

学習理論の練習 6

渡辺澄夫

いつも何度でも復習

復習：カルバック・ライブラ情報量

定義. \mathbf{R}^N 上の確率密度関数 $q(x)$, $p(x) > 0$ が与えられたとき

$D(q||p) = \int q(x) \log (q(x)/p(x)) dx$ を $q(x)$ と $p(x)$ のカルバック・ライブラ情報量(KL情報量)という。

定理. 連続な確率密度 $q(x)$, $p(x)$ について、次が成り立つ。

- (1) 任意の $q(x)$, $p(x) > 0$ について $D(q||p) \geq 0$.
- (2) $D(q||p) = 0 \Leftrightarrow q(x) = p(x) (\forall x)$

$$\begin{aligned} D(q||p) &= \int q(x) \log (q(x)/p(x)) dx \\ &= - \int q(x) \log p(x) dx - \left\{ - \int q(x) \log q(x) dx \right\} \\ &= G(p) - S \end{aligned}$$

KL情報量 = 汎化損失 - エントロピー

復習： ベイズ推定

1. 未知の $q(x)$ からデータ X^n が発生
2. 解析者が $\varphi(w)$ と $p(x|w)$ をテキストに(人間力で)準備

3. w の事後分布 $p(w|X^n) = (1/Z) \varphi(w) \prod_i p(X_i|w)$

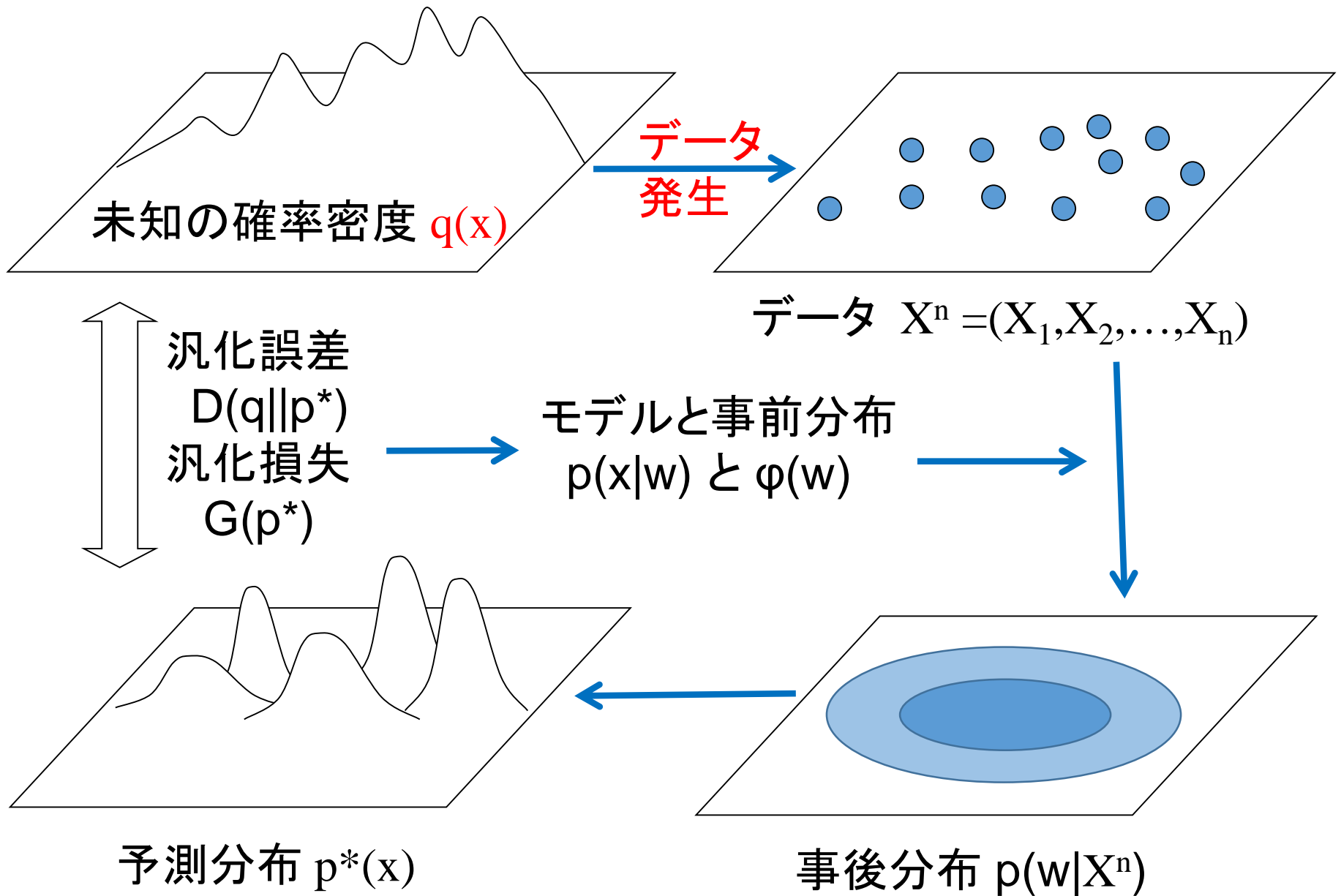
ここで $Z = \int \varphi(w) \prod_i p(X_i|w) dw$

