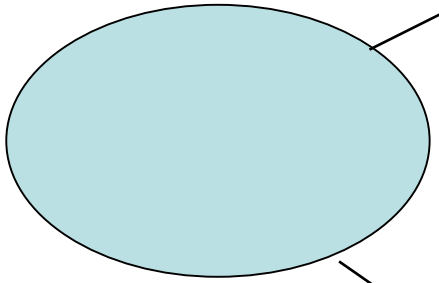
情報学習理論

渡辺澄夫
東京工業大学

教師なし学習

学習データ

$X_1, X_2, \cdots, X_n$



<u>真の情報源</u>

テストデータ

X

真の確率分布を
探求する・・・

教師なし学習の目標の例

たくさんの例が与えられたとき

(1) 代表例をあげる　← 先週まで
　　K-Means, 競合学習

(2) 空間の地図を作る　← 先週まで
　　自己組織化写像

(3) 情報源の確率分布を知る　← 今週ここ
　　混合正規分布　ボルツマンマシン

真の情報源を知る

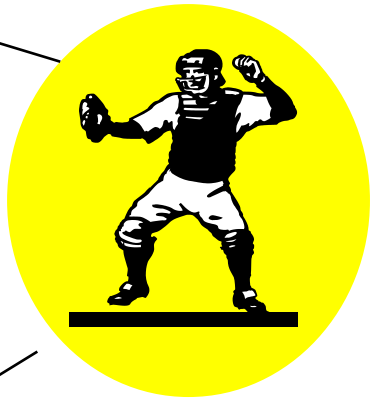
学習データ

$X_1, X_2, \cdots, X_n$

テストデータ

X

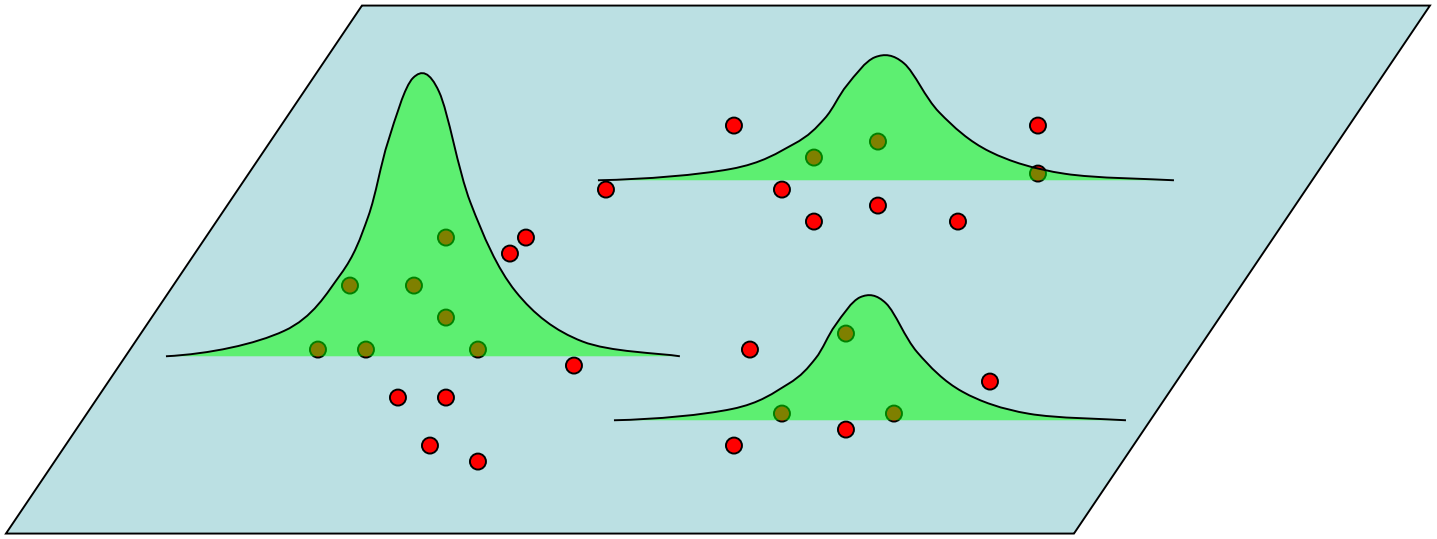
情報源
q(x)



学習モデル
p(x|w)

