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

モデル選択

- あるデータを説明する、複数個の(確率的)モデルがある時、その中の一つを選択すること。「最良の」モデルを選びたい
- 「最良」とはどういうことか
- それを実現するにはどうするか？
- 最良とは、未知のデータに対して(他のモデルより)、誤差の小さい予測をする。
- 方法:「未知のデータに対する誤差」を推定する
  - 実データを用いる方法
    - Validation dataset を用いる方法
    - Cross validation (学習データの一部を用いる方法)
  - 理論的に予測する方法

復習

モデル選択

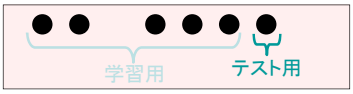
- 一つの方法
  - 予測(汎化)誤差の推定値が最も小さいところ(複雑度、学習回数)の学習器を使う
    - Validation set を用いる。Cross validation を行う
- 他の方法
  - 情報量規準に基づいて最適な複雑度を推定する。

方法:「未知のデータに対する誤差」を推定する
実データを用いる方法
Validation dataset を用いる方法
Cross validation (学習データの一部を用いる方法)
理論的に予測する方法

再掲

k 重クロスバリデーション
k-fold cross validation

訓練データをk 群に分け、$(k-1)$ 群で学習し、残りで予測誤差を計測する。これを全ての k 種類の組み合わせに対して行なう



万能ではないが、多くの場合に結構うまくいく
アルゴリズムや構造の適切さを測ることになる
構造や構造のパラメータ(複雑度)を決める目的で用いる

代表的な情報量規準

- AIC
- MDL

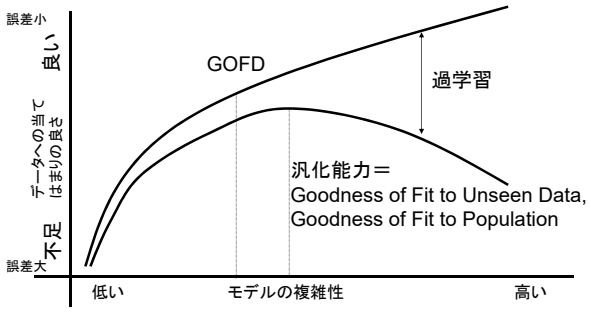
汎化能力

- 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、データへの当てはまりの良さ(goodness of fit to Data, 長いので GOFD 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

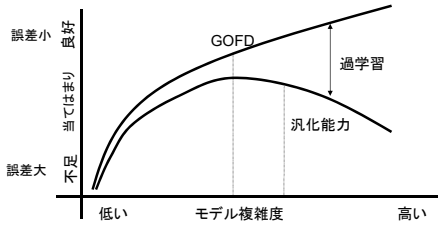
- GOFD
  = 規則性への当てはまり (汎化能力)
  + 規則性からのずれへの当てはまり
  (過学習)

汎化能力

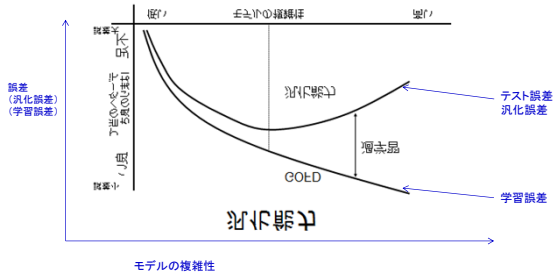


汎化能力

- モデル複雑度が高いほど、過学習することになる.



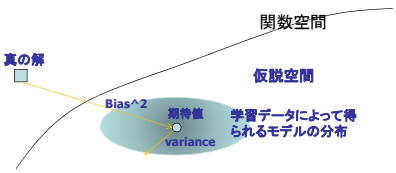
汎化誤差



Bias-variance decomposition

- 学習によって得られたモデルを用い、未知の入力値に対して出力値を得たとするときの、その出力値の2乗誤差の値を、bias^2とvarianceの和（＋ノイズ）に分ける。直観的に
  - モデルと元の関数（回帰の場合）との違い
  - モデルの学習データセットによる揺れ

  の和（？）に分解できることが分かる



Bias-variance decomposition

仮定 $Y = f(X) + \varepsilon$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

損失の期待値

$L(x_0) = E\left[\left(Y - \hat{f}(x_0)\right)^2 \middle| X = x_0\right]$

x_0 は未知の値
E は ε に関する期待値

$= E\left[\left((Y - f(X)) + \left(f(X) - \hat{f}(x_0)\right)\right)^2 \middle| X = x_0\right]$

