ベイズ統計　演習 ②

前回の復習

確率空間$(\Omega,\mathcal{B},P)$ を復習した

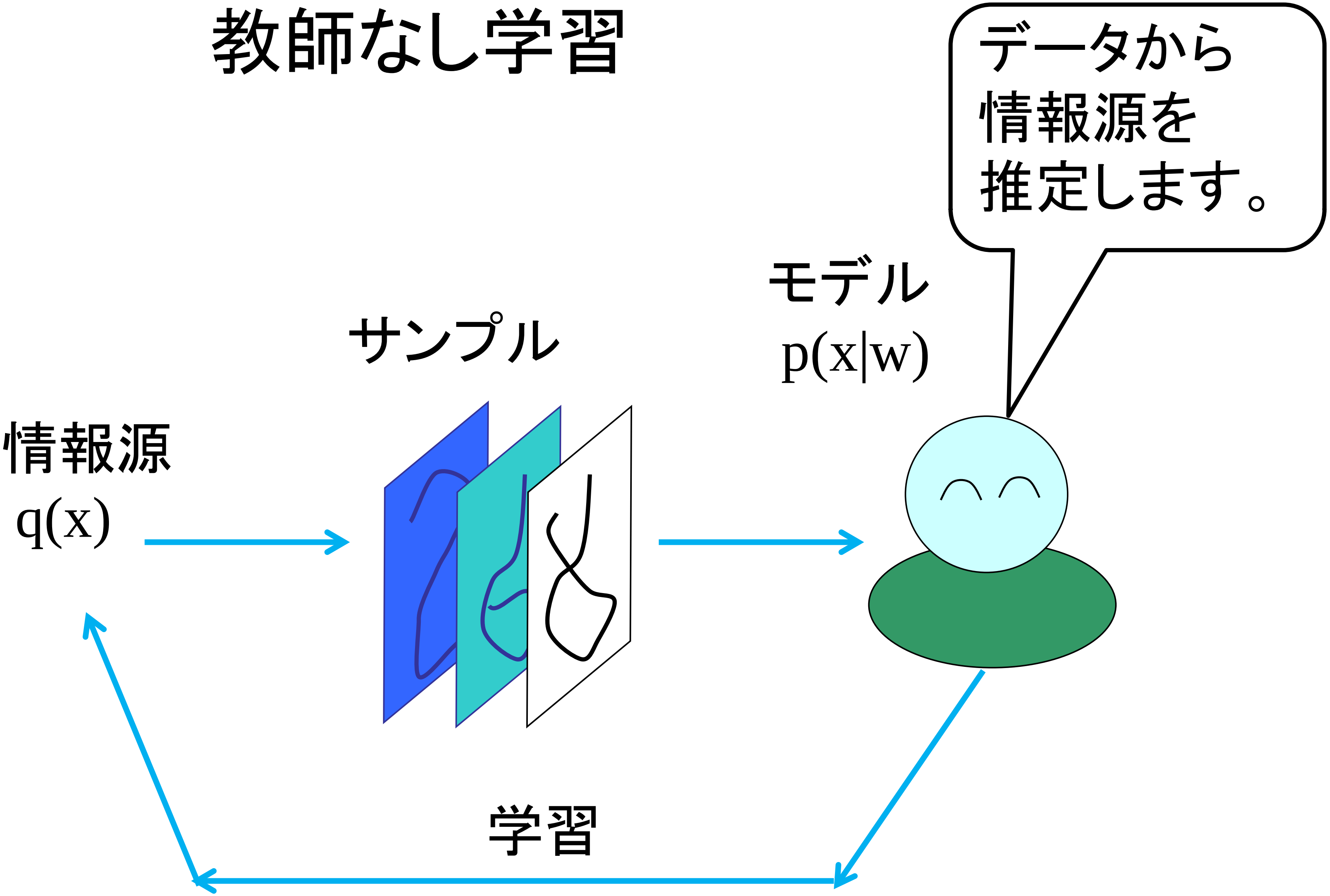
同時確率 $p(x,y)$
周辺確率$p(x)$
条件つき確率 $p(y|x)$