牧場の草の減り方から牛の分布を推定する

混合正規分布



平均 b_k ,分散σ_k^2 の正規分布の重み $\{a_k\}$ の和

混合正規分布の定義式

x : M 次元ベクトル

パラメータ　$w = (a_k , b_k , \sigma_k)$

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right)$$

$$\sum_{k=1}^{K} a_k = 1$$

平均 b_k ,分散σ_k^2 の
正規分布

混合正規分布の使い方

高次元空間上に複数の要因に基づくと思われるデータがあるとき、そのデータを発生した確率分布を推測する。K-means と似ているが、混合正規分布では、情報源の確率分布まで推測できる。

（例）1000人の中学生について、5次元のデータ(国語、数学、理科、社会、英語）があったとき、どのような群があるか知りたい。また、それぞれの群の平均と分散を知りたい。

隠れ変数（潜在変数）の導入

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{M/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right)$$

y について和をとる

$$p(x,y|w) = \prod_{k=1}^{K} \left[a_k \frac{1}{(2\pi\sigma_k^2)^{M/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right) \right]^{y_k}$$

$y = (y_1, y_2, .., y_k)$ はひとつだけ1で残りは0。つまり
$y \in \{(1,0,..,0), (0,1,0,\ldots,0), \ldots, (0,0,\ldots,1)\} \equiv C^K$

