データ解析

第13回：モデルの評価

渡辺澄夫

モデルの評価

実世界

評価法

観測データ

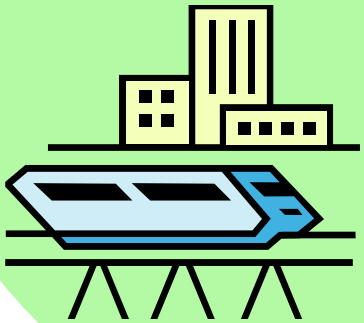
回帰・判別分析

推定と検定

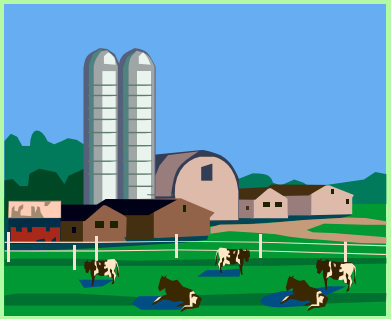
解析方法

統計的検定

ベイズ法
・階層ベイズ法



時系列予測



因子・主成分
・クラスタ分析

統計学における評価とは

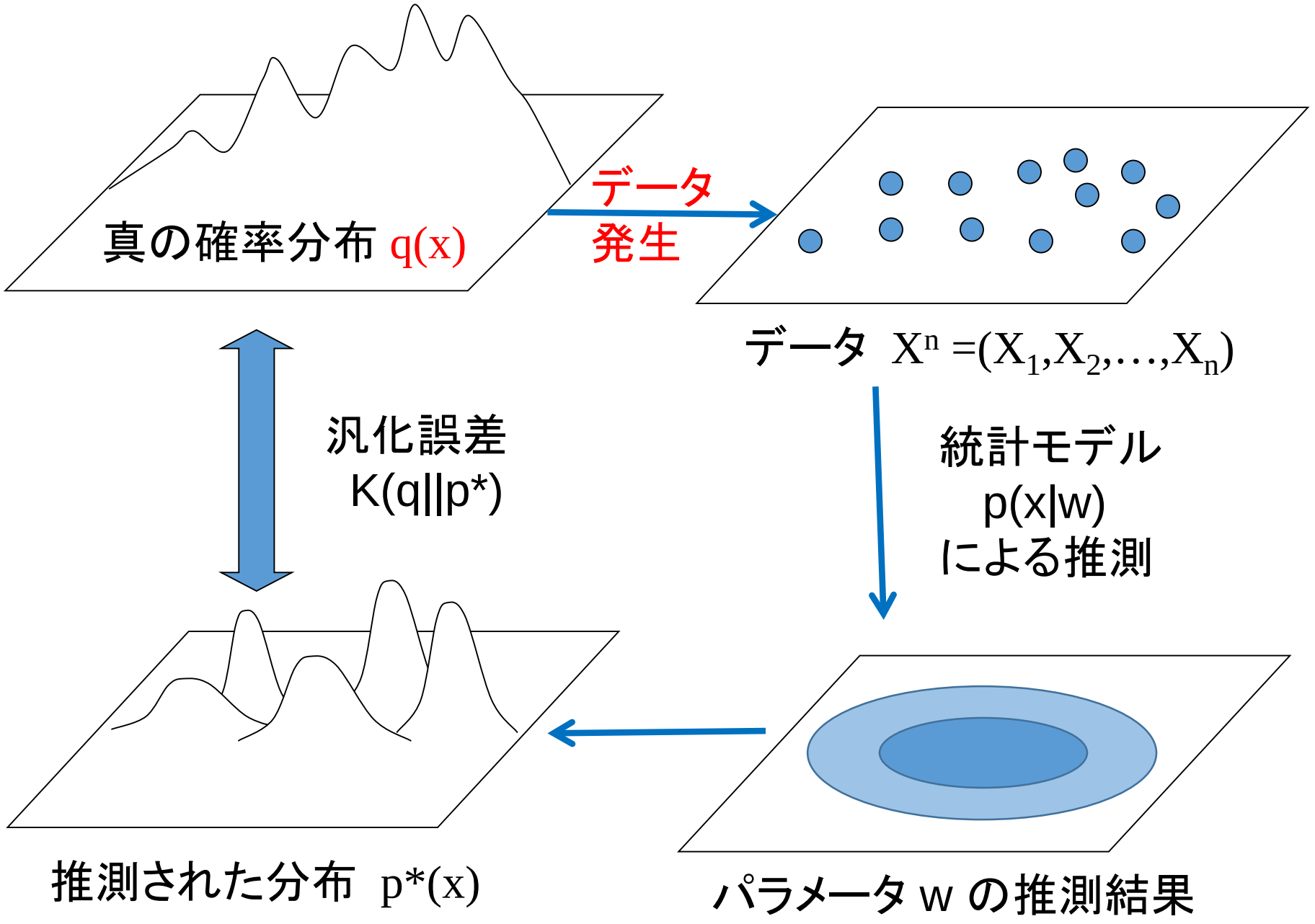
１．基本的な疑問：

「自分の用いたモデルは適切だったのだろうか」

２．モデルの適切さは、データを発生している真の分布を規準にして考察されるべきである。

３．しかしながら、真の分布はわからないので、較べることができない。

４．数学的な法則を利用すればよい。


真の確率分布 q(x)
データ
発生
データ $X^n=(X_1,X_2,\ldots,X_n)$
統計モデル
p(x|w)
による推測
パラメータ w の推測結果
推測された分布 p*(x)
汎化誤差
K(q||p*)

最尤推定量の漸近分布

フィッシャー情報行列

定義. 統計モデル $p(x|w)$ $(x \in R^N, w \in R^d)$のフィッシャー情報行列を

$$I_{jk}(w) = \int [(\partial / \partial w_j) \log p(x|w)][(\partial / \partial w_k) \log p(x|w)]\, p(x|w)dx$$

と定義する。

（注意）定義より $I(w)$ は非負定値行列である。すなわち任意の $v=(v_i)$ について $(v, I(w)v) \geqq 0$.

補題. $M(x) = \sup_w |(\partial^2 / \partial w_j \partial w_k) p(x|w)|$ とおくとき

$\int M(x)\, dx < \infty$が成り立つとする。このとき

$$I_{jk}(w) = - \int [(\partial^2 / \partial w_j \partial w_k) \log p(x|w)]\, p(x|w)dx.$$

補題の証明

証明. $p=p(x|w)$, $\partial_j = \partial / \partial w_j$ とかく。

$$(\partial^2 / \partial w_j \partial w_k) \log p(x|w) = \partial_j \partial_k \log p$$

$$= \partial_j (\partial_k p / p) = (\partial_j \partial_k p) / p - \partial_j p \partial_k p / p^2$$

$$= (\partial_j \partial_k p) / p - (\partial_j \log p)(\partial_k \log p)$$

パラメータ w によらず $\int p(x|w)dx=1$ が成り立つので、ルベーグの優収束定理を用いると前半の積分は0。後半の積分はフィッシャー情報量列の定義である（証明終）。

