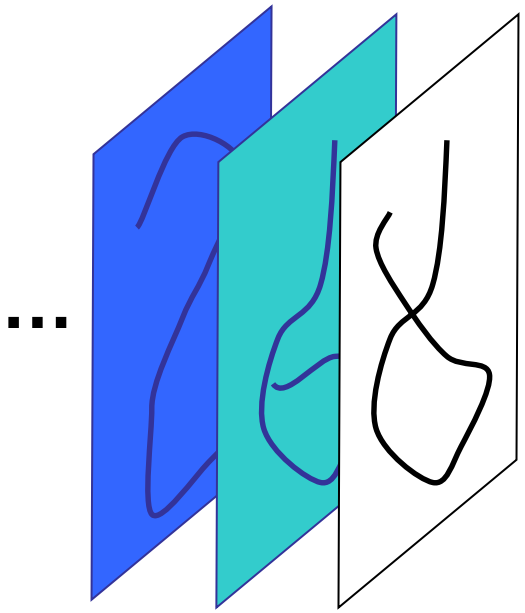
学習理論の練習　４

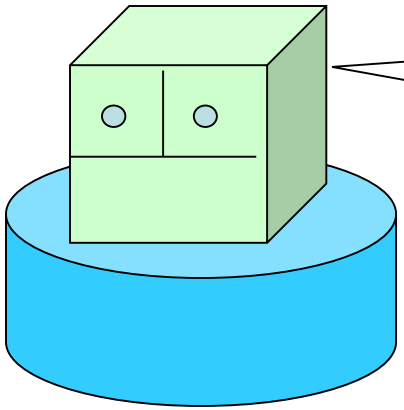
１　漫画でわかる学習理論

教師あり学習

例