4. 予測分布 $p^*(x) = \int p(x|w) p(w|X^n) dw$

5. 汎化損失 $G(p^*) = - \int q(x) \log p^*(x) dx$

6. $G(p^*)$ が小さくなるように $\varphi(w)$ と $p(x|w)$ を改良

ベイズ推測の手順



汎化損失と学習損失

何が未知で 何が既知なのか

真の分布 $q(x)$ はわからない。

予測分布 $p^*(x)$ がどのくらい正しいかはわからない。

汎化損失 $G(p^*) = - \int q(x) \log p^*(x) dx$

が小さいほど良い予測であるが、 $G(p^*)$ は $q(x)$ がわからなければ計算できない。

モデルと事前分布は人間がテキトーに決めたものである。

もしも $G(p^*)$ が求められれば、モデルと事前分布を

$G(p^*)$ が小さくなるように改良できるが...

いろいろな平均があるので表記に慣れよう

学習データ $X^n = (X_1, X_2, \dots, X_n)$ についての平均

$$E[f(X_1, X_2, \dots, X_n)]$$

$$= \iint \cdots \int f(x_1, x_2, \dots, x_n) q(x_1) q(x_2) \cdots q(x_n) dx_1 dx_2 \cdots dx_n$$

テストデータ X についての平均 $E_x[f(X)] = \int f(x) q(x) dx$

事後分布による平均

$$E_w[f(w)] = \frac{\int f(w) \prod_{i=1}^n p(X_i|w) \varphi(w) dw}{\int \prod_{i=1}^n p(X_i|w) \varphi(w) dw}$$

損失と平均

$$\text{予測分布} \quad p^*(x) = E_w[p(x|w)] = p(x|X^n)$$

$$\text{汎化損失} \quad G_n = - E_x[\log p(X|X^n)]$$

$$\text{学習損失} \quad T_n = -(1/n) \sum_{i=1}^n \log p(X_i|X^n)$$

(注意)

一般に関数 $f(x)$ について $E[(1/n) \sum_{i=1}^n f(X_i)] = E_x[f(X)]$ が成り立つ。しかし学習損失は条件部分と推測部分の両方に学習データがあるので $E[T_n] \neq E[G_n]$ である。学習損失で汎化損失を推測することはできない。

交差検証

ひとつめき交差損失(LOOCV)

ひとつめき交差損失の定義

集合の引き算

$$\text{LOOCV} = - (1/n) \sum_{i=1}^n \log p(X_i | X^n - X_i)$$

定理. X^n が独立なら $E[\text{LOOCV}] = E[G_{n-1}]$ (平均の存在は仮定)

(証明)

$$\begin{aligned} E[\text{LOOCV}] &= - (1/n) \sum_{i=1}^n E[\log p(X_i | X^n - X_i)] \\ &= E[\log p(X_n | X^n - X_n)] = E[\log p(X_n | X^{n-1})] \\ &= E E_x[\log p(X | X^{n-1})] = E[G_{n-1}] \end{aligned}$$