フィッシャー情報行列の例

例．正規分布（平均a,分散1/s) では

$p(x|a,s) = (s/2\pi)^{1/2} \exp(-(s/2)(x-a)^2)$ であり

$\log p(x|a,s) = -(s/2)(x-a)^2 + (1/2)\log s +$ 定数　　なので

$I_{11}(a,s) = -\int \{ -s \} p(x|a,s)dx = s$

$I_{12}(a,s) = -\int \{ x-a\} p(x|a,s)dx = 0$

$I_{22}(a,s) = \int \{-1/2s^2\} p(x|a,s)dx = 1/2s^2$

例．２項分布 $p(x|a) = a^x(1-a)^{1-x}$　では

$\log p(x|a) = x \log a +(1-x)\log(1-a)$ なので

$I(a) = -\int \{-x/a^2 - (1-x)/(1-a)^2\} p(x|a)\, dx = 1/a+1/(1-a)$

最尤推定量

定義. データ $X^n=(X_1,X_2,\ldots,X_n)$ が与えられたとき

$\Pi_i\, p(X_i|w)$ を最大にする w を w* と書いて最尤推定量という。

注意. データは確率変数なので、その関数である最尤推定量も
確率変数である(データ X^n が出るたびに確率的に変動する)。

例. 正規分布 $p(x|a,s) = (s/2\pi)^{1/2} \exp(-(s/2)(x-a)^2)$ では

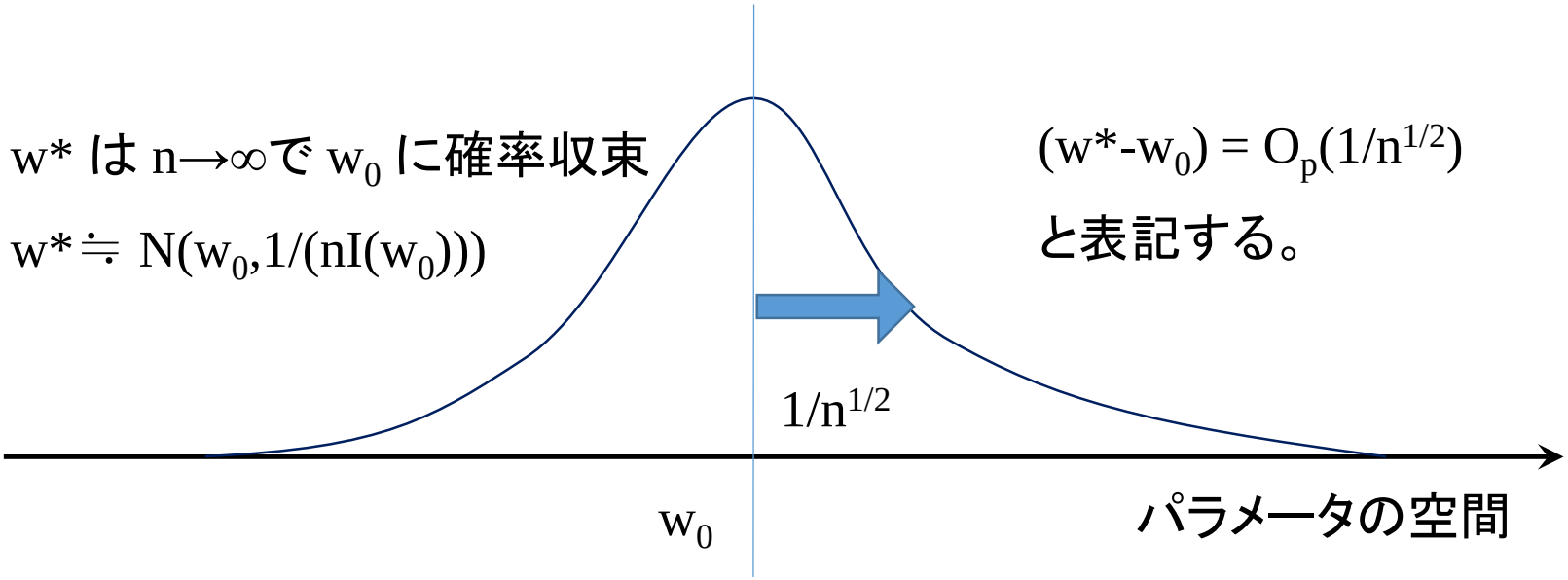
$$a^* = (\Sigma_i X_i) / n$$

$$1/s^* = \Sigma_i (X_i - a^*)^2 / n$$

2項分布 $p(x|a) = a^x(1-a)^{1-x}$ では $a^* = (\Sigma_i X_i) / n$ である。

最尤推定量の漸近正規性

定理（クラメールの定理）. データ $X^n=(X_1,X_2,\ldots,X_n)$ が独立に $p(x|w_0)$ に従うものとする。統計モデル $p(x|w)$ のフィッシャー情報行列 $I(w)$ が正定値とする。ある十分条件のもとで、データの数 n が∞に近づく極限で w^* は w_0 に確率収束し、$n^{1/2}(w^*-w_0)$ の確率分布は $N(0,I(w_0)^{-1})$ に収束することを示すことができる（最尤推定量の漸近正規性）。



最尤推定量の漸近正規性

注意. 最尤推定量が漸近正規性を満たすための十分条件には様々なものがある。例えば「対数尤度 $\log p(x|w)$ が3階までパラメータで微分可能で、それらの全ての絶対値が積分可能な関数でバウンドできる」などがある。ただし、最尤推定量が漸近正規性を満たすための必要十分条件というものはいまでも分かっていない。