を定義した

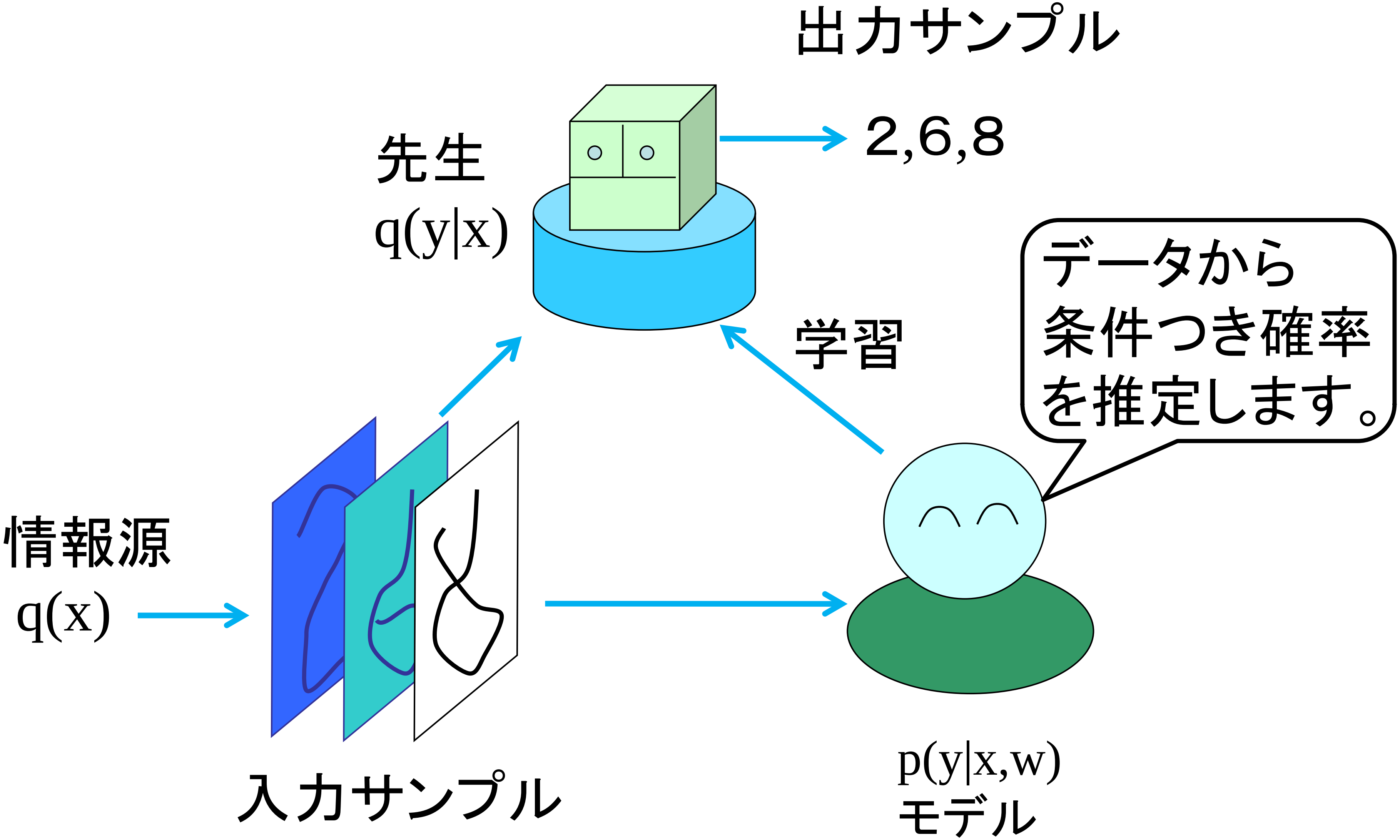
カルバック・ライブラ情報量を定義した

教師なし学習と教師あり学習

教師なし学習



教師あり学習


出力サンプル
2,6,8
先生
q(y|x)
学習
データから
条件つき確率
を推定します。
情報源
q(x)
入力サンプル
p(y|x,w)
モデル

教師あり学習と教師なし学習は数学的に等価

教師なし学習は
「 $q(x)$ からのデータを用いて $q(x)$ を $p(x|w)$ で推定」

教師あり学習は
「 $q(x)q(y|x)$ からのデータを用いて $q(y|x)$ を $p(y|x,w)$で推定」

数理的には、教師あり学習は教師なし学習の特殊な場合であり
また、教師なし学習は教師あり学習の特殊な場合です。

理論を作るとき記述は教師なし学習のほうがシンプルになるので、
教師なし学習の数式を考えますが、教師あり学習にそのまま適
用できます。

ベイズ推測の定義

統計的推測と統計的学習

$\mathbf{R}^N$に値をとる確率変数 $X^n = (X_1, X_2, \ldots, X_n)$ の実現値（データ）が得られたとき、このデータを発生している確率密度関数 $q(x)$ （真の分布）を知りたい。