混合分布の性質

混合正規分布の学習は

各 x がどのコンポーネントから発生したかの情報 y (隠れ変数, 潜在変数）

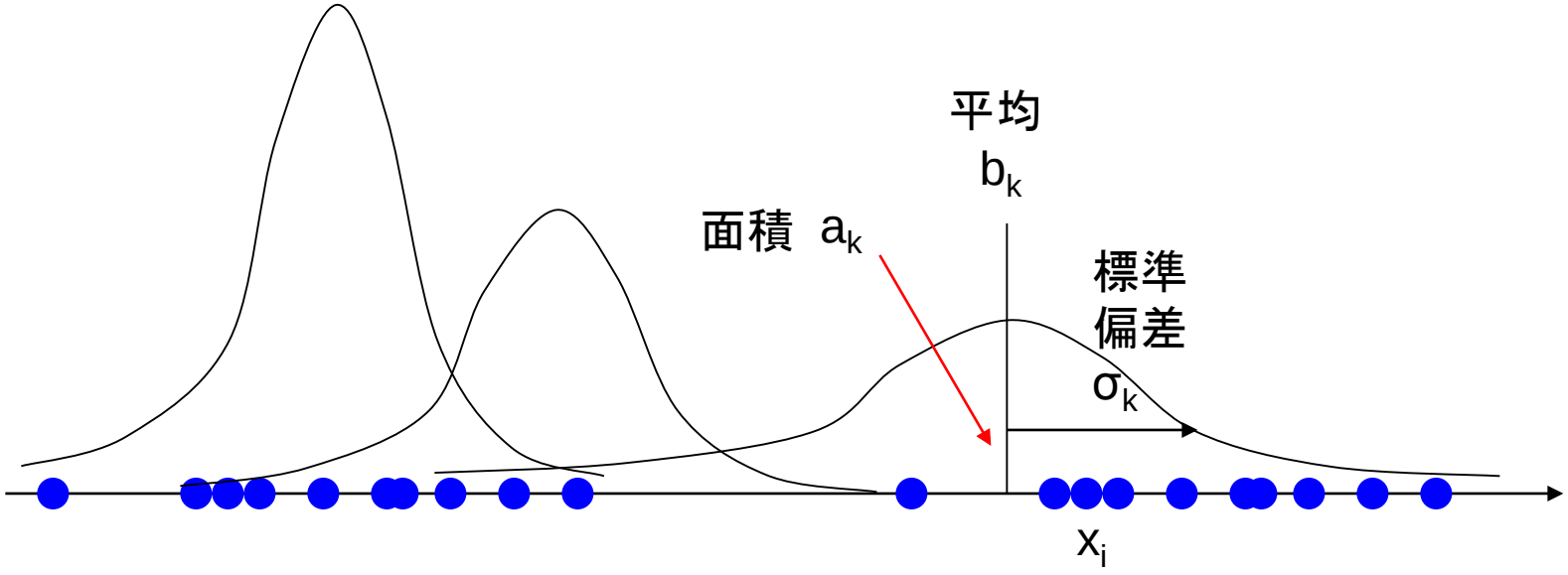
を計測できない場合の学習と等価である。

$$p(x|w) \quad \Leftrightarrow \quad p(x,y|w),\ y \text{ は隠れ変数}$$

Expectation -Maximization法

(1) パラメータ初期化
(2) 隠れ変数推定→パラメータ推定
　　→隠れ変数推定･･･の繰り返し

このとき尤度が単調非減少であることを証明できる。
局所的に尤度の大きな解が得られる。



パラメータ→隠れ変数推定

パラメータ w がわかれば
データ x_i に対する隠れ変数が求められる。

$$E[\, y_k | \, x_i \, , \, w]$$

$$= \Sigma_y \; y_k \, p(y | \, x_i, \, w)$$

$$= \Sigma_y \; y_k \, p(x_i, \, y | \, w) \, / \, p(x_i | w)$$

$$= \frac{a_k \; \frac{1}{(2\pi\sigma_k{}^2)^{N/2}} \exp\left(- \; \frac{\| \; x_i - b_k \|^2}{2\sigma_k{}^2} \right)}{p(x_i | w)}$$

隠れ変数→パラメータ推定

隠れ変数がわかるとパラメータが推定できる。

$$a_k = \frac{\Sigma\, E[y_k|\, x_i\, ,\, w]}{n}$$

（Σは i=1,..,n の和）

$$b_k = \frac{\Sigma\, E[y_k|\, x_i\, ,\, w]\, x_i}{\Sigma\, E[y_k|\, x_i\, ,\, w]}$$

$$\sigma_k^{\,2} = \frac{\Sigma\, E[y_k|\, x_i\, ,\, w]\, \| x_i - b_k \|^2}{\Sigma\, M\, E[y_k|\, x_i\, ,\, w]}$$

EM法の計算

両方の手続きを行う　＝　次の計算を繰り返せはよい。

$$a_k = \frac{\Sigma\, E[y_k | x_i , w_1]}{\Sigma\, 1}$$

$$b_k = \frac{\Sigma\, E[y_k | x_i , w_1]\, x_i}{\Sigma\, E[y_k | x_i , w_1]}$$

$$\sigma_k^{\,2} = \frac{\Sigma\, E[y_k | x_i , w_1] \,\| x_i - b_k \|^2}{M\Sigma\, E[y_k | x_i , w_1]}$$

$w_1 \rightarrow w_2$ が計算できる

（Σはすべてi=1,..,n の和を表す）

アルゴリズムの発展

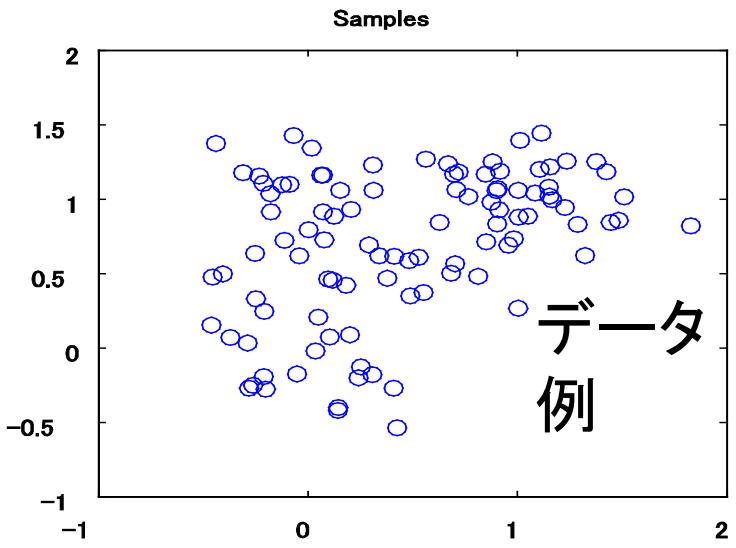
(1) EM 法(1977)の問題点

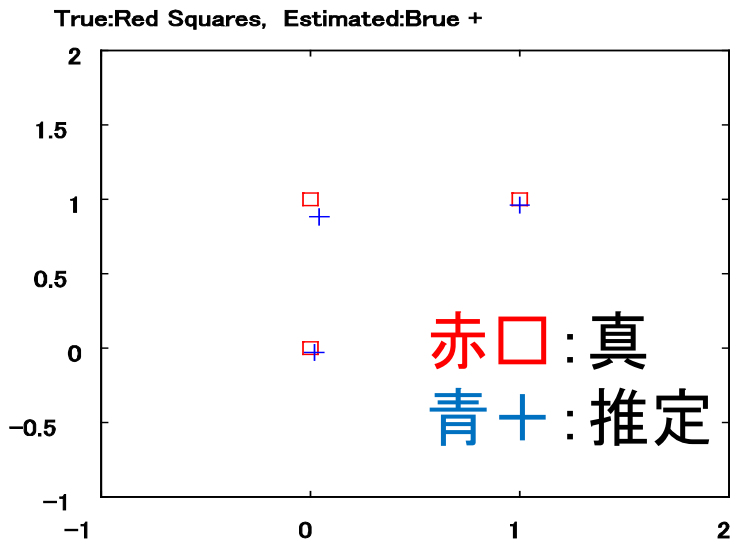
EM法は最尤推定量をみつけようとする方法であるが
混合正規分布では最尤推定量は存在しないので
EM法が何をしているのかは実はよくわからない。
データの揺らぎに対して弱く、汎化誤差も大きい。