簡単な統計モデルであれば最尤推定量は漸近正規性を満たすため、統計学の教科書には「ほとんどのモデルでOK」と記載されていることも多いが、今日の機械学習で用いられるカーネル法や深層学習やボルツマンマシンなどでは、最尤推定量は漸近正規性を満たさない。それらのモデルでは最尤推定量は精度の悪い推定を与えるので実務的にも好ましくない（がそれを理解できる統計学者は少ない）。

もちろん、「簡単なモデルの最尤推定量を計算せよ」という問題は統計関係の検定試験や期末試験における頻出問題であるから、計算はできるようになっておいたほうが良いと思います。

学習損失と汎化損失

学習損失と汎化損失

定義. データ $X^n=(X_1,X_2,\ldots,X_n)$ が独立に $p(x|w_0)$ に従うものとする。

学習損失（経験対数尤度の符号反転）と汎化損失（期待対数尤度の符号反転）をそれぞれ次式で定義する。

$$T(w) = - (1/n) \Sigma_i \log p(X_i|w)$$

$$G(w) = - \int \log p(x|w)\, p(x|w_0)\, dx$$

（注意） 関係式 $G(w) = - \int \log p(x|w_0)\, p(x|w_0)\, dx + KL(p(x|w_0) \,||\, p(x|w))$ が成り立つので、$G(w)$ が小さいほどKL情報量が小さい。

最尤推定量 w^* を用いたときの学習損失と汎化損失はそれぞれ

$$T(w^*) = - (1/n) \Sigma_i \log p(X_i|w^*)$$

$$G(w^*) = - \int \log p(x|w^*)\, p(x|w_0)\, dx$$

である。

汎化損失

定理.　汎化損失は次の漸近挙動を持つ(データ数 n, パラメータ数 d).

$$G(w^*) = G(w_0) + (1/2)\,\mathrm{tr}\,[\, I(w_0)\,(w_0 - w^*)\,(w_0 - w^*)^T\,] + O_p(1/n).$$

$$E[G(w^*)] = G(w_0) + (d/2n) + O(1/n).$$

(証明）　平均値の定理から（$|w_0-w^*|>|w_+-w^*|$を満たすw^+ が存在して

$$G(w^*) = G(w_0)+(w^*- w_0)\cdot\nabla G(w_0)+(1/2)(w^*- w_0)\cdot\nabla^2 G(w^+)(w^*- w_0).$$

パラメータ w_0 はKL情報量を最小化し $G(w)$ も最小にするので $\nabla G(w_0)=0$.

$$G(w^*) = G(w_0)+(1/2)(w^*- w_0)\cdot\nabla^2 G(w^+)(w^*- w_0)$$

$\nabla^2 G(w^+)$ は $I(w_0)$ に確率収束する.　$(w^*-w_0) = O_p(1/n^{1/2})$ を用いると

$$G(w^*) = G(w_0)+(1/2)(w^*- w_0)\cdot I(w_0)\,(w^*- w_0) + O_p(1/n).$$

$(w_0 - w^*)\,(w_0 - w^*)^T$ の平均は w_0 の分散共分散であるから $I(w_0)$ に収束する。

(証明終）

学習損失

定理.　学習損失は次の漸近挙動を持つ.

$$T(w^*) = T(w_0) - (1/2)\,\mathrm{tr}\,[\, I(w_0)\,(w_0 - w^*)\,(w_0 - w^*)^T \,] + O_p(1/n).$$

$$E[T(w^*)] = G(w_0) - (d/2n) + O(1/n).$$

(証明）　平均値の定理から（$|w_0-w^*|>|w_+-w^*|$を満たすw^+ が存在して

$$T(w_0) = T(w^*) + (w_0 - w^*)\cdot\nabla T(w^*) + (1/2)(w_0 - w^*)\cdot\nabla^2 T(w^+)(w_0 - w^*).$$