しかし確率分布の集合は一般に無限次元であり、データは有限かつランダムであるから $q(x)$ を完全に知ることはできない。

ここでは、パラメータ w をもつ**統計モデル** $p(x|w)$ とパラメータ w の**事前密度関数** $\varphi(w)$ を使って推測を行なう方法を考える。

（注意）あるパラメータ w_0 が存在して $q(x)=p(x|w_0)$ となるとき、$q(x)$は $p(x|w)$ で実現可能という。一般には $q(x)$ は $p(x|w)$ で実現可能ではない。

事後分布の定義

もしも 「パラメータ w が事前密度 φ(w) から発生し、データ X^n がモデル p(x|w) から独立に発生した」ならば (w,X^n) の同時密度関数は

$$\varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

である。データ X^n が与えられたときのパラメータの条件つき密度関数は

$$p(w|X^n) = \quad (1/Z)\, \varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

である。これを**事後確率密度関数**という。ここで正規化定数

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

を X^n の**周辺密度関数**という。

(注意)モデルは真実ではないから、事後確率は人間が定義したものにすぎない。

ベイズ推測の定義

人間が準備した φ(w) と p(x|w) を用いて真の密度関数 q(x) を推測する密度関数 p*(x) を次式で定義する。

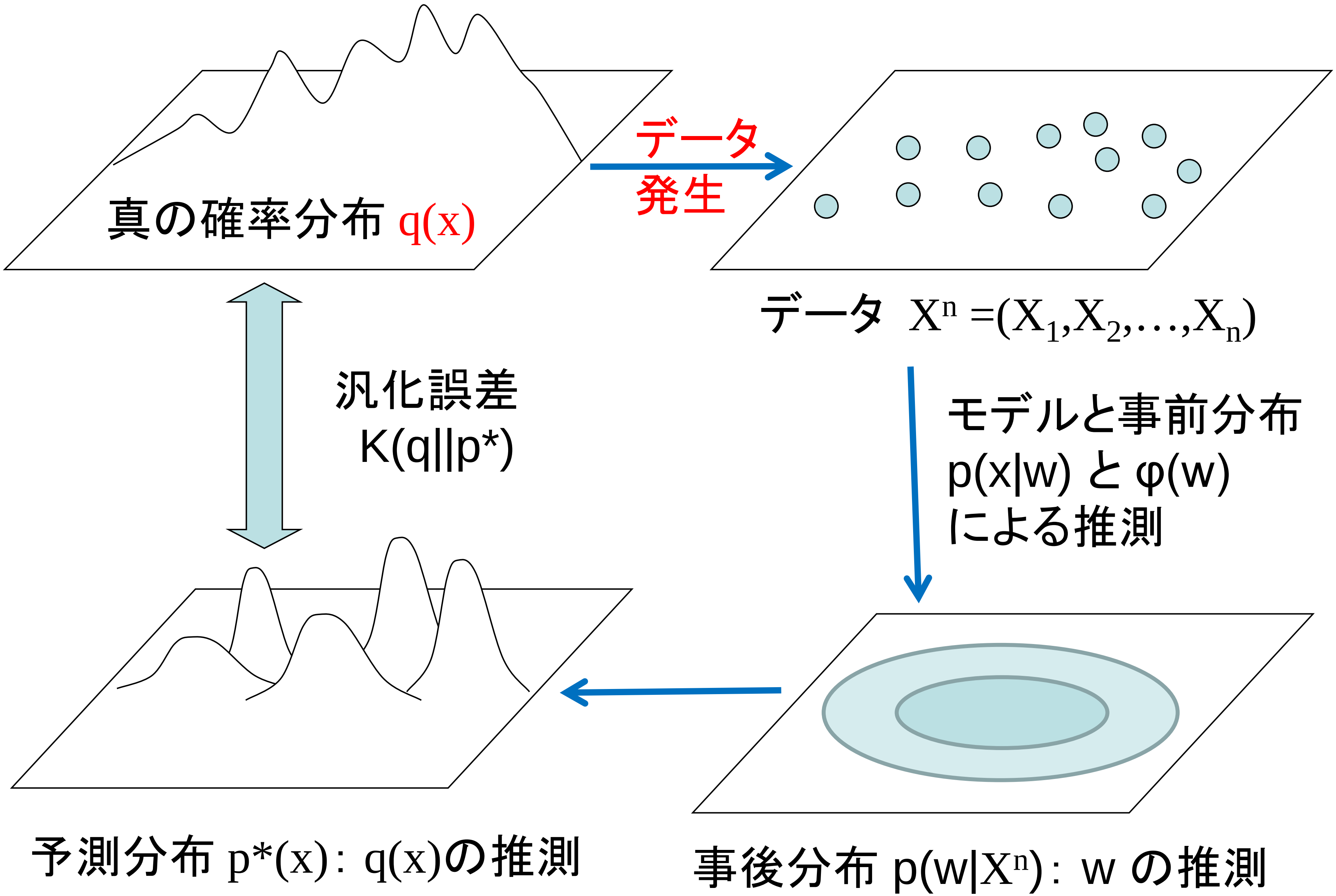
<u>**ベイズ推測**</u>. 統計モデルを事後分布で平均して推測結果とする.