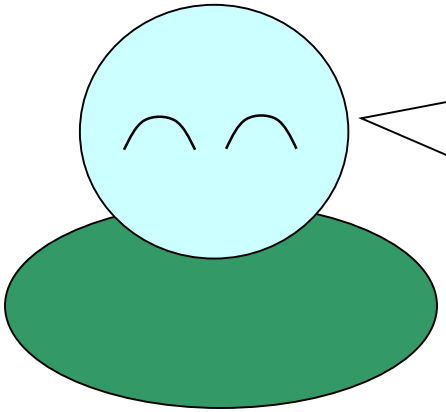


…



先生

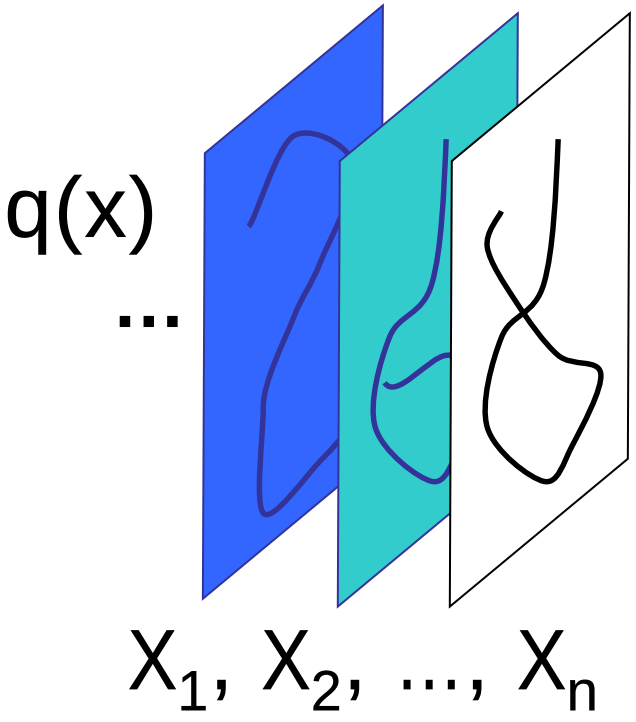
8, 6, 2…

