情報意味論 (第6回)
モデル選択

慶應義塾大学理工学部
櫻井 彰人

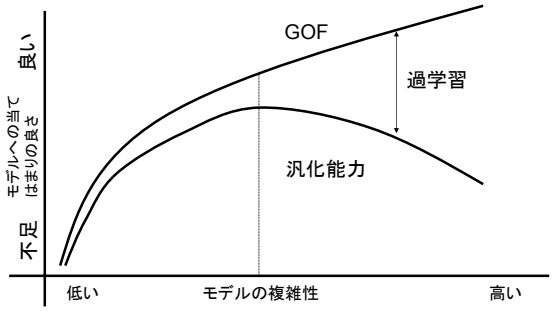
汎化能力

- 汎化能力は、(学習データではなく)未知データに対して正解することができる能力をいう.
- 学習データは、一般に、(分類なら)クラス値の誤り、(回帰なら)出力値の誤り(これらを、略してノイズということにしよう)で劣化している。
- 従って、当てはまりの良さ(goodness of fit, 長いので GOF と略)には、規則性だけでなく、ノイズへの当てはまりが反映することになる.

汎化能力

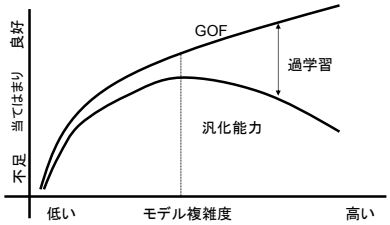
- GOF = 規則性への当てはまり (汎化能力) + データへの当てはまり (過学習)

汎化能力



汎化能力

- モデル複雑度が高いほど、過学習することになる.



汎化能力

モデル	M1	M2 (真)	M3	M4
	$y = w_1x+e$	$y = \ln(x + w_1) + e$	$y = w_1 \ln(x + w_2) + w_3 + e$	$y = w_1x + w_2x^2 + w_3 + e$
学習 GOF	2.14	1.85	1.71	1.62
テスト GOF	2.29	2.05	6.48	3.44

汎化能力

- モデルの自由度が高ければ、よいGOFが得られる（当てはまりがよい）.
- 当てはまりがよいことは必要である, しかし、それだけでは、隠れた構造を獲得したとこにはならない.
- 当てはまりがよいことは、次の検討ができるために条件である.

汎化能力

- 開発された手法の多くは、未知データによく当てはまるモデルを探す方法であり、必ずしも真のモデルを探す方法ではない.
  - 真のモデルが同定できるほどに多くのデータがあることは、まず、ない.
  - 万が一あったにしても、真のモデルは、考慮中のモデル集合にはない可能性だってある.
- 勿論、真のモデルがいらないと言っているわけではない.

モデル選択

- 関心を持つべきは、汎化能力のなさである.
- 本質（いい加減ですが）:
  - GOF = 規則性への当てはまり (汎化能力) + ノイズへの当てはまり (過学習)
  - 汎化能力 = GOF – 過学習
  - 汎化能力 = GOF – 複雑度
  - よって, –汎化能力 = – GOF + 複雑度

AIC（赤池の情報量規準）

- Akaike Information Criterion (AIC)
  - 赤池氏自身は An Information Criterion と命名した。他の情報量規準が提案されるに従い、上記の名前が普通に用いられるようになった
- AIC は汎化能力のなさの測度, 従って, 大きい方が悪い.

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. Proc. 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki eds .) Akademiai Kiado, Budapest, (1973) 267-281.

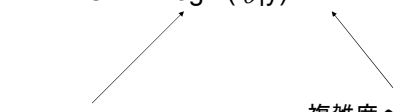
最初はこちら Hirotugu Akaike. Determination of the number of factors by an extended maximum likelihood principle. Research Memorandum 44, Inst. Statist. Math. (March 1971).

AIC

- AIC = -2 log L($\hat{\theta}$ |y) + 2k
  - y はデータ
  - w* は最尤推定量 (MLE)
  - L は尤度（ L($\hat{\theta}$ |y) = Prob(y| $\hat{\theta}$) ）
  - k はモデルパラメータの個数
  - log は自然対数

AIC

- AIC = -2 log L($\hat{\theta}$|y) + 2k



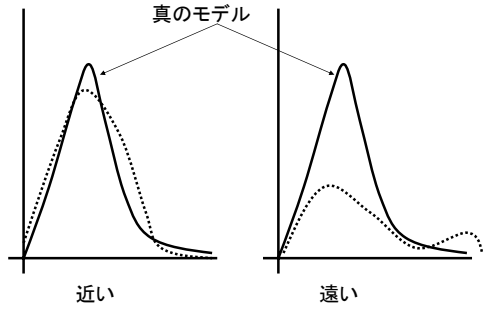
当てはまりの悪さ:
パラメータ数が増えると一般に減少する.