$$p^*(x) = \int p(x|w)\, p(w|X^n)\, dw.$$

推測された密度関数 p*(x) を**予測分布**と呼ぶ（あくまでも推測である）。

真の分布 q(x) は不明である。人間の推測 p*(x) が、不明な q(x) に対して、どのくらい正しいかを数学で調べていく。

ベイズ推測の手順



真の確率分布 q(x)
データ
発生
データ X^n =(X_1,X_2,…,X_n)
汎化誤差
K(q||p*)
モデルと事前分布
p(x|w) と φ(w)
による推測
予測分布 p*(x): q(x)の推測
事後分布 p(w|X^n): w の推測

真の分布と推測された分布

事前密度関数 φ(w) も統計モデル p(x|w) も人間が仮に定めたものに過ぎないので、そこから定義される予測密度関数 p*(x) は、真の密度関数 q(x) と同じではない。**汎化損失**を次の式で定義する。

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

汎化損失は p*(x)=q(x) のときに限り最小値（真の分布のエントロピー）

$$S = -\int q(x) \log q(x)\, dx$$

をとる。カルバック・ライブラ情報量 K(q||p*) を**汎化誤差**という。

$$K(q\|p^*) = G(p^*) - S.$$

汎化誤差が小さいほど、予測分布は真の分布に近いと考えられる。

推測の目標

次の二つの問題を考えたい。

１　順問題（学習理論）

真 $q(x)$、モデル $p(x|w)$、事前 $\varphi(w)$ が与えられたとき、確率変数である汎化誤差の挙動を解明する。

２　逆問題（実データの分析）

真が不明であるとき、汎化誤差の値を推定したい。

汎化誤差の値が小さくなるようにモデルと事前分布を最適化し、真の分布をできる限り正確に知りたい。

順問題の実験の手順

(1) 真 $q(x)$、モデル $p(x|w)$、事前 $\varphi(w)$ を決める。

(2) 真から n 個のデータ X^n を発生する（独立に）。

(3) 事後分布 $p(w|X^n)$ から $\{w_k;k=1,2,\ldots,K\}$ を発生。

(4) 予測分布 $p(x|X^n)$ を $(1/K)\ \Sigma_{k=1}^{K} p(x|w_k)$ とする。

(5) 学習用のデータとは独立にテスト用データ Y^T を発生して汎化誤差 $(1/T)\ \Sigma_{t=1}^{T} \log\{q(Y_t)/p(Y_t|X^n)\}$ を求める。