$= E\left[(Y - f(X))^2 \middle| X = x_0\right] + E\left[\left(f(X) - \hat{f}(x_0)\right)^2 \middle| X = x_0\right]$

$+ 2E\left[(Y - f(X))\left(f(X) - \hat{f}(x_0)\right) \middle| X = x_0\right]$

$= \sigma_\varepsilon^2 + E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]$

$+ 2\left(f(x_0) - \hat{f}(x_0)\right) E[(Y - f(X)) | X = x_0]$

$= \sigma_\varepsilon^2 + \left(f(x_0) - \hat{f}(x_0)\right)^2$

注: $f(x_0) = E[Y | X = x_0]$

Bias-variance decomposition

学習データセット D に関する期待値を計算する

$$E_D[L(x_0;D)] = \sigma_\varepsilon^2 + E_D\left[\left(f(x_0) - \hat{f}(x_0;D)\right)^2\right]$$

$\hat{f}$ が D に依存している

右辺第2項は

$$E_D\left[\left(\left(f(x_0) - E_D[\hat{f}(x_0;D)]\right) + \left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)\right)^2\right]$$

$$= E_D\left[\left(f(x_0) - E_D[\hat{f}(x_0;D)]\right)^2\right] + E_D\left[\left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)^2\right]$$

$$+ 2E_D\left[\left(f(x_0) - E_D[\hat{f}(x_0;D)]\right)\left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)\right]$$

ところが、右辺第3項は

$$2\left(f(x_0) - E_D[\hat{f}(x_0;D)]\right)E_D\left[\left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)\right] = 0$$

結局 $E_D[L(x_0;D)]$ は

$$\sigma_\varepsilon^2 + \left(f(x_0) - E_D[\hat{f}(x_0;D)]\right)^2 + E_D\left[\left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)^2\right]$$

Bias-variance decomposition

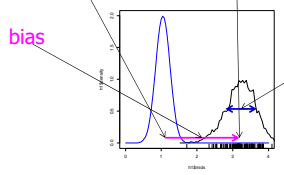
結局 $E_D[L(x_0;D)]$ は

$$\underbrace{\sigma_\varepsilon^2}_{\text{ノイズ}} + \underbrace{\left(f(x_0) - E_D[\hat{f}(x_0;D)]\right)^2}_{\text{bias の二乗}} + \underbrace{E_D\left[\left(E_D[\hat{f}(x_0;D)] - \hat{f}(x_0;D)\right)^2\right]}_{\text{variance}}$$

$f(x_0)$ の真値

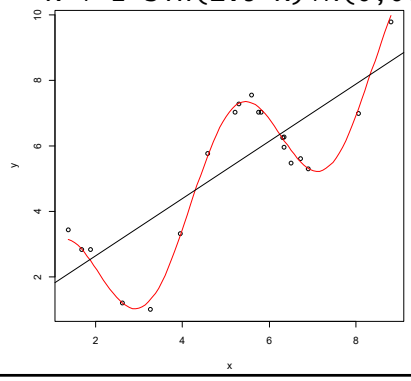
$f(x_0)$ の予測値の、D を振った時の期待値

$f(x_0)$ の予測値。D に依存

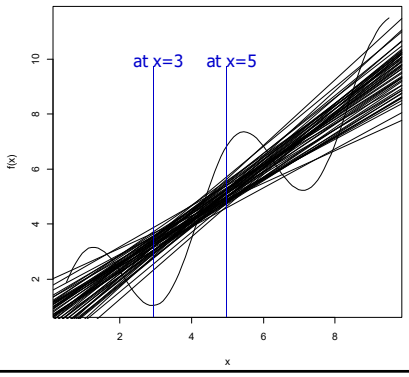


簡単な例（20事例）

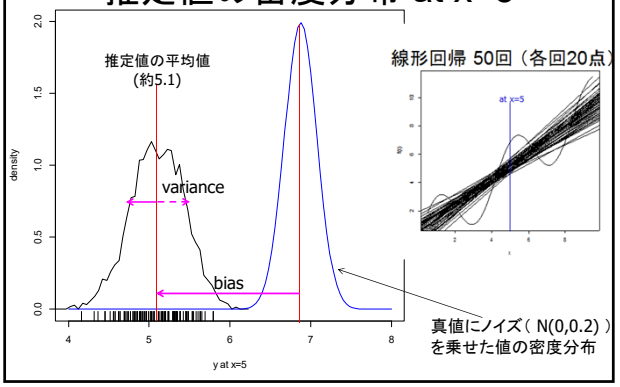
y = x + 2*sin(1.5*x)+N(0,0.2)