最尤推定量は $T(w)$ を最小にするので $\nabla T(w^*)=0$. 従って

$$T(w_0) = T(w^*) + (1/2)(w_0 - w^*)\cdot\nabla^2 T(w^+)(w_0 - w^*)$$

$\nabla^2 T(w^+)$ は $I(w_0)$ に確率収束する.　$(w^*-w_0) = O_p(1/n^{1/2})$ を用いると

$$T(w_0) = T(w^*) + (1/2)(w_0 - w^*)\cdot I(w_0)\,(w_0 - w^*) + O_p(1/n)$$

従って　$T(w^*) = T(w_0) - (1/2)(w_0 - w^*)\cdot I(w_0)\,(w_0 - w^*) + O_p(1/n).$

$T(w_0)$の平均値は $G(w_0)$である。　(証明終）

何がわかったか

最尤推定量 w^* を用いたときの学習損失と汎化損失

$$T(w^*) = - (1/n) \Sigma_i \log p(X_i|w^*)$$

$$G(w^*) = - \int \log p(x|w^*) \, p(x|w_0) \, dx$$

である。$T(w^*)$ も $G(w^*)$ も $n \rightarrow \infty$ のとき $G(w_0)$ に収束するが、

有限の n においては

$$E[T(w^*)] = G(w_0) - (d/2n) + O(1/n).$$

$$E[G(w^*)] = G(w_0) + (d/2n) + O(1/n).$$

すなわち、汎化誤差は学習誤差よりも平均的に d/n 程度大きい。

$$E[G(w^*)] = E[T(w^*)] + (d/n) + O(1/n).$$

学習誤差の平均値から汎化誤差の平均値が予測できる（赤池弘次）。

注意

機械学習で用いられるカーネル法や深層学習やボルツマンマシンなどでは、最尤推定量は漸近正規性を満たさないので、汎化誤差や学習誤差も上記の定理を満たさない。機械学習では汎化誤差を小さくするパラメータを見つけることが大きな目標のひとつであり、そのときに学習誤差から汎化誤差を予測する数学的公式を見出すことが望まれているが、それはいまだにできていない。

深層学習は今日、非常に多くの問題に適用されているが、その基礎にある数学が構築されていないため、理論とアルゴリズムがヒューリスティクスの段階から進歩できないでいる。