逆問題の実験の手順

(1) n 個のデータ X^n を得る。

(2) モデル $p(x|w)$、事前 $\varphi(w)$ を決める。

(3) 事後分布 $p(w|X^n)$ から $\{w_k;k=1,2,\ldots,K\}$ を発生。

(4) 予測分布 $p(x|X^n)$ を $(1/K)\ \Sigma_{k=1}^{K} p(x|w_k)$ とする。

(5) 予測分布が、真に対してどの程度に適切かを
調べ、予測分布から真の分布を推測する。

計算科学として難しいポイント

(3) 事後分布 $p(w|X^n)$ から $\{w_k;k=1,2,\ldots,K\}$ を発生。

○　事後分布は式では書けるものの、事後分布に従うパラメータを取り出すことは一般的に容易ではない。

○　マルコフ連鎖モンテカルロ法が必要になるためその研究が多数行われている。

○　数年前にどんなモデルに対しても実行できるプログラム言語STANが発表され広く使われている。

練習問題

$X \in \{0,1\}$, $0<a<1$. （表が出る確率が a のコイン）

学習モデル $p(x|a) = a^x(1-a)^{1-x}$　事前分布 $\varphi(a) =1$

とする。$\Sigma X_i=n_1$ と書く（表の回数）。

事後分布と予測分布を n と n_1 で表せ。ただし

$\int_0^1 a^m(1-a)^n \, da =m!n!/(m+n+1)!$ を用いてもよい。

準備

確率論や統計学で必要になる初等的な知識として
いくつかの基本的な分布を

「*式で表すこと*」
「*その分布に従う乱数を発生すること*」

があります。

「式で表すこと」と「乱数を発生すること」は、場合に
よってはまったく別々の課題のときもあります。