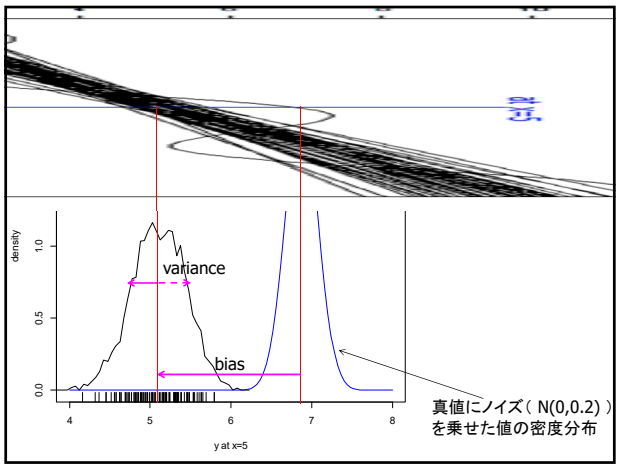


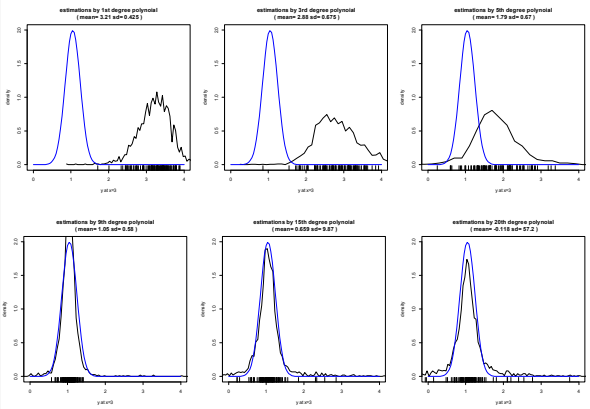
線形回帰 50回（各回20点）



推定値の密度分布 at x=5

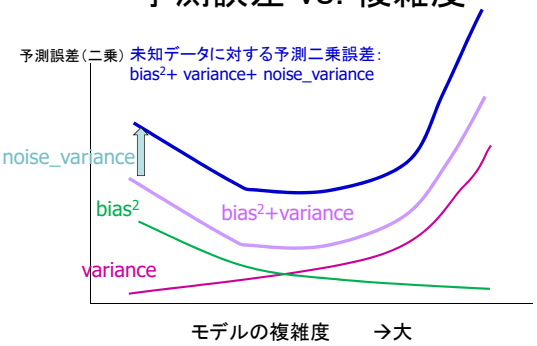




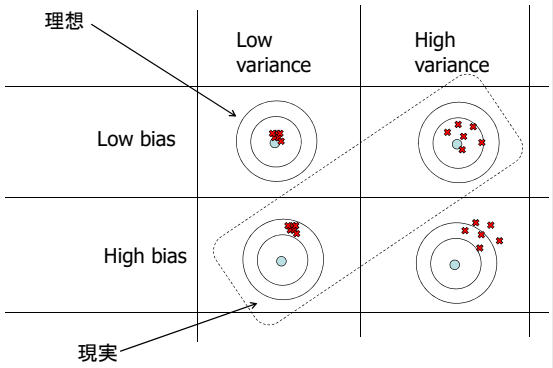


注: at x=3; 高次でも、外れた点が多い

予測誤差 vs. 複雑度



蛇足: 推定量が2次元だとして書いた図



汎化能力

- モデルの自由度が高ければ、よいGOFDが得られる（当てはまりがよい）.
- 学習データへの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、テストデータへの当てはまりがよいことにはならない. すなわち、隠れた構造を獲得したとこにはならない.
- しかし、当てはまりがよいことは、更なる検討を進める上で必要である.

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（多くの場合、パラメータ）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.
  - 多くのデータがあっても、ノイズで汚されることはある.
  - 真のモデルが同定できるには、ノイズがノイズであることが分かるほど、多くのデータが必要.
  - （真のモデルはないかもしれない。しかし）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（仮説集合）にはないかもしれない.
- だからといって、真のモデルがいらないと言っているわけでは、勿論、ない.

モデル選択

- 考えるべきは、汎化能力の大小である.
- 本質（いい加減ですが）:
  - GOFD = 規則性への当てはまり (汎化能力) + 偏りへの当てはまり (過学習)
  - 汎化能力 = GOFD – 過学習
  - 汎化能力 ≈ GOFD – 複雑度
  - よって, –汎化能力 ≈ – GOFD + 複雑度

複雑度

- そこで、複雑度を適切に定義することにより、より良い汎化誤差を推定することはできないか？
  - 推定された汎化誤差に基づきモデルを選択する（最小値のものを選ぶ）ことにより、本当の汎化誤差の小さいモデルを選ぶことにならないか？

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, 大きすぎると都合が悪い.