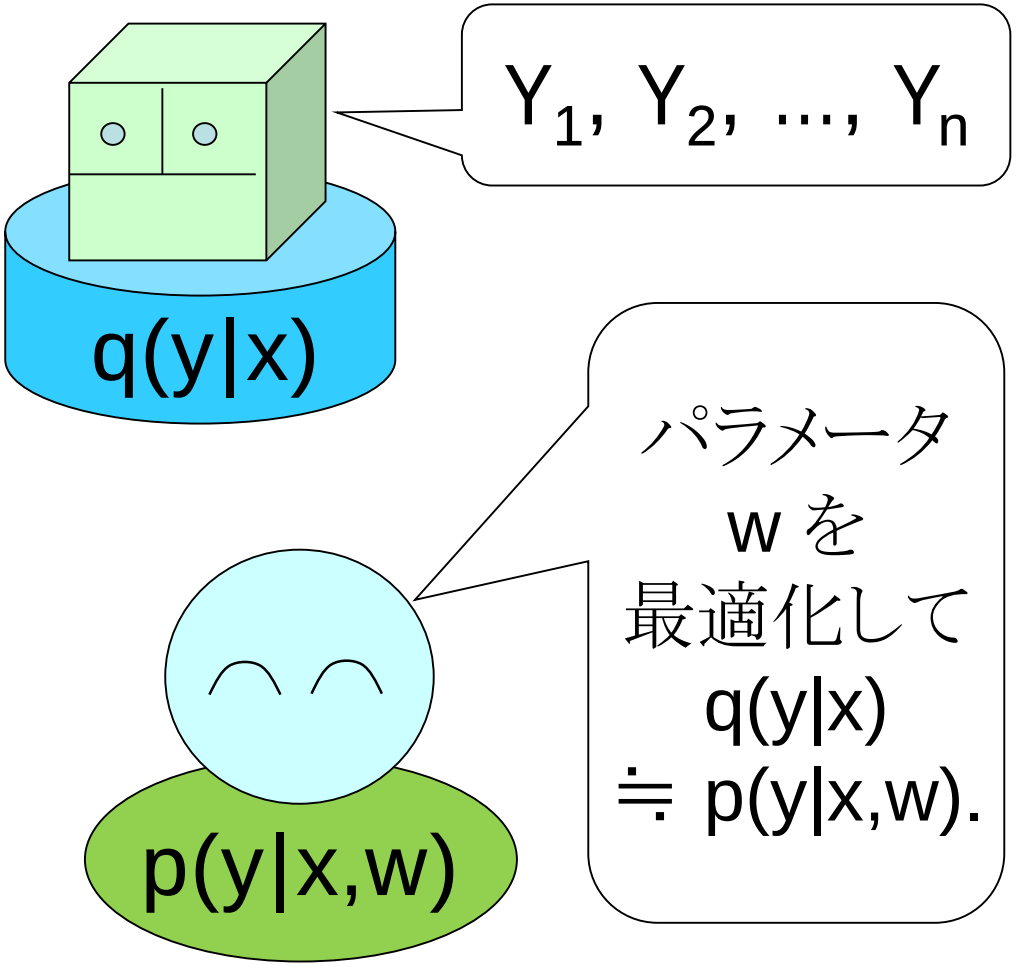


生徒

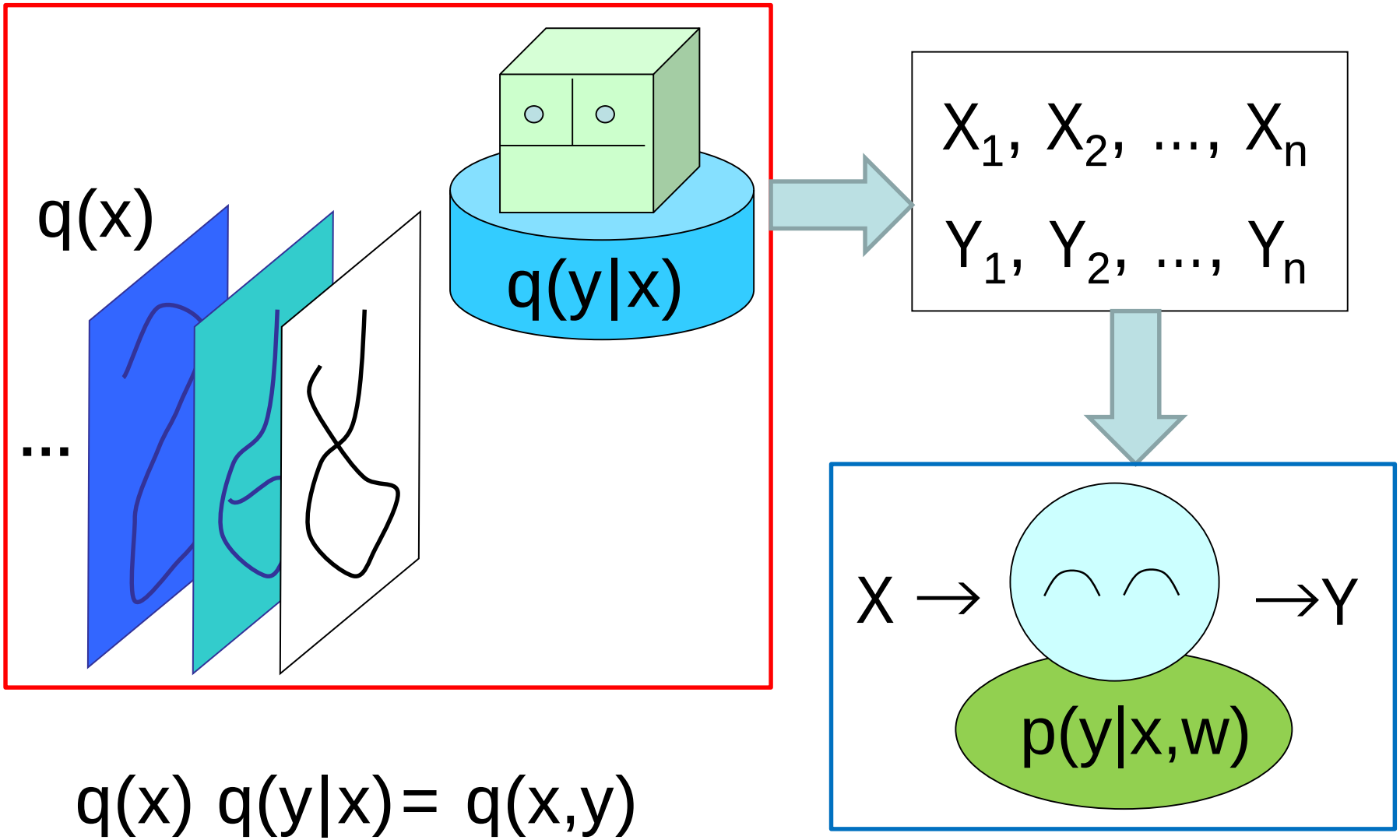
文字を
読みます

数学的に記述すると


q(x)
…
$X_1, X_2, \ldots, X_n$