従来の統計理論はすべて利用できない。→　新しい数学を零から作って欲しい。

→若いみなさんの目標です。→他人を頼らず自分でやれば。

→代数幾何やゼータ関数がどうしても必要なので道は険しい。←いまここ

赤池情報量規準

AICの定義

最尤推定量 w^* を用いたときの学習損失と汎化損失

$$T(w^*) = - (1/n) \Sigma_i \log p(X_i|w^*)$$

$$G(w^*) = - \int \log p(x|w^*) \, p(x|w_0) \, dx$$

すなわち、汎化誤差は学習誤差よりも平均的に d/n 程度大きい。

$$E[G(w^*)] = E[T(w^*)] + (d/n) + O(1/n).$$

赤池情報量規準 (AIC, 1974) を

$$AIC = 2n \, T(w^*) + 2d = - 2 \Sigma_i \log p(X_i|w^*) + 2d$$

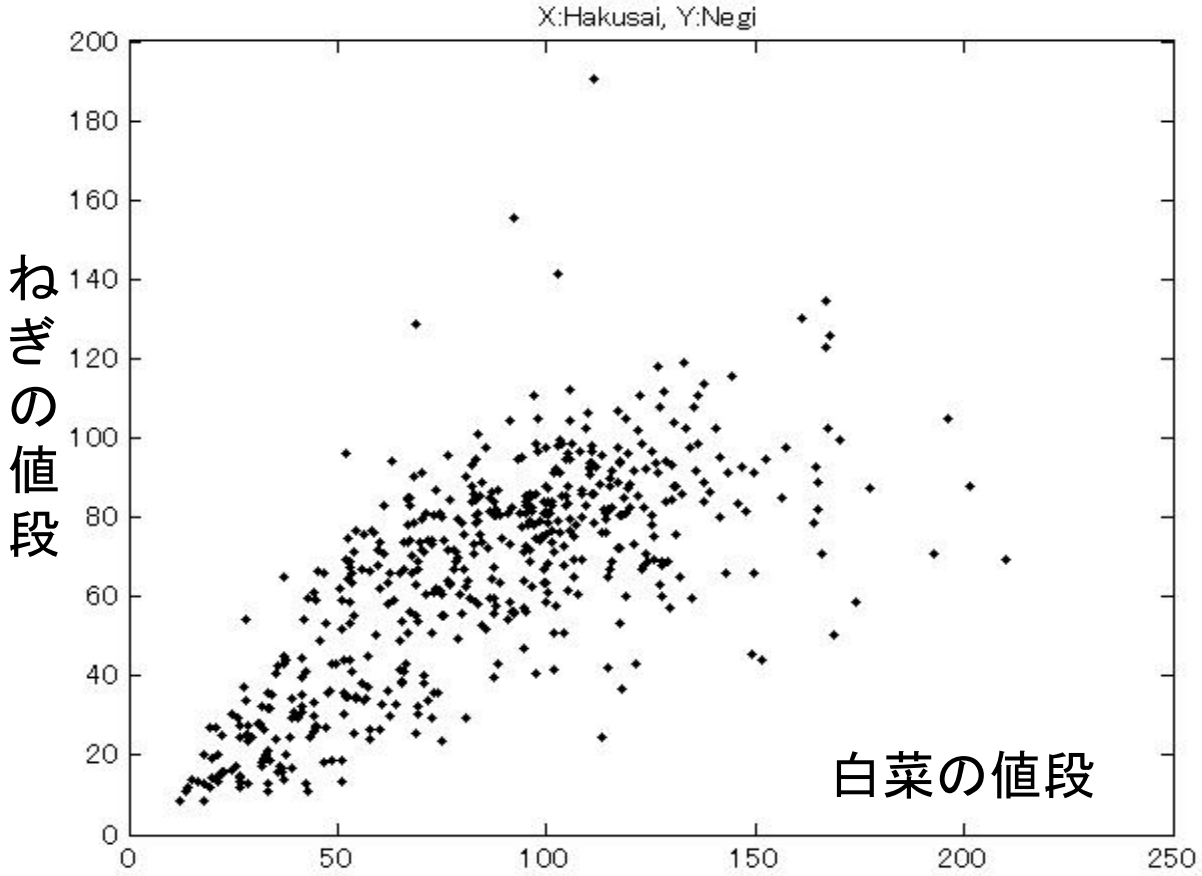
と定義すれば　$E[AIC] = 2n \, E[G(w^*)] + O(1)$ が成り立つ。

異なる統計モデルを AIC を用いて比較することができる

（小さいほうが平均的に汎化損失を小さくする）。

実データに使ってみる

時系列予測の例：1970年1月から2013年12月までの白菜とねぎの値段
「政府統計の総合窓口」のデータを使用しています。
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do



白菜の値段からねぎの値段を知りたいとき、
どんなモデルを使うとよいのだろうか。(n=50)