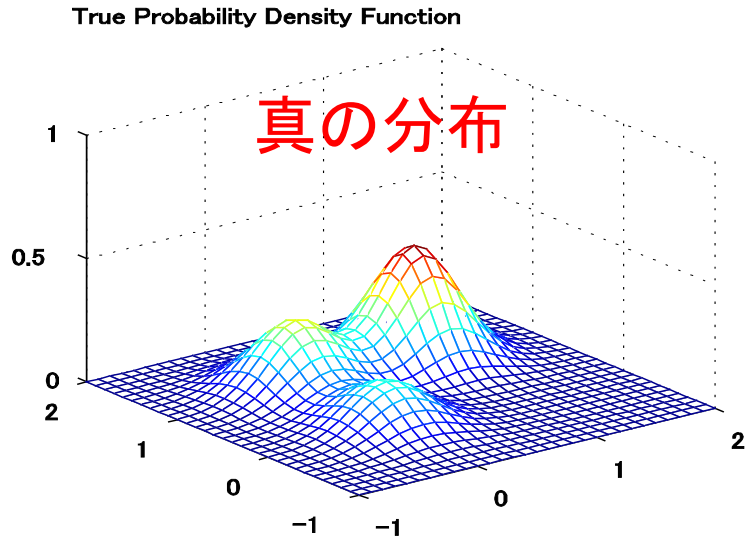
(2) VB 変分ベイズ法(1999)の発見

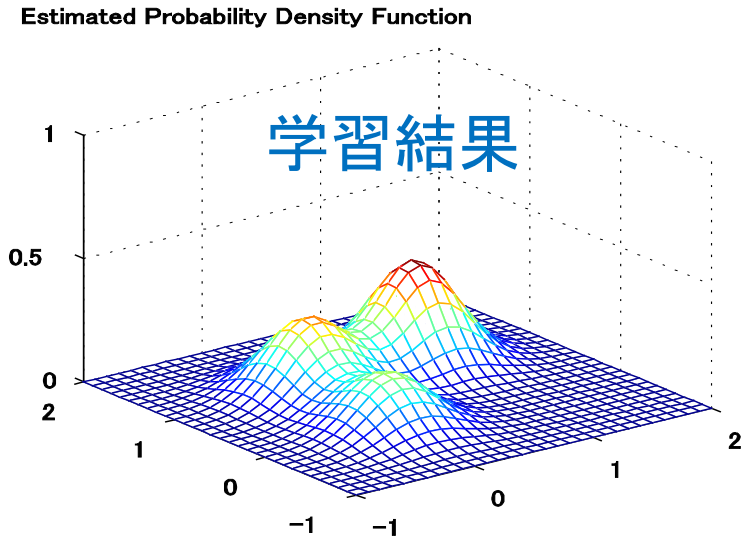
EM法とよく似た学習アルゴリズムを与える変分ベイズ法は
ベイズ事後分布の平均場近似により導かれるが、
データの揺らぎに対して強く、汎化誤差も小さくできる。
隠れ変数を持つシステムにおいて広く使われている。