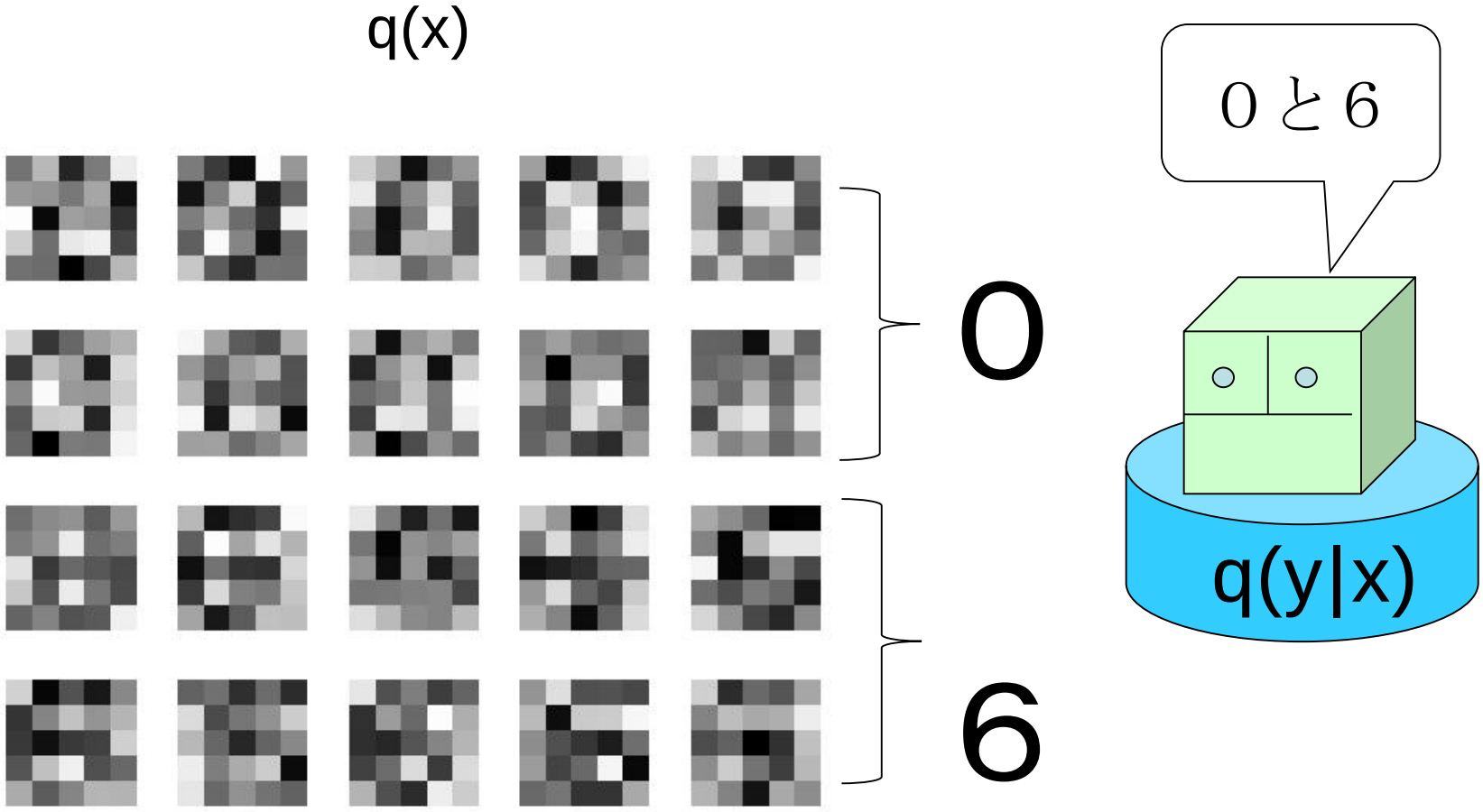

$Y_1, Y_2, \ldots, Y_n$
q(y|x)
パラメータ
w を
最適化して
q(y|x)
≒ p(y|x,w).
p(y|x,w)