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.

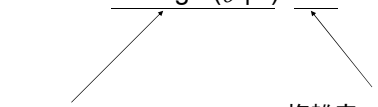
最初はこちら Hirotugu Akaike. Determination of the number of factors by an extended maximum likelihood principle. Research Memorandum 44, Inst. Statist. Math. (March 1971).

AIC

- AIC = -2 log L($\hat{\theta}$ |D) + 2k
  - D は学習データ
  - $\hat{\theta}$ は最尤推定量 (MLE)
  - L は尤度（ L($\hat{\theta}$ |D) = Prob(D| $\hat{\theta}$) ）
  - k はモデルを規定するパラメータの個数
  - log は自然対数

AIC

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



当てはまりの悪さ（誤差）: 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

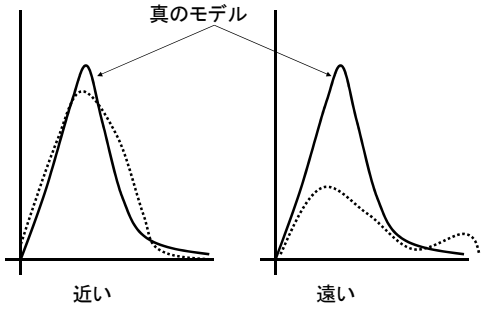
- AIC はパラメータ数を用いて、モデルの複雑度を測る.
- （パラメータ→確率という）関数の形（に関する複雑度）は考慮していない

「関数の形」の複雑度とは何か？
そもそも「関数の形」とは何か？
しかし、考えたくはなる

AIC

- AIC最小化 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.
  - 今持っている学習データについて、今のモデルについては、何も言っていない
- AIC は、真のモデルに関する情報は用いない.

モデル間の距離



KL-情報量

- KL-情報量(Kullback-Leibler 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）

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

注: クロスエントロピー

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

$$H(P,Q) = H(P) + D(P,Q)$$

- $D(P,Q)$ 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（ラフスケッチ）

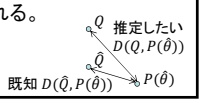
- 真の分布 Q と推定モデル$P(\hat{\theta})$ の間のKL-情報量$D(Q, P(\hat{\theta}))$ を最小にするモデルを選びたい。
- ところが、データから得られるのは、経験分布$\hat{Q}$ であり、これからは$D(Q, P(\hat{\theta}))$ が計算できない。
- そこで、経験分布$\hat{Q}$ と推定モデルの間のKL-情報量$D(\hat{Q}, P(\hat{\theta}))$ から、$\hat{\theta}$の漸近正規性を用いて$D(Q, P(\hat{\theta}))$ の平均を評価した
- その結果は、

漸近正規性: 漸近的に正規分布に従う

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

- AICは、上記の$2N$倍であり、次式で与えられる。

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



参考

AICの導出(ラフスケッチ1)

- 真の分布 Q と推定モデル $P(\hat{\theta})$ の間のKL-情報量 $D(Q, P(\hat{\theta}))$ を最小にするモデル i.e. $\hat{\theta}$ を選びたい。

g: 真の分布
f: 近似分布

$$D(g; f) = E_g[\log g(X)/f(X;\theta)]$$
$$= E_g[\log g(X)] - E_g[\log f(X;\theta)]$$

$$\hat{\theta}_{best} = \arg\min_\theta D(g; f)$$
$$= \arg\max_\theta E_g[\log f(X;\theta)]$$

参考

AICの導出(ラフスケッチ2)

- ところが、データから得られるのは、経験分布 $\hat{Q}$ であり、これからは $D(Q, P(\hat{\theta}))$ を計算できない。
- そこで、経験分布 $\hat{Q}$ と推定モデルの間のKL-情報量 $D(\hat{Q}, P(\hat{\theta}))$ から、$\hat{\theta}$ の漸近正規性を用いて $D(Q, P(\hat{\theta}))$ の平均を評価した

g: 真の分布 Q
f: 近似分布 $P(\theta)$

$$\hat{\theta}_{best} = \arg\min_\theta D(g; f)$$
$$= \arg\max_\theta E_g[\log f(X;\theta)]$$

$$\max_\theta E_g[\log f(X;\theta)] = \max_\theta E_{\hat{g}}[\log f(X;\theta)] + \text{補正項}$$
$$= \max_\theta \left(\frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\theta) + \text{補正項} \right)$$

$$\hat{\theta}_{ML} = \arg\max_\theta E_{\hat{g}}[\log f(X;\theta)] = \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\theta)$$