混合正規分布で推定してみた（変分ベイズ法）









変分ベイズ法　あれこれ

(1)「パラメータと隠れ変数が独立な分布」から事後分布までのKL情報量を最小化する方法を平均場近似という。

(2) 事後分布の数値実現にはマルコフ連鎖モンテカルロ法が用いられるが、平均場近似は繰り返し代入法で実行できる。

(3) 平均場近似によるパラメータの分布は真の事後分布よりも縮んでいる。

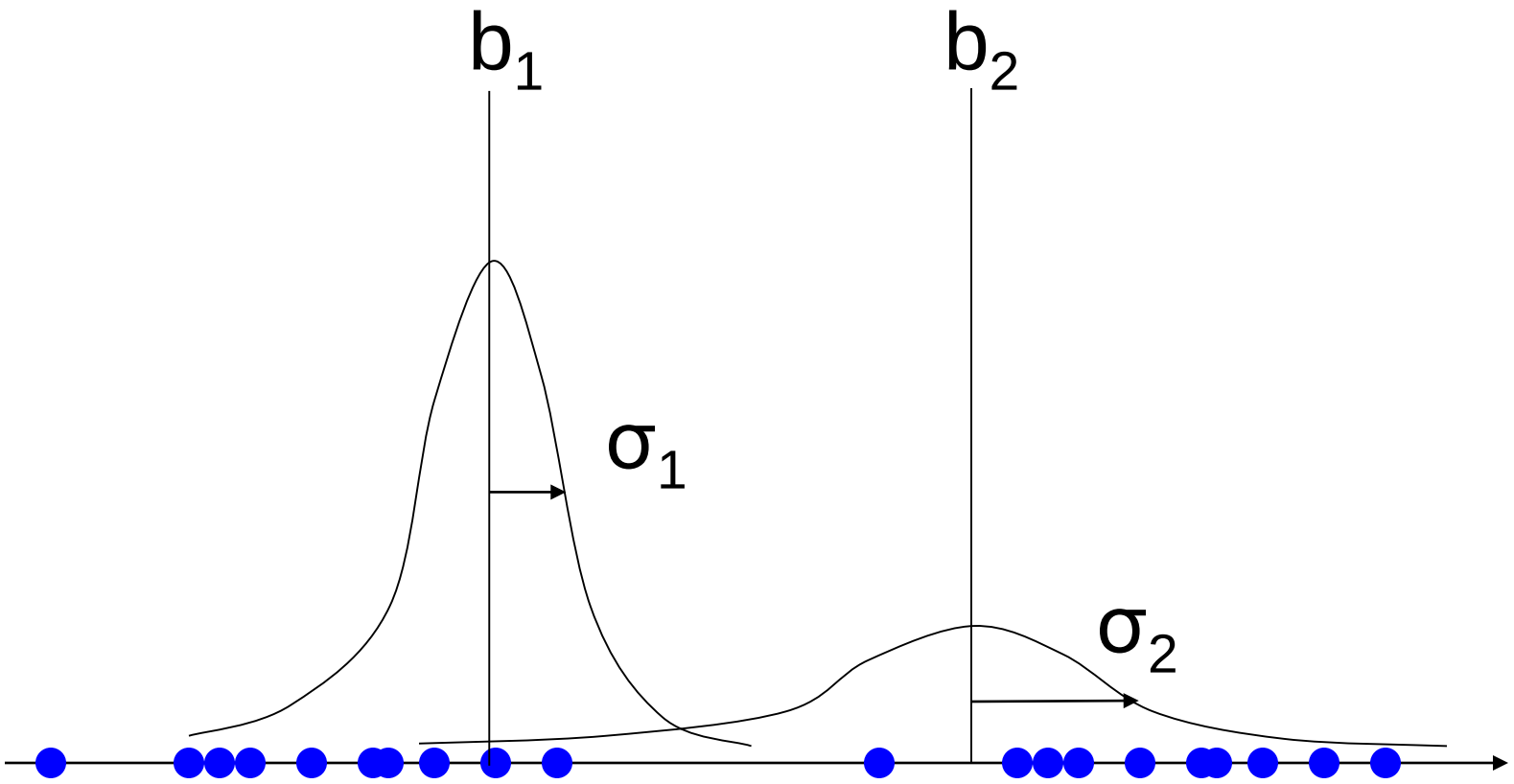
(4) 平均場近似の自由エネルギーは真の自由エネルギーよりも少し大きい。ハイパーパラメータの変化による相転移が存在する（渡辺一帆）。