準備1　正規分布を式で書く

一次元 正規分布 $N(a,\sigma^2)$（平均a 標準偏差 σ）

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

密度関数を描画
（matlab で書きますがプログラムは何でもOKです）

```
a=1;
s=1;
x = -5:0.1:5 ;
y=1/(2*pi*s^2)^(1/2)*exp(-1/(2*s^2)*(x-a).^2);
plot(x,y,'b-');
```

準備２　正規分布に従う乱数を発生する

正規分布 $N(0,1^2)$ に従う独立な n 個を横ベクトルで発生

x = randn(1,n) ;

正規分布 $N(a,\sigma^2)$ に従う独立な n 個を横ベクトルで発生

x = a+ σ*randn(1,n);

（注意）一様乱数を使って正規分布を発生する方法は
ありますが、今日では、正規乱数を発生するライブラリを
使うということでよろしいかと思います。

準備3　ガンマ分布を式で書く

ガンマ分布 (x>0, a>0, b>0)

$$p(x|a,b) = \{1/(b^a\Gamma(a))\}\, x^{a-1} \exp(-x / b)$$

（知っててほしい）ガンマ関数：$\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)\, dx$

密度関数を描画

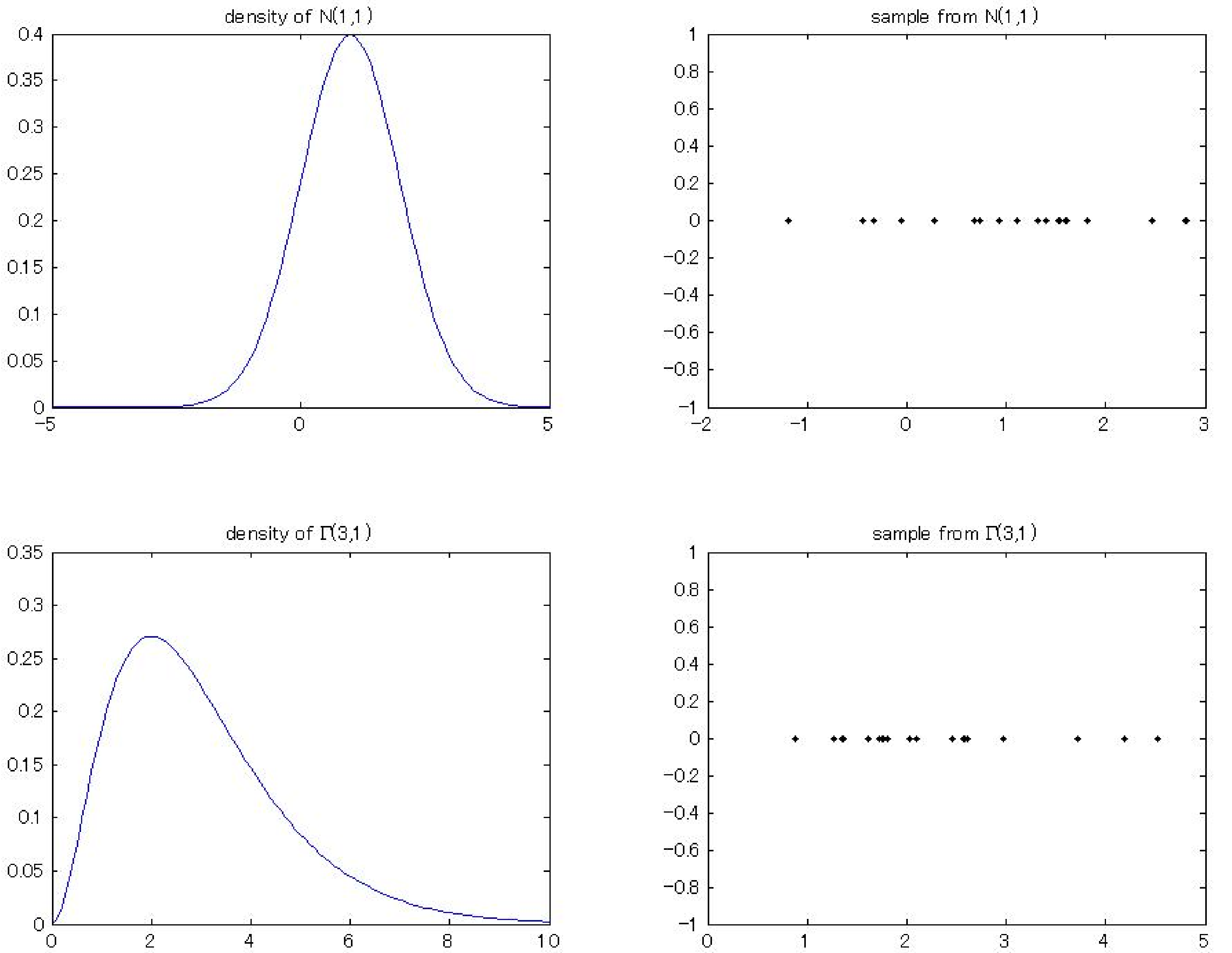
```
a=3;
b=1;
x = 0:0.1:10 ;
y= 1/(b^a*gamma(a))*x.^(a-1).*exp(-x/b);
plot(x,y,'bo-');
```