参考

AICの導出(ラフスケッチ3)

- そこで、経験分布 $\hat{Q}$ と推定モデルの間のKL-情報量 $D(\hat{Q}, P(\hat{\theta}))$ から、$\hat{\theta}$ の漸近正規性を用いて $D(Q, P(\hat{\theta}))$ の平均を評価した

$$\max_\theta E_g[\log f(X;\theta)] = \max_\theta E_{\hat{g}}[\log f(X;\theta)] + \text{補正項}$$
$$= \max_\theta \left(\frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\theta) + \text{補正項} \right)$$

$$E_g\left[\log f(X;\hat{\theta}_{best})\right] \approx E_{\hat{g}}\left[\log f(X;\hat{\theta}_{ML})\right] + \left(-\frac{k}{N}\right)$$

$$\hat{\theta}_{best} = \arg\max_\theta E_g[\log f(X;\theta)]$$
$$\hat{\theta}_{ML} = \arg\max_\theta E_{\hat{g}}[\log f(X;\theta)] = \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\theta)$$

参考AICの導出（スケッチ）：改めて

最尤推定された統計モデル群（の中のモデル同士）を比較する情報量基準を考える。

$g(y)$: 真のモデル、真の分布（未知）
$f(y)$: 推定した一つの統計モデル

Kullback-Leibler divergence

$$D(g;f) = E_g\left[\log g(y)/f(y)\right]$$
$$= \int_{-\infty}^{\infty} \left(\log g(y)/f(y)\right) g(y)dy$$
$$= \sum_{i=1}^{k} g_i \log_2 \frac{g_i}{f_i}$$

1 $D(g;f) \geq 0$
2 $D(g;f) = 0 \text{ iff } g = f$

参考

$$D(g;f) = 0 \text{ iff } g = f$$

つまり、KL情報量の値が小さければ、推定したモデルと真のモデルとが近いことになる

しかし、現実のデータのモデル化には、KL情報量の値を用いることはできない。真のモデルが未知だからである。

しかし、（真のモデルから得られた）データはある。

参考

$$D(g;f) = E_g\left[\log g(y)/f(y)\right]$$
$$= \boxed{E_g\left[\log g(y)\right]} - \boxed{E_g\left[\log f(y)\right]}$$

$g(y)$ は知らないので計算はできない。しかし、fによらない一定値

この値が大きければ大きいほどKL情報量値は小さくなる（「この」値は負なので、0に近いほどよい）

つまり、右辺第二項を最大にするようなモデル f を選べばよいことになる

$$E_g\left[\log f(y)\right] = \int_{-\infty}^{\infty} \left(\log f(y)\right) g(y)dy$$

$$\frac{1}{N}\sum_{i=1}^{N} \log f(y_i) \to E_g\left[\log f(y)\right]$$ 大数の法則

参考

記号の定義

$g(y)$: 真のモデル、真の分布（未知）
$f(y|\theta)$: 推定した一つの統計モデル

$X = (x_1, ..., x_N)$ データ、学習データ、訓練データ

$\hat{\theta} = \hat{\theta}(X)$ データ X に基づくパラメータθの最尤推定量

$l(\theta) = \sum_{i=1}^{N} \log f(x_i|\theta)$ （対数尤度）を最大にするθ

θ_0 $E_g\left[\log f(y|\theta)\right]$ を最大にするθ

真の分布に最近接する分布を与えるパラメータ

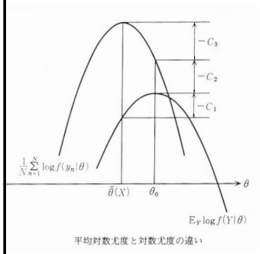
この値は知りようがないので、途中の計算に用いるのみ

$E_g\left[\log f(y|\hat{\theta})\right]$ を $\frac{1}{N}l(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\hat{\theta})$ で近似したい

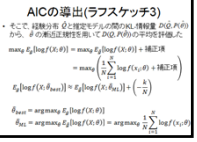
その差（推定の偏り）の期待値を求めよう $\hat{\theta} = \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\theta)$

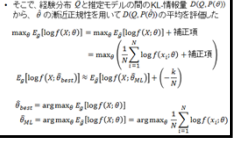
$$C \equiv E_X\left[E_g\left[\log f(y|\hat{\theta})\right] - \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\hat{\theta})\right]$$

$$C \equiv E_X\left[E_g\left[\log f(y|\hat{\theta})\right] - E_g\left[\log f(y|\theta_0)\right]\right]$$
$$+ E_X\left[E_g\left[\log f(y|\theta_0)\right] - \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\theta_0)\right]$$
$$+ E_X\left[\frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\theta_0) - \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|\hat{\theta})\right]$$
$$\equiv C_1 + C_2 + C_3$$