情報源と学習モデル



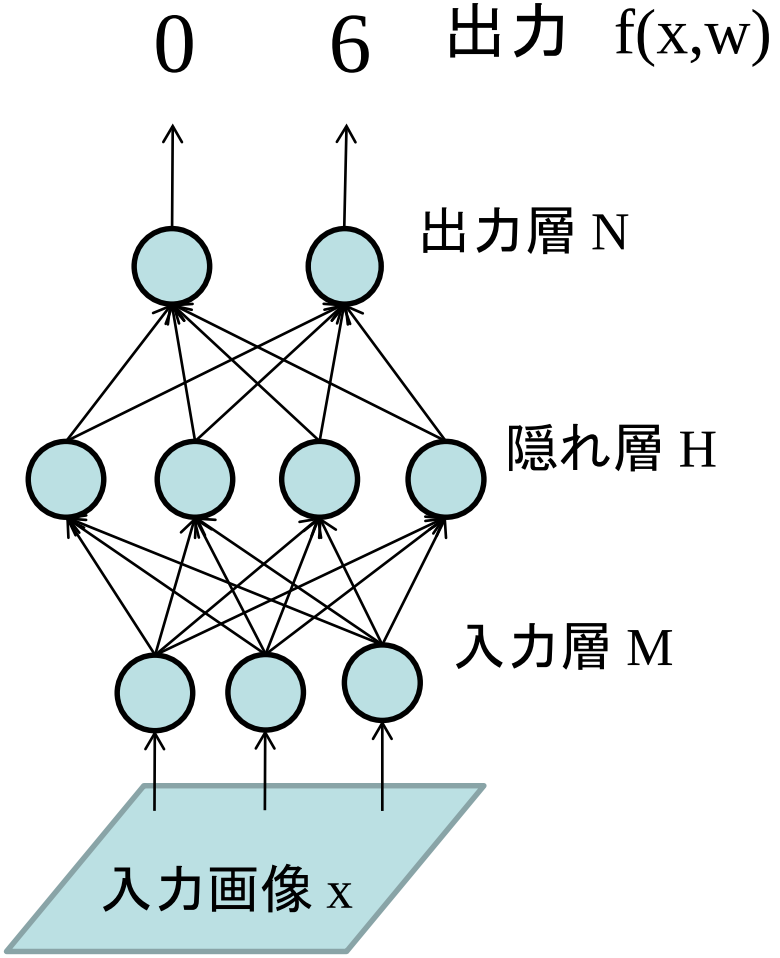
$q(x)\ q(y|x) = q(x,y)$

情報源 q(x)q(y|x) の例

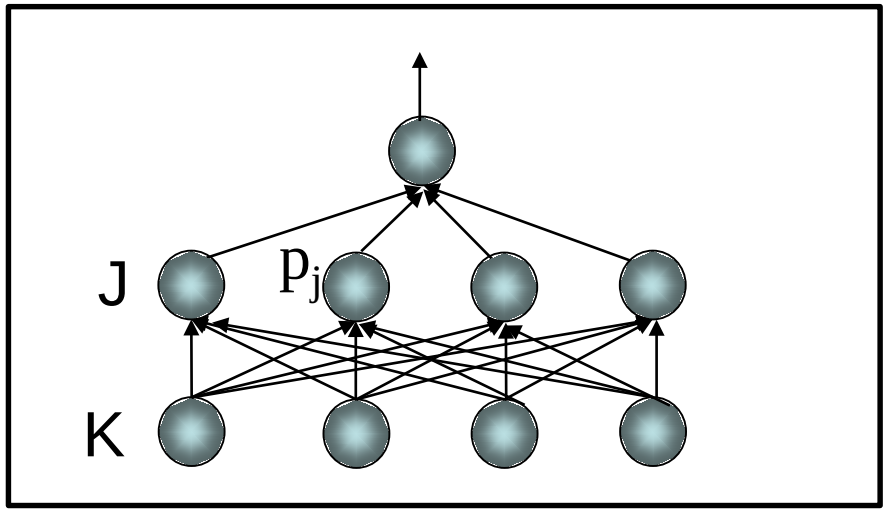


学習モデルの例

=



ここ数年の話題1：敵対的データ生成

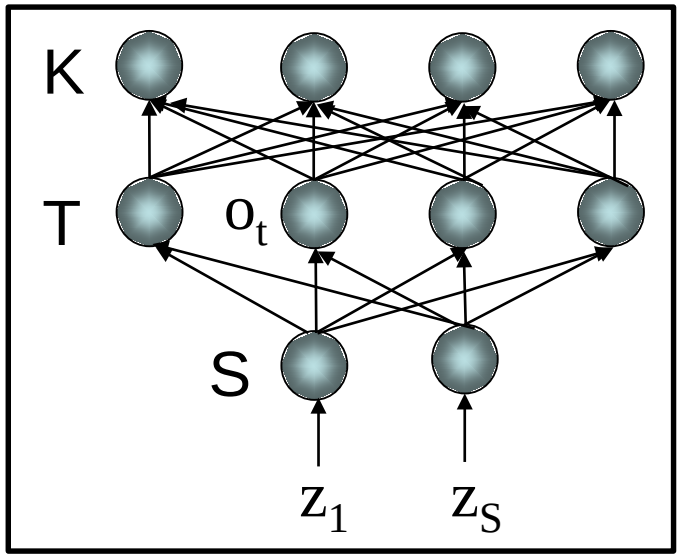


データと偽データを
識別するネットの
学習1

学習1と
学習2を
交互に
繰りかえす

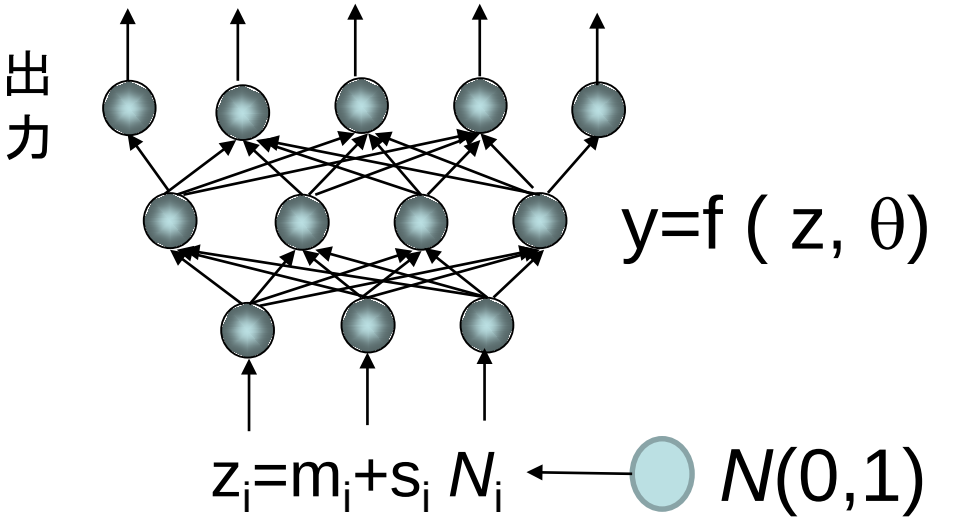
学習データ

$x=\{x_k\}$



偽データを
生成するネットの
学習2

ここ数年の話題2：変分オートエンコーダー

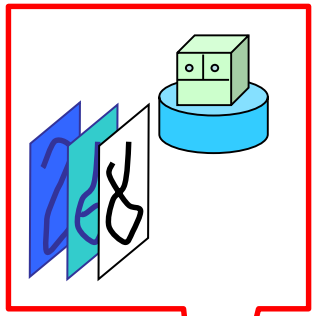


オートエンコーダーの中間部分に正規雑音を入れてその平均と分散を学習すると

データがあるところだけでなくその周りを確率的に補完することができる

２　漫画から数式へ

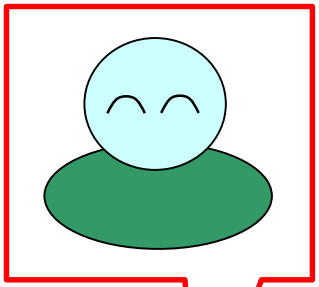
教師あり学習の枠組み



学習用データ

$X_1, X_2, \ldots, X_n$

$Y_1, Y_2, \ldots, Y_n$



q(x,y)

p(y|x,w)