交差損失の計算法

LOOCV を計算するには事後分布を n 回作らなくてはならないので計算量が多大になります。ベイズ法では次の式を用いると1個の事後分布で計算できます。

$$C_n = (1/n) \sum_{i=1}^n \log E_w[1/p(X_i|w)]$$

定理. LOOCV = C_n

(証明)

$$\begin{aligned} p(X_k | X^n - X_k) &= \frac{\int p(X_k|w) \prod_{i=1, i \neq k}^n p(X_i|w) \varphi(w) dw}{\int \prod_{i=1, i \neq k}^n p(X_i|w) \varphi(w) dw} \\ &= \frac{\int \prod_{i=1}^n p(X_i|w) \varphi(w) dw}{\int 1/p(X_k|w) \prod_{i=1, i \neq k}^n p(X_i|w) \varphi(w) dw} = \frac{1}{E_w[1/p(X_k|w)]} \end{aligned}$$

学習損失は交差損失より小さい

定理. $T_n \leq C_n$. 等号成立は $p(x|w)$ が w の定数関数

(証明) 学習損失 $T_n = -\frac{1}{n} \sum_{i=1}^n \log E_w[p(X_i|w)]$

交差損失 $C_n = \frac{1}{n} \sum_{i=1}^n \log E_w[1/p(X_i|w)]$

$$C_n - T_n = \frac{1}{n} \sum_{i=1}^n \log \{ E_w[1/p(X_i|w)] E_w[p(X_i|w)] \}$$

コーシーシュワルツの不等式を使う。 $f(w) > 0$ なら

$$1 = E_w[(1/f(w)) f(w)] \leq E_w[1/f(w)^2]^{1/2} E_w[f(w)^2]^{1/2}$$

「等号成立 $\Leftrightarrow f(w) = \text{定数}$ 」より定理が得られる。

実験例

$N(0, 1/s)$ を平均 0, 分散 $1/s$ の正規分布を表すものとする。

$f_m(x, a)$: $(m-1)$ 次多項式 a :パラメータ (m 個)

回帰モデル $Y = f_m(x, a) + N(0, 1/s)$ を表す確率密度関数は

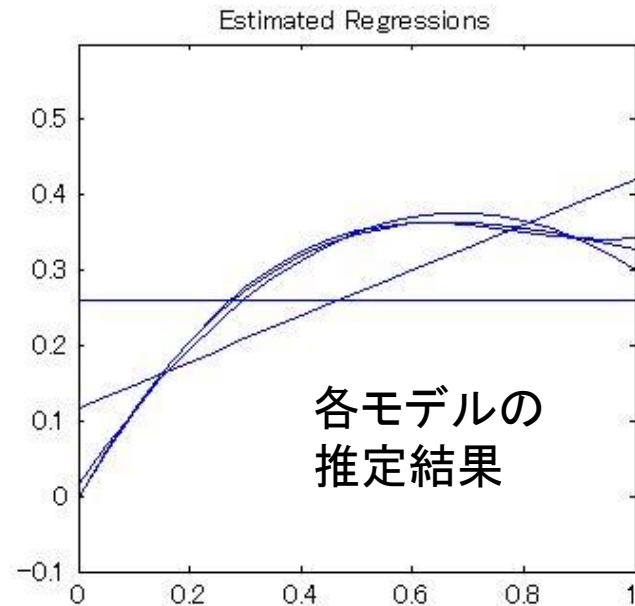
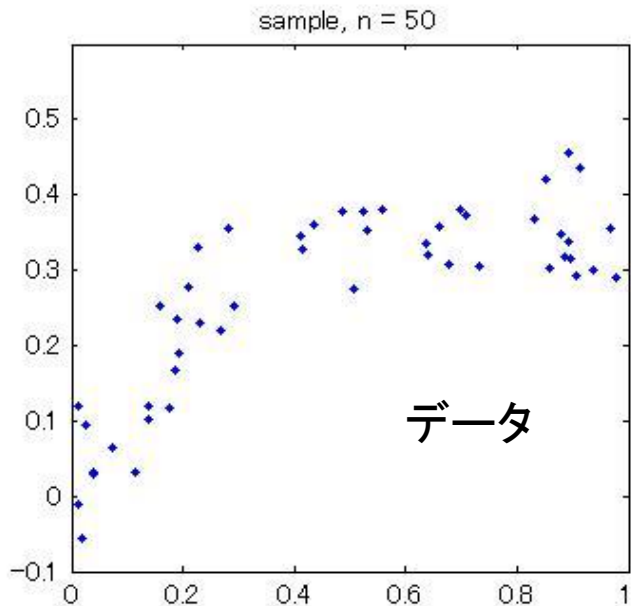
$$p(y|x, a, s) = (s / 2\pi)^{1/2} \exp(-s(y - f_m(x, a))^2 / 2).$$