準備4　ガンマ分布に従う乱数を発生

ガンマ分布 に従う独立な n 個を横ベクトルで発生

```
x = gamrnd(a,b,1,n) ;
```

描画の例



density of N(1,1)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−5
0
5
sample from N(1,1)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−2
−1
0
1
2
3
density of Γ(3,1)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
sample from Γ(3,1)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
1
2
3
4
5

具体例

正規分布による推定

学習モデル

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

事前分布

$$\varphi(a,s) \propto s^{1/2} \exp(- sa^2/2 - s)$$

推定は、真の分布が不明でもできる。

データと事後分布

データ　$\{ x_i ; i=1,2,\ldots,n\}$

事後分布　：　定義にモデルの式を代入すると

$$p(a,s|X^n) \propto \varphi(a,s) \Pi_{i=1}^{n} p(x_i|a,s)$$

補題　$A = (\Sigma_{i=1}^{n} x_i)/(n+1)$, $B=2+\Sigma_{i=1}^{n} x_i^2$ とおくと

$$p(s|X^n) \propto s^{n/2} \exp[- (s/2)\{B-(n+1) A^2 \}]$$

$$p(a|s,X^n) \propto \exp[- (s(n+1)/2)(a - A)^2]$$

これより事後分布は次の手続きで発生できる。

$s \sim \Gamma(n/2+1,2/(B-(n+1) A^2))$,　$a \sim N(A,1/(s(n+1)))$

補題の証明

事後分布の定義にモデルの式を代入すると

$$p(a,s) \propto \varphi(a,s) \Pi_{i=1}^{n} p(x_i|a,s)$$