平均対数尤度と対数尤度の違い





$C_1 \equiv E_X\left[E_g\left[\log f(y|\hat{\theta})\right] - E_g\left[\log f(y|\theta_0)\right]\right]$ を評価する

$$E_g\left[\log f(y|\hat{\theta})\right] \approx E_g\left[\log f(y|\theta_0)\right] + \frac{\partial}{\partial\theta} E_g\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}(\hat{\theta}-\theta_0)$$

$f(y|\theta_0)$ は停留点

$$+\frac{1}{2}(\hat{\theta}-\theta_0)^T\left\{\frac{\partial}{\partial\theta\partial\theta^T} E_g\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}\right\}(\hat{\theta}-\theta_0)$$

$$= E_g\left[\log f(y|\theta_0)\right] + \frac{1}{2}(\hat{\theta}-\theta_0)^T E_g\left\{\frac{\partial}{\partial\theta\partial\theta^T}\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}\right\}(\hat{\theta}-\theta_0)$$

$$= E_g\left[\log f(y|\theta_0)\right] - \frac{1}{2}(\hat{\theta}-\theta_0)^T J(\hat{\theta}-\theta_0)$$

$$J = -E_g\left\{\frac{\partial^2}{\partial\theta\partial\theta^T}\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}\right\}$$

$$I = -E_g\left\{\frac{\partial}{\partial\theta}\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}\right\}\left\{\frac{\partial}{\partial\theta}\left[\log f(y|\theta)\right]\Big|_{\theta=\theta_0}\right\}^T$$

とすると、適当な条件のもと

$\sqrt{N}(\hat{\theta}-\theta_0) \sim \mathrm{N}(0, J^{-1}IJ^{-1})$

参考

準備

x: $n \times 1$ ベクトル、A: $n \times n$ 正方行列

→ $x^T Ax = \mathrm{tr}(Axx^T)$

なぜなら両辺とも $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$

x: $n \times 1$ ランダムベクトルで、$\mathrm{E}(x)=\theta$, $\mathrm{Var}(x)=\Sigma$

→ $E(x^T Ax) = \mathrm{tr}(A\Sigma) + \theta^T A\theta$

なぜなら $E(x^T Ax) = E(\mathrm{tr}(Axx^T)) = \mathrm{tr}(A \cdot E(xx^T)) = E(\mathrm{tr}(A(\Sigma + \theta\theta^T))$

$= \mathrm{tr}(A\Sigma) + \mathrm{tr}(A\theta\theta^T) = \mathrm{tr}(A\Sigma) + \theta^T A\theta$

$X \sim N(\mu, \Sigma)$, C: 正則行列

→ $CX \sim \mathrm{N}(C\mu, C\Sigma C^T)$

C_1 の評価の続き

$$\hat{\theta} - \theta_0 \sim \mathrm{N}(0, J^{-1}IJ^{-1}/N)$$

$$E_X\left\{(\hat{\theta}-\theta_0)^T J(\hat{\theta}-\theta_0)\right\} = E_X\left\{\mathrm{tr}(J(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T)\right\}$$

$$= \mathrm{tr}(J \cdot J^{-1}IJ^{-1}/N) = \frac{1}{N}\mathrm{tr}(IJ^{-1}) \qquad k = \dim(I)$$

$$= \frac{k}{N} \quad \text{if } g(y) = f(y|\theta_0)$$

$$\approx \frac{k}{N} \quad \text{otherwise}$$

$$E_g\left[\log f(y|\hat{\theta})\right] - E_g\left[\log f(y|\theta_0)\right] = -\frac{1}{2}(\hat{\theta}-\theta_0)^T J(\hat{\theta}-\theta_0)$$

であるので

$$C_1 \equiv E_X\left[E_g\left[\log f(y|\hat{\theta})\right] - E_g\left[\log f(y|\theta_0)\right]\right]$$

$$= E_X\left[-\frac{1}{2}(\hat{\theta}-\theta_0)^T J(\hat{\theta}-\theta_0)\right] \approx -\frac{k}{2N}$$