テストデータ

X

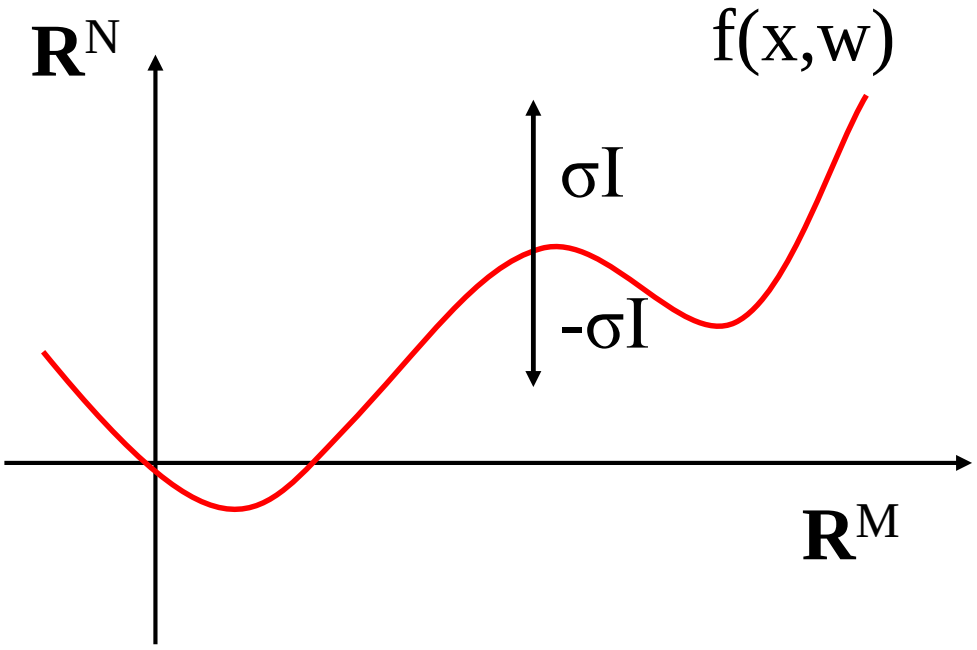
Y

未知の情報源

学習モデル

学習モデルの例1

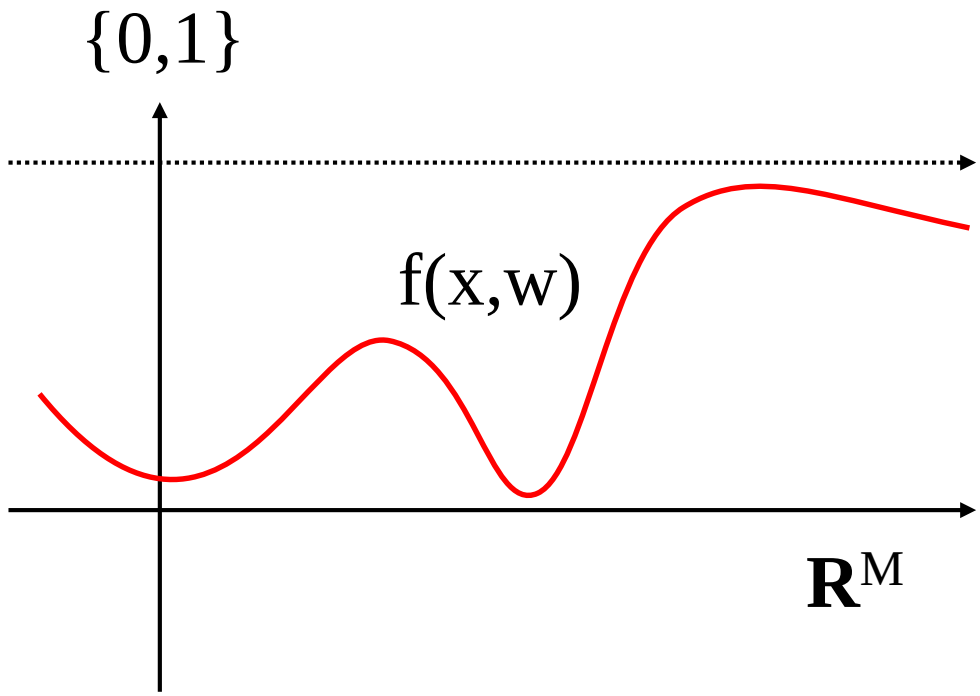
$$p(y|x,w,\sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \| y - f(x,w) \|^2 \right\}$$



- $p(y|x,w,\sigma)$ はX が与えられたときのYの条件つき確率密度
- Y は実数全体をとりうる
- Y は平均 f(x,w) 分散$\sigma^2 I$ の正規分布
- w と σ がパラメータ