$$\propto s^{1/2} \exp(- sa^2/2 - s)$$

$$\times \Pi_{i=1}^{n} \; s^{1/2} \exp[-s(x_i-a)^2/2]$$

$$\propto s^{(n+1)/2} \exp(- sa^2/2 - s)$$

$$\times \; \exp[- s \, \Sigma_{i=1}^{n} (x_i - a)^2/2].$$

$A = (\Sigma_{i=1}^{n} x_i)/(n+1)$,　　$B=2+\Sigma_{i=1}^{n} x_i^2$ とおくと

補題の証明（続き）

$$p(a,s) \propto s^{(n+1)/2} \exp[- (s/2)\{ (n+1) (a^2 – 2Aa) +B\}].$$

a について平方完成すると

$$p(a,s) \propto s^{(n+1)/2} \exp[-(s/2)\{(n+1)(a-A)^2-(n+1) A^2 +B\}].$$

a について積分すると s の周辺分布が得られるので

$$p(s) \propto s^{n/2} \exp[- (s/2)\{B-(n+1) A^2 \}].$$

また $p(a|s) = p(a,s) / p(s)$

$$\propto \exp[- (s/2) (n+1)(a – A)^2].$$

（証明終）

実験例

真

$$q(x) = (1/2\pi)^{1/2} \exp[-x^2/2]$$

学習モデル

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

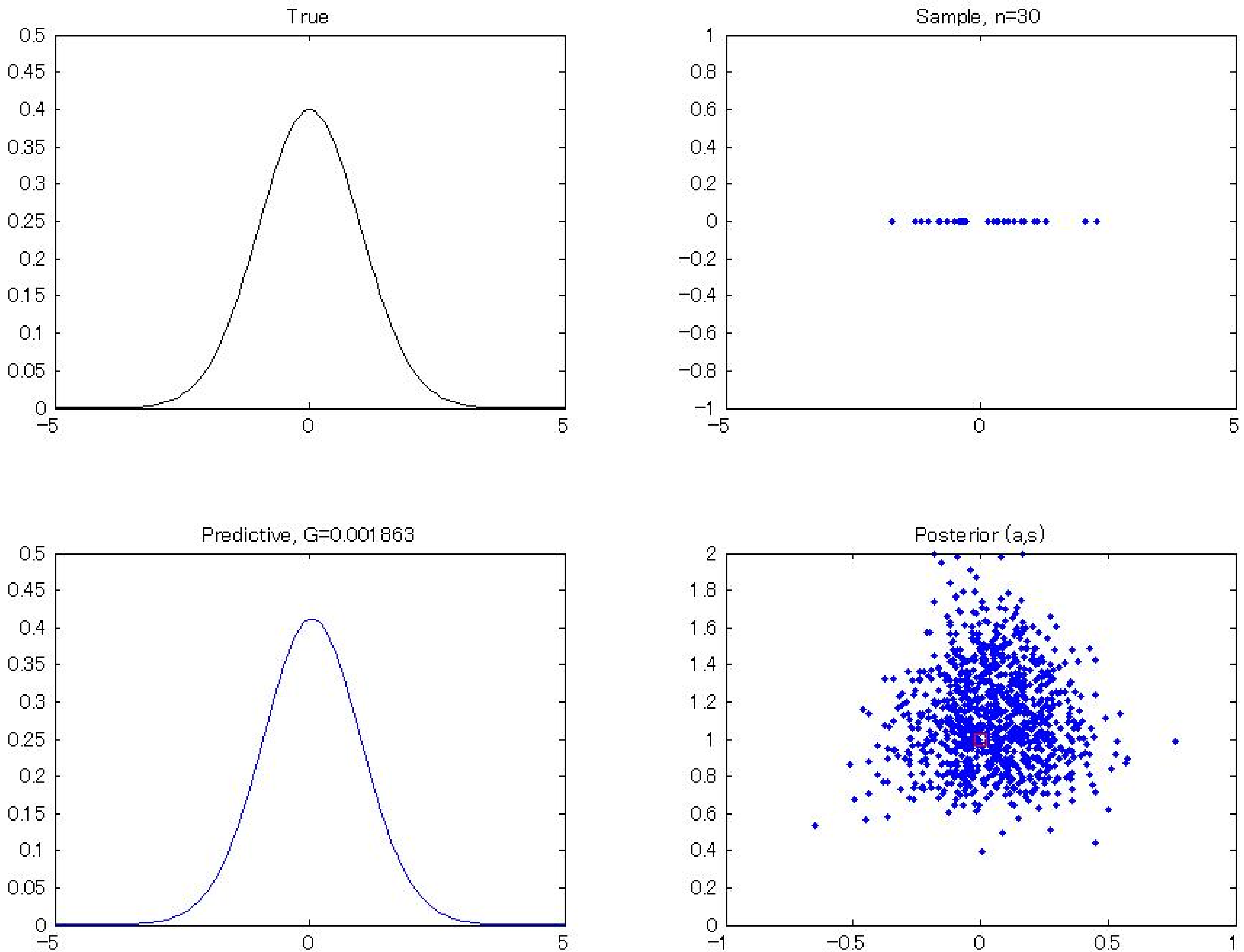
事前分布

$$\varphi(a,s) \propto s^{1/2} \exp(-sa^2/2 - s)$$

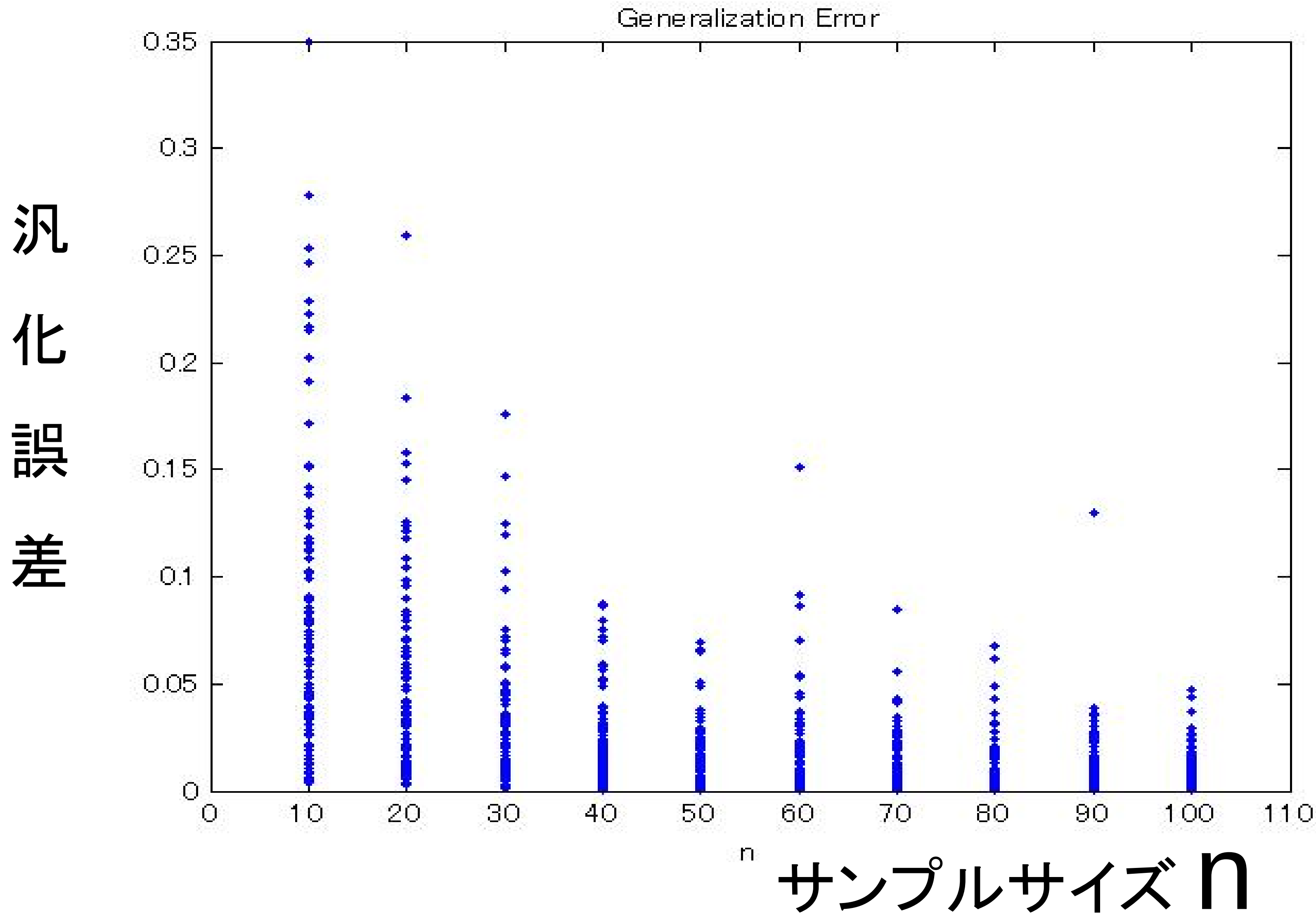
ここでは演習のため真の分布を正規分布として

いますが、実問題では真の分布は不明です。

実験例



実験例　学習曲線


Generalization Error
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
90
100
110
n
汎
化
誤
差
サンプルサイズ n

練習問題

真の分布が正規分布では実現できないもの
であるとき、学習曲線はどのようになると考えるか
図示せよ。