参考

C_3 の評価

$$\frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\theta_0) \approx \frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\hat{\theta}) + \left\{\frac{1}{N}\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\log f(y_i|\hat{\theta})\right\}(\theta_0-\hat{\theta})$$

$$+\frac{1}{2}(\theta_0-\hat{\theta})^T\left\{\frac{1}{N}\frac{\partial^2}{\partial\theta\partial\theta^T}\sum_{i=1}^{N}\log f(y_i|\hat{\theta})\right\}(\theta_0-\hat{\theta})$$

大数の法則より

$$\frac{1}{N}\frac{\partial^2}{\partial\theta\partial\theta^T}\sum_{i=1}^{N}\log f(y_i|\hat{\theta}) \to E_g\left\{\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(y|\theta_0)\right\} = -J$$

したがって

$$\frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\theta_0) - \frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\hat{\theta}) \approx -\frac{1}{2}(\theta_0-\hat{\theta})^T J(\theta_0-\hat{\theta})$$

C_1 の議論と同様に

$$C_3 = E_X\left[\frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\theta_0) - \frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\hat{\theta})\right] \approx -\frac{k}{2N}$$

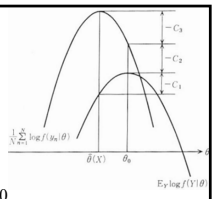
参考

C_2 の評価

$$C_2 = E_X\left[E_g\left[\log f(y|\theta_0)\right] - \frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\theta_0)\right]$$

$$= E_g\left[\log f(y|\theta_0)\right] - \frac{1}{N}\sum_{i=1}^{N}E_X\left[\log f(y_i|\theta_0)\right]$$

$$= E_g\left[\log f(y|\theta_0)\right] - \frac{1}{N}\cdot N \cdot E_X\left[\log f(X|\theta_0)\right] = 0$$



以上より

$$C = C_1 + C_2 + C_3 \approx 2\times\left(-\frac{k}{2N}\right) = -\frac{k}{N}$$

$$C \equiv E_X\left[E_g\left[\log f(y|\hat{\theta})\right] - \frac{1}{N}\sum_{i=1}^{N}\log f(y_i|\hat{\theta})\right] \approx -\frac{k}{N}$$

すなわち、最尤推定量 $\hat{\theta}$ で評価した対数尤度 $\frac{1}{N}l(\hat{\theta})$ は、

平均対数尤度よりも平均的に $\frac{k}{N}$ だけ大きな値となる

参考

したがって、$\frac{1}{N}l(\hat{\theta}) - \frac{k}{N}$ は、最尤モデルの平均対数尤度 $E_g \log f(y|\hat{\theta})$ の不偏推定量になっている。

赤池情報量基準は、これを $-2N$ 倍して、

AIC $= -2l(\hat{\theta}) + 2k$

$= -2$(最大対数尤度)$+2$(パラメータ数)

と定義する

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:
  - 000100010001000100010001000100010001
  - 011101001101000010101010101011

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:
  - 000100010001000100010001000100010001
    - 7.times{ print "0001" }
  - 011101001101000010101010101011
    - puts("011101001101000010101010101011")

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.
  - 000100010001000100010001000100010001
    - 7.times{ print "0001" }
  - 011101001101000010101010101011
    - puts("011101001101000010101010101011")

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.

```
000100010001000100010001000100010001
7.times{ print "0001" }
puts("000100010001000100010001000100010001")
```

Occam の剃刀

- 人口に膾炙しているのは
  - Entities should not be multiplied beyond necessity.
- Bertrand Russell によれば
  - It is vain to do with more what can be done with fewer.
- 最も普通の解釈
  - Among the theories that are consistent with the observed phenomena, one should select the simplest theory.

Occam の剃刀: 蛇足

- もっと以前から言われていたといわれる。表現も複数ある。
  - Wikipedia参照
  - 最近であれば "The philosophy of John Duns Scotus" の第8.2節
  - 古ければ、"The Myth of Occam's Razor" Mind, 27(107), 345-353 (1918)
- Albert Einstein: "Theories should be as simple as it is, but not simpler."
- 残差があるときは単に「単純に」とはいえない
- 「仮定」を入れても、それにより複数の現象が一つの理論で説明可能となるなら、これも単純化
  - 例: 未だ見えなかった分子（原子、素粒子）による、現象の説明
- 単に「概念の個数」だけを数えるのでは、誤る。理論全体の長さを考えするべき
  - そして、残差の「長さ」も

最小記述長(minimum description length)

- Occam's razor: "最短仮説を選べ"

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D \mid h)$$

ex. 決定木を記述するビット数 ∝ 記述する符号の長さ

h が所与のとき、D を記述するビット数 ∝ 誤分類データの個数

このままでは、使えない。使うようにした方法がある
1. Rissanen による統計的MDL
2. Kolmogorov/Chaitin のプログラム複雑度に基づくMDL

