↩

6. 山西健司: 情報論的学習理論, 共立出版 (2010), pp.1-34.↩↩↩

7. T. M. Cover and J. A. Thomas, Elements of Information Theory Second Edition, Wiley-Interscience, 1991, pp.107-109.↩

8. クロスバリデーション：n 個のデータを学習用データと評価用データに分け、学習用データで学習させたモデルの性能を評価用データで評価する手法。データを k 組に分割して、組の単位でクロスバリデーションを行い、性能の平均値をみる k-fold cross validation と呼ばれる手法や、$k=n$ として、$n-1$ 個を学習用にして残りの 1 個のデータで性能を評価するのを n 個のデータ全体で繰り返す leave-one-out cross validation と呼ばれる手法があります。↩

9. Bayesian information criterion: $\mathrm{BIC}_k = -\log p(x^n; \hat{\theta}(x^n), k) + \frac{k}{2}\log n$ を最小化する k を選ぶ、という基準。↩

10. g-関数という関数を用いた計算法もあります。詳細は文献[4]を参照ください。↩