事前分布として次のものを用いる。

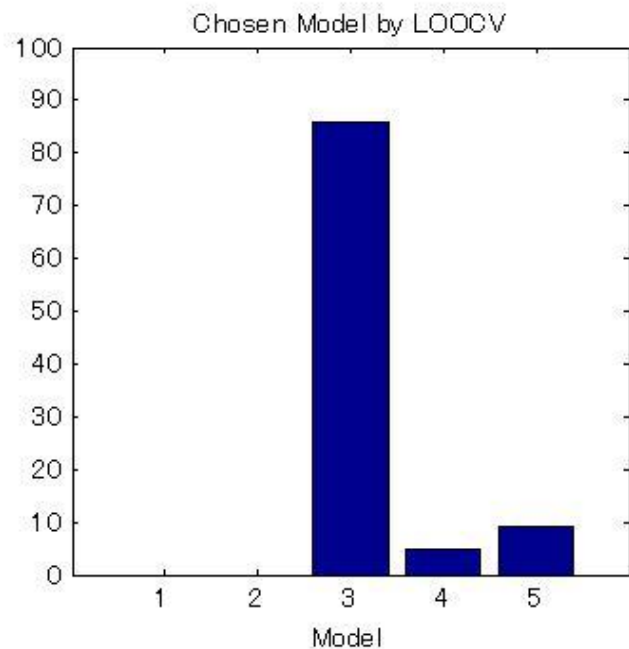
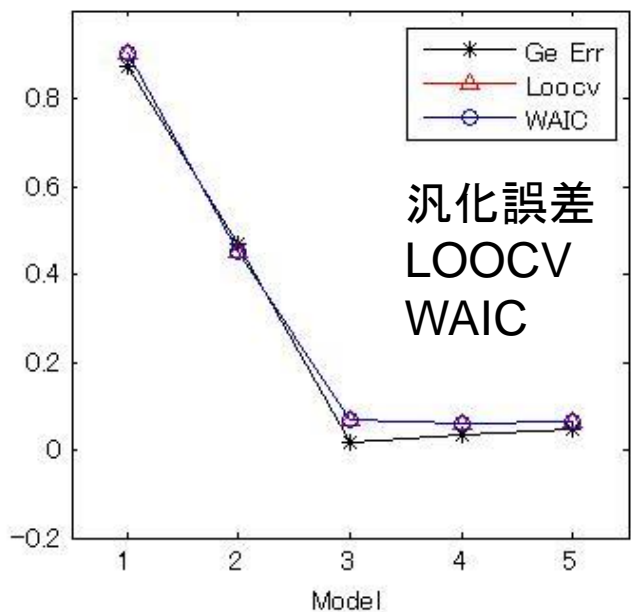
$$\varphi(a, s) \propto \exp(-\tau s (r + \|a\|^2) / 2)$$

ただし τ は十分に小さく設定した。

実験の結果



LOOCVで選ばれたモデル



ここまでの物語

- ◎ 人間は真 $q(x)$ を知ることはできない。
- ◎ データ X^n は得られる。
- ◎ とりあえず $\varphi(w)$ と $p(x|w)$ をテキトーに決めて予測 $p(x|X^n)$ は作れる。
- ◎ 汎化損失 G_n が小さくなるように $\varphi(w)$ と $p(x|w)$ を改良したい。
- ◎ 人間は汎化損失 G_n を求めることはできない。
- ◎ 学習損失 T_n を求めることはできるが、それで G_n は推測できない。
- ◎ ところが、ひとつめき交差損失 C_n を求めることはできる。
- ◎ C_n と G_{n-1} は平均値が等しい ←いまここ

だんだん厳しく難しくなる

- ◎ C_n を小さくすると G_n も小さくなるのか？
- ◎ C_n よりも精度よく G_n を推定する方法はないのか？
- ◎ そもそも G_n を理論で計算できればいいのに。それは無理？