であり、Lin & Vitanyi グループによるもの

最小記述長 符号的解釈

- MDL: 次を最小化する仮説を選ぶ

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D \mid h) P(h) \\ &= \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h) \\ &= \arg\min_{h \in H} L_{C_2}(D \mid h) + L_{C_1}(h) \end{aligned}$$

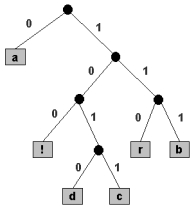
蛇足: 確率と符号長

- 有限または可付番無限集合 X を考える
  - X の符号 $C(x)$ とは
    - X から $U_{n>0}\{0,1\}^n$ への1-to-1 写像
    - $L_C(x)$: 符号C を用いた時の符合長（ビット）
  - P: X 上で定義した確率分布
    - $P(x)$: x の確率
    - 観測値の系列（通常は iid） $x_1, x_2, \ldots, x_n$: x^n

$$P(x^n) = \prod_{i=1}^{n} P(x_i)$$

蛇足:確率と符号長: 接頭符号（語頭符号）

- 接頭符号: 瞬時復号可能な符号の例
  - どの符号も他の符号の語頭にはなっていない



a	0
b	111
c	1011
d	1010
r	110
!	100

http://www.cs.princeton.edu/courses/archive/spring04/cos126/

蛇足: 確率と符号長: 最適符号

- ある符号 C の符号長の期待値

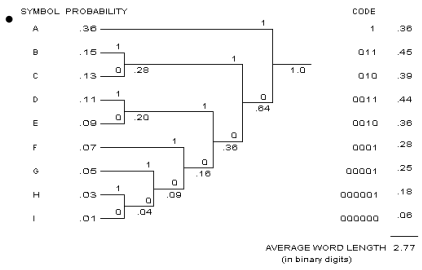
$$E_P(L_C(x)) = \sum_{x \in X} P(x) L_C(x)$$

  - 下界:

$$H(x) = -\sum_{x \in X} P(x) \log_2 P(x) = \sum_{x \in X} P(x)(-\log_2 P(x))$$

- 最適符号
  - 瞬時復号可能な符号の中で期待符号長が最小
  - 仮に分布 P が与えられた時、どう設計せればよいか?
    - Huffman 符号

蛇足: 確率と符号長: ハフマン符号



http://star.itc.it/caprile/teaching/algebra-superiore-2001/

蛇足: 確率と符号長: 有限個数

- {1, 2, …, M} の符号語を設計するには?
  - 一様分布を仮定すれば: それぞれの数に *1/M*
  - ~log*M* ビット

蛇足: 確率と符号長: 無限集合なら

- 正整数すべての符号を設計するには?
  - それぞれの *k* について
    - まず先頭に $\lceil \log k \rceil$ 個の0をおき
    - 次に一個の 1 をおき
    - そして *k* を符号化する。ただし $\{1,\ldots,2^{\lceil \log k \rceil}\}$ の符号
    - 長さは合計 ~ 2log*k* + 1 ビット
  - 勿論、改善は可能…

蛇足: 確率と符号長: 双対性(?)

- *P* を *X* 上の確率分布としよう。そうすると *X* に対する符号 *C* で次の条件を満たすものがある:

$$L_C(x) = \lceil -\log P(x) \rceil$$

- *C* を *X* 上の即時復号可能な符号とする. そうすると確率分布 *P* で次の条件を満たすものがある:

$$L_C(x) = -\log P(x)$$

$$L_C(x^n) = -\log P(x^n)$$

再掲: 最小記述長 符号的解釈

- MDL: 次を最小化する仮説を選ぶ

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D \mid h) P(h) \\ &= \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h) \\ &= \arg\min_{h \in H} L_{C_2}(D \mid h) + L_{C_1}(h) \end{aligned}$$

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:

$$MDL = \underbrace{-\ln f(x \mid \hat{\theta})}_{\text{当てはまりの悪さ(誤差)}} + \underbrace{\frac{k}{2} \ln \frac{N}{2\pi}}_{\text{パラメータ数の多さに対する罰金}} + \underbrace{\ln \int \sqrt{\det I(\theta)} d\theta}_{\text{確率分布形に依存する罰金}}$$

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

参考

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 $\theta = \hat{\theta}(X_1, X_2, \cdots, X_N)$ 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、

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

参考

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。

参考

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。

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

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

- 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）

これが比較的簡単にわかるのは、Bayesian network を学習するときである。
Wekaで試してみるとよい