(5) 平均場近似の汎化誤差の挙動は未解決。縮小ランク回帰モデルではランダム行列理論を用いて数学的に解ける（中島伸一）。

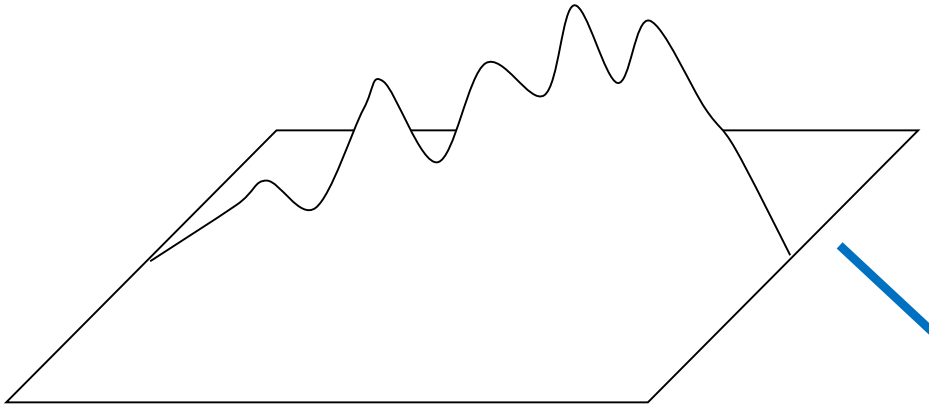
(6) 自由エネルギー・汎化誤差・相転移は数値実験では解析できないので専門書でも論文でも誤った記載が多い（PRMLなども間違っている）。間違っていることを知らない研究者も多いので注意しましょう。