いろいろなモデル

真の関係はわからない。多項式モデルを複数考えてみる。

$$p(y|x,w,\sigma) = 1/(2\pi\sigma^2)^{1/2} \exp(- (1/2\sigma^2)(y-f(x,w))^2)$$

$Y=a+$雑音$(0,\sigma^2)$

$Y=a+bX+$雑音$(0,\sigma^2)$

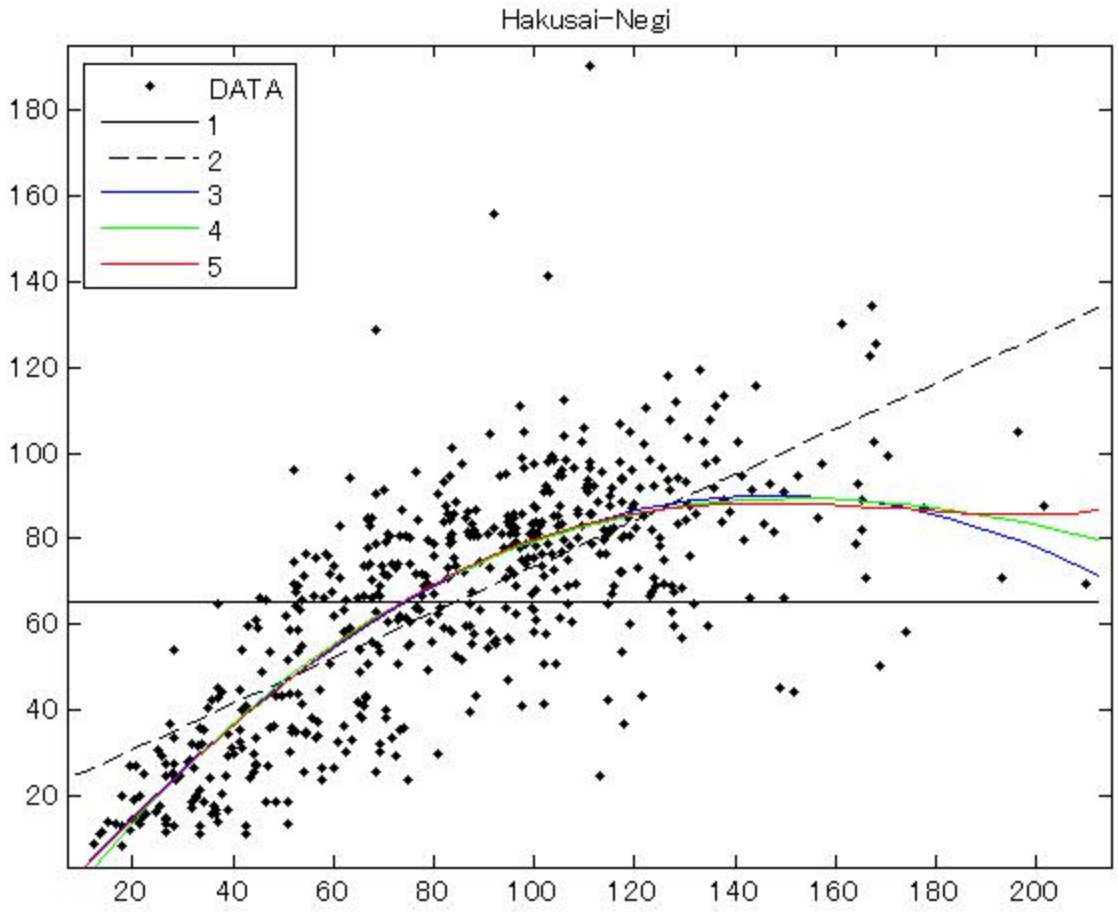
$Y=a+bX+cX^2+$雑音$(0,\sigma^2)$

$Y=a+bX+cX^2+dX^3+$雑音$(0,\sigma^2)$

$Y=a+bX+cX^2+dX^3+eX^4+$雑音$(0,\sigma^2)$

二乗誤差を最小にするパラメータは簡単に求められるが・・・。

学習してみた



グラフを見てもどれが良いのかわからない・・・。
高次元だとグラフを見ることもできない・・・。

AICの計算法

統計モデル

$$p(y|x,w,\sigma) = 1/(2\pi\sigma^2)^{1/2} \exp(- (1/2\sigma^2)(y-f(x,w))^2)$$

に対して最尤推定量 (w^*,σ^*) を計算して

$$AIC = -2 \ \Sigma \log p(Yi|X,w^*,\sigma^*) + 2d$$

を求めればよい。

モデル番号	AIC(赤池情報量規準)
1	1751.5
2	1560.2
3	1508.2
4	1509.6
5	1511.3

AICの性質

AICは1974年、日本人の統計学者赤池弘次博士により考案された。それよりも前は、合理的に統計モデルを評価するという考え方自体が存在しなかったので、この業績は「合理的に統計モデルを評価することを提案し、数学的な解決を与えたもの」として今日の統計学の大きな基礎になったものである。

☆　想定しているモデルの集合の中に真のデータを発生している分布があるとき、
データの数 $n \to \infty$のとき、真のデータを発生している（パラメータ数が一番少ない）
モデルを選ぶことができる時、そのモデル選択法を「一致性を持つ」という。
☆　考察している有限のモデルの集合の中にデータを発生している分布がないとき
データの数 $n \to \infty$のとき、最も汎化誤差を小さくするモデルを選ぶことができるとき
そのモデル選択法を「有効性を持つ」という。
AICは一致性は持たないが、有効性をもつケースがあることが知られている。