複雑度への罰金:
パラメータ数が増加すると増加する.

AIC

- AIC はパラメータ数を用いて、モデルの複雑度を測る.
- 関数形は考慮していない

モデル間の距離



AIC

- AIC はモデル集合の中から、真のモデルとの距離を平均的に最小化するモデルを選択する.
- AIC は、真のモデルに関する情報は用いない.

KL-情報量

- KL-情報量(Kullback-Liebler divergence)は、二つの分布間の距離のような量である(数学的な距離ではない。そこで、擬距離と呼ばれる)。
- 二つの分布を P_i と Q_iとするとき($Q_i \neq 0$ とする)

$$D(P,Q) = \sum_{i=1}^{k} P_i \log_2 \frac{P_i}{Q_i}$$

- $D(P,Q)$ の性質:

1 $D(P,Q) \geq 0$
2 $D(P,Q) = 0$ iff $P = Q$

- 対称性はない。三角不等式も満足しない

AICの導出手順

- 真の分布 P と推定モデル$P(\hat{\theta})$ の間のKL-情報量 $D(P,P(\hat{\theta}))$ を最小にするモデルを選びたい。
- ところが、データから得られるのは、経験分布 $\hat{P}$ であり、これからは $D(P,P(\hat{\theta}))$ を計算できない。
- そこで、経験分布 $\hat{P}$ と推定モデルの間のKL-情報量 $D(\hat{P},P(\hat{\theta}))$ から、 $\hat{\theta}$ の漸近正規性を用いて $D(P,P(\hat{\theta}))$ の平均を評価した
- そうして得られたのがAICである。
- なお、その結果は、

$$D\ (\ P\ ,\ P\ (\hat{\theta}\ ))\ \text{の推定量} = D\ (\ \hat{P}\ ,\ P\ (\hat{\theta}\ ))\ +\ \frac{k}{N}$$

- AICは、対数尤度の−2倍を補正するもので、次式で与えられる。

$$AIC = -2\sum_{i=1}^{N} \log p(X_i;\hat{\theta}) + 2k$$

Given a set of estimates $\hat{\theta}$'s of the vector of parameters θ of a probability distribution with density function $f(x \mid \theta)$ we adopt as our final estimate the one which will give the maximum of the expected log-likelihood, which is by definition

$$E \log f(X \mid \hat{\theta}) = E \int f(x \mid \theta) \log f(x \mid \hat{\theta})\, dx, \tag{1.1}$$

where X is a random variable following the distribution with the density function $f(x \mid \theta)$ and is independent of $\hat{\theta}$.

This seems to be a formal extension of the classical maximum likelihood principle but a simple reflection shows that this is equivalent to maximizing an information theoretic quantity which is given by the definition

$$E \log \left[\frac{f(X \mid \hat{\theta})}{f(X \mid \theta)}\right] = E \int f(x \mid \theta) \log \left[\frac{f(x \mid \hat{\theta})}{f(x \mid \theta)}\right] dx. \tag{1.2}$$

The integral in the right-hand side of the above equation gives the Kullback–Leibler's mean information for discrimination between $f(x \mid \hat{\theta})$ and $f(x \mid \theta)$ and is known to give a measure of separation or distance between the two distributions [15]. This observation makes it clear that what we are propos-

Since the purpose of estimating the parameters of $f(x \mid \theta)$ is to base our decision on $f(x \mid \hat{\theta})$, where $\hat{\theta}$ is an estimate of θ, the discussion in the preceding section suggests the adoption of the following loss and risk functions:

$$W(\theta, \hat{\theta}) = (-2) \int f(x \mid \theta) \log \left[\frac{f(x \mid \hat{\theta})}{f(x \mid \theta)}\right] dx \tag{3.1}$$

$$R(\theta, \hat{\theta}) = EW(\theta, \hat{\theta}), \tag{3.2}$$

where the expectation in the right-hand side of (3.2) is taken with respect to the distribution of $\hat{\theta}$. As $W(\theta, \hat{\theta})$ is equal to 2 times the Kullback–Leibler's

MDL: Minimum Description Length

- 次のデータ(ビット列)が与えられているとしよう:
  - 00010001000100010001000010001
  - 01110100110100000101010101011

MDL

- これらは、プログラムを使って、次のように符号化できる:
  - 00010001000100010001000010001
    - 7.times{ print "0001" }
  - 01110100110100000101010101011
    - puts("01110100110100000101010101011")

MDL

- データを圧縮するのに、規則性を活用することができる.
- データがより規則的であるほど、一般には符号化方法に依存するが、「プログラム」は短くなる.
  - 符号化方法は、理論としては、それほど問題にならない(符号化法に依存する項は定数で抑えられる).