問1

EMアルゴリズムを1回動かしたとき
各パラメータがどのように変化するか図示せよ。

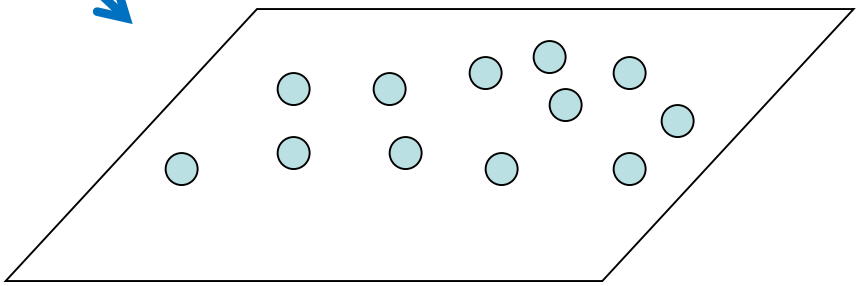


学習の良さをどのように判定しよう



真の確率分布 $q(x)$

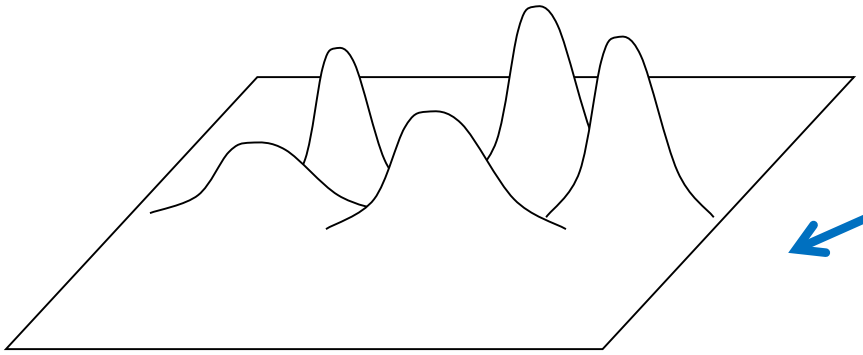
データを発生



例 $X^n = (X_1, X_2, \ldots, X_n)$

推測は真と
どれくらい近い？

学習



推測された分布 $p(x|w)$

学習の問題とは

現実の問題では真の分布は不明であり、得られた学習結果がどのくらい正しいかはわからない。

現実の問題では真の分布は不明であり、自分の用いた学習モデルが適切であったかどうかもわからない。

学習におけるこの原理的な問題に対して統計学で考案されたモデルの良さを知るための量は二つある。

汎化誤差：真の分布と推定された分布のKL情報量

自由エネルギー：推定に用いたモデルの -log（確率）

汎化誤差

真の分布を q(x) とし学習の結果を p(x) とするとき汎化誤差は

$$G = \int q(x) \log (q(x) / p(x)) \, dx$$

汎化誤差は真の分布がわからないときには計算できないが

$$G = \int q(x) \log \, q(x) \, dx \quad - \int q(x) \log p(x) \, dx$$

なので、対数損失（右辺第２項）が小さいことと汎化誤差が小さいことは等価。対数損失は、クロスバリデーションや情報量規準によって推測できる。

（注1）混合正規分布は正則でないので AIC, TIC, DIC は使えない。

自由エネルギー

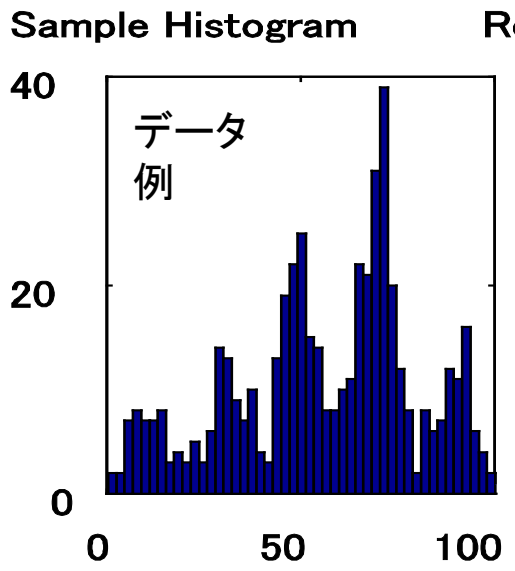
確率モデル p(x|w) と事前分布 φ(w)が与えられたとき、その自由エネルギーを下記で定義する。

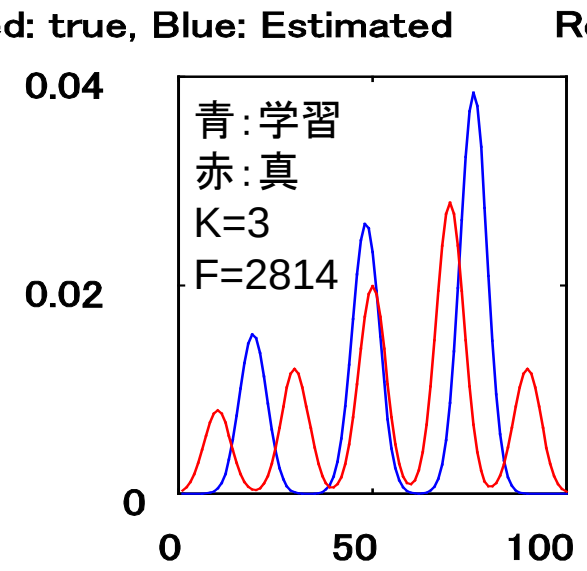
$$F = -\log \int \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw$$

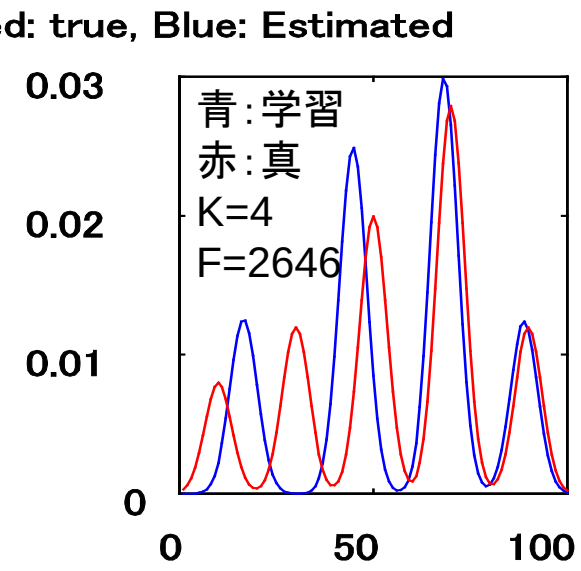
exp(-F) は、データ $X_1, X_2, \ldots, X_n$ が与えられたときの確率モデルと事前分布の尤度であり、これを最大化（Fを最小化）することでモデルと事前分布の評価ができる。

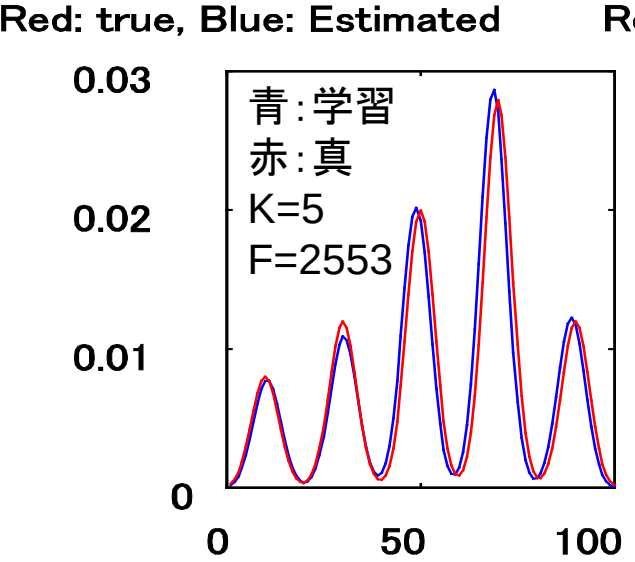
（注1）F が小さいことと G が小さいこととは等価ではない。
(注2）一般に F の計算は大きな計算量が必要になる。
今日は平均場近似を使う。

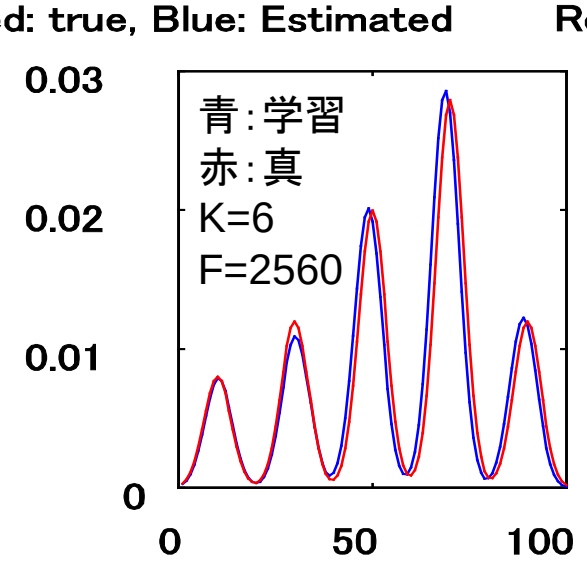
学習の例(変分ベイズ法)

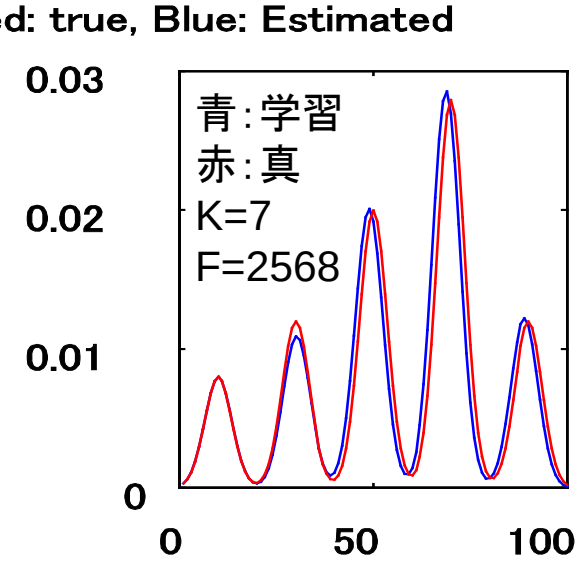

Sample Histogram
データ
例
40
20
0
0
50
100


Red: true, Blue: Estimated
青:学習
赤:真
K=3
F=2814
0.04
0.02
0
0
50
100


Red: true, Blue: Estimated
青:学習
赤:真
K=4
F=2646
0.03
0.02
0.01
0
0
50
100


Red: true, Blue: Estimated
青:学習
赤:真
K=5
F=2553
0.03
0.02
0.01
0
0
50
100


Red: true, Blue: Estimated
青:学習
赤:真
K=6
F=2560
0.03
0.02
0.01
0
0
50
100


Red: true, Blue: Estimated
青:学習
赤:真
K=7
F=2568
0.03
0.02
0.01
0
0
50
100

現実の問題では

現実のデータが、ぴったりＫ個の正規分布から出ているということはめったに起こらない。

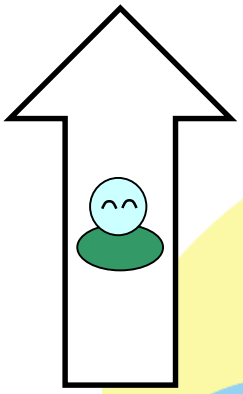
データの個数は有限なので、その個数に応じた解像度で真の情報源がわかる。

データの個数が増えるにつれて少しずつ詳しい情報が分かっていく。適切なクラスターの個数も少しずつ増えていく。

汎化誤差は予測の正確さを、自由エネルギーはモデルの確率的な適切さをそれぞれ表している。

次週予告:「天空城ボルツマンマシン」

学習理論の秘密を求めて
旅する主人公はついに
天空城ボルツマンマシンへ。
輝きを放つ伝説の剣・・・。
最後の敵が姿を現す・・・。



ボルツマンマシン

深層学習

ニューラルネットワーク

教師なし学習

自己組織化

競合学習

教師あり学習

サポートベクタマシン

問2

混合正規分布モデルの混合数を

自由エネルギー　Fを見ることで評価してみよう。

混合数	3	4	5	6	7
F					
F					
F					
最小は					