MDL

- この「プログラム」をモデルだと考えよう.
- データ中の規則性を最もよく捉えたプログラムが、最短のプログラム、すなわち、最短の符号となる.
  - 00010001000100010001000010001
    - 7.times{ print "0001" }
  - 01110100110100000101010101011
    - puts("01110100110100000101010101011")

MDL

- データの規則性が獲得できれば、次に来るデータが予想できる。すなわち、よい汎化能力が獲得できる.
- すなわち、記述長最小のモデルを見つけることができれば、それは予測能力が最もあるモデルということになる.

統計的MDL

- 統計的な状況では、すなわち、データが分布する場合、MDL原理は:

$$MDL = -\ln f(x \mid \hat{\theta}) + \frac{k}{2}\ln\frac{N}{2\pi} + \ln\int\sqrt{\det I(\theta)}\,d\theta$$

当てはまりの悪さ

パラメータ数の多さに対する罰金

確率分布形に依存する罰金

J.Rissanen, Modeling by shortest data description. *Automatica,* vol. 14 (1978), pp. 465-471.
J.Rissanen, Fisher information and stochastic complexity. *IEEE Trans. Information Theory,* vol. 42 (1996), pp. 40-47.

MDL Reading http://www.mdl-research.org/reading.html

MDL規準の導出

- 情報の符号化を行うために、はじめに情報源の確率分布のパラメータ（正規分布の平均と分散など）を推定する。
- 推定した確率分布に従って、データの符号化を行う。
- （1）符号化したデータと、（2）データの符号化に用いた確率モデルのパラメータを適当な方法で符号化したもの、の両者を合わせたものが、求める記述長である。

- 確率モデルを $p(X;\theta)$ とする。
- 情報源からの*N*個の観測データを $X_1, X_2, \cdots, X_N$ とし、これらを用いたパラメータの最尤推定量を $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_N)$ とする。
- このときの（1）最適に選んだ符号化データの長さは、

$$\sum_{i=1}^{N} \log \frac{1}{p(X_i;\hat{\theta})} = -\sum_{i=1}^{N} \log p(X_i;\hat{\theta})$$

MDL規準の導出

- 次に（2）確率モデルのパラメータの符号化に必要なビット長を推定する。
- モデルのパラメータは連続値であるので、その符号化のために離散化して、近似することを考える。
- 推定されたパラメータは、その標準偏差程度、真の値から離れている可能性がある。そのため、離散化の幅は、標準偏差程度とするのが妥当。（離散化の幅が大きすぎると、近似精度が悪くなる。一方、離散化の幅が小さすぎると、パラメータの記述長が長すぎる。このトレードオフをちゃんと計算すると、標準偏差値程度となる）
- パラメータを1変量とすると、その推定量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_N)$ の標準偏差は、そのサンプル数*N*の平方根に反比例する。
- パラメータ数を k とすると、その標準偏差は $O(1/N^{k/2})$ となる。
- 全体を1として、その中から、標準偏差で区分けした点の一つを指定するのに必要な情報量は、$p = 1/N^{k/2}$ の一様分布に従う情報源からの1つの標本の符号化長 $k/2 \log N$ となる。

MDL規準の導出

- （1）、（2）を合計したものが、最小記述長となる。すなわち、

$$MDL = -\sum_{i=1}^{N} \log p(X_i;\hat{\theta}) + \frac{k}{2} \log N$$

となる。ここで、Nはデータ数、 k はパラメータ数である。

AICとMDLの比較

- AICとMDLを比較する。

$$AIC = -2\sum_{i=1}^{N} \log p(X_i;\hat{\theta}) + 2k$$

$$2\,MDL = -2\sum_{i=1}^{N} \log\ p(X_i;\hat{\theta}) + k\ \log\ N$$

- 第2項の補正項を比べると、Nが大きい場合にはMDLの方が大きくなる。このため、MDLはAICに比べて、パラメータ数が多いモデルが選びにくくなっている。逆に言えば、同じデータにたいしてはMDLの方が小さいモデルを選ぶ傾向があると予想される。（「情報理論の基礎」村田昇著、サイエンス社から引用）