学習モデルの例2

$$p(y|x,w) = f(x,w)^{y} \left(1-f(x,w)\right)^{1-y}$$



- $p(y|x,w)$ は
  X が与えられたとき
  のYの条件つき確率
- Y は {0,1} をとる
- Y は 確率 f(x,w) で1
  確率 (1-f(x,w))で0
- w がパラメータ

条件つき確率の尤度関数

データ $\{(X_i,Y_i)\ ;\ i=1,2,\ldots,n\}$, モデル $p(y|x,w)$ が与えられたとき

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

を(条件つき確率の）尤度関数という。

- ○ もしも「モデル＝真」であれば、$L(w)$ は $\{ X_i \}$ が与えられたときの $\{ Y_i \}$ の確率密度関数
- ○ 尤度関数を最大にする w を最尤推定量という。
- ○ 最尤推定量は良い推定量ではないがよく利用される。深層学習では悪すぎて使えない。学習理論の問題は一般には最適化問題には帰着しない。

対数尤度関数

尤度関数の対数をとって

$$\log L(w) = \sum_{i=1}^{n} \log p(Y_i|X_i,w)$$

を(条件つき確率の）対数尤度関数という。

- ○ log は単調増大関数だから対数尤度関数を
  最大にするパラメータが最尤推定量

- ○ $H(w) = -(1/n) \log L(w)$ と書けば $L(w)= \exp(-nH(w))$

- ○ 尤度関数と対数尤度関数はどちらを考えても同じだが
  計算上対数尤度関数のほうが便利なことが多い。

対数尤度関数の例1

学習モデル

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

の対数尤度関数は

$$\log L(w) = - (1/2\sigma^2) \sum_{i=1}^{n} \| Y_i - f(X_i,w) \|^2 - Nn \log \sigma - (N/2) \log(2\pi)$$

最尤推定量(w*,s*) は

$$w^* = \mathrm{argmin}_w \sum_{i=1}^{n} \| Y_i - f(X_i,w) \|^2$$

$$s^* = \mathrm{sqrt} \left\{ (1/nN) \sum_{i=1}^{n} \| Y_i - f(X_i,w^*) \|^2 \right\}$$

対数尤度関数の例2

学習モデル　$p(y|x,w) = f(x,w)^{y} (1-f(x,w))^{1-y}$

の対数尤度関数は

$$\log L(w) = \sum_{i=1}^{n} \{ Y_i \log f(X_i,w) + (1-Y_i) \log (1-f(X_i,w)) \}$$

最尤推定量 w^* はこれを最大にするパラメータ

- ○ Xを0と1に識別する問題ではどちらの学習モデルを用いることもできる。同じ神経回路網を使っていても学習モデルとしては異なる。
- ○ Xを $\{(0,0),(1,0),(0,1),(1,1)\}$ に識別したい場合は多次元に拡張すればよい。個々に問題を解いてもよい。

３　数式から数理へ

対数尤度関数の符号反転

対数尤度関数を(1/n)倍して符号反転したものを H(w) と書く

$$H(w) = -(1/n) \sum_{i=1}^{n} \log p(Y_i|X_i,w)$$

データを発生している確率密度が q(x) q(y|x) であるとする。
データの数 n が大きいとき 大数の法則から

$$H(w) \fallingdotseq -\iint q(x)\, q(y|x) \log p(y|x,w)\, dxdy$$

条件つきエントロピーを下記で定義する

$$S = -\iint q(x)\, q(y|x) \log q(y|x)\, dxdy$$

尤度とKL情報量

$$H(w) \fallingdotseq -\iint q(x)\, q(y|x) \log q(y|x)\, dxdy$$

$$+ \iint q(x)\, q(y|x) \log \{ q(y|x) / p(y|x,w) \}\, dxdy$$

$$= S + D(q(y|x) \,\|\, p(y|x,w))$$

ここで $D(q(y|x)||p(y|x,w))$ は条件つき確率のKL情報量であり、常に非負で、零になるのは $q(y|x)=p(y|x,w)$ のときだけである。

データの数 n が十分大きければ

尤度が大きい＝対数尤度が大きい＝H(w)が小さい
≒ 真とモデルの条件つき確率のKL情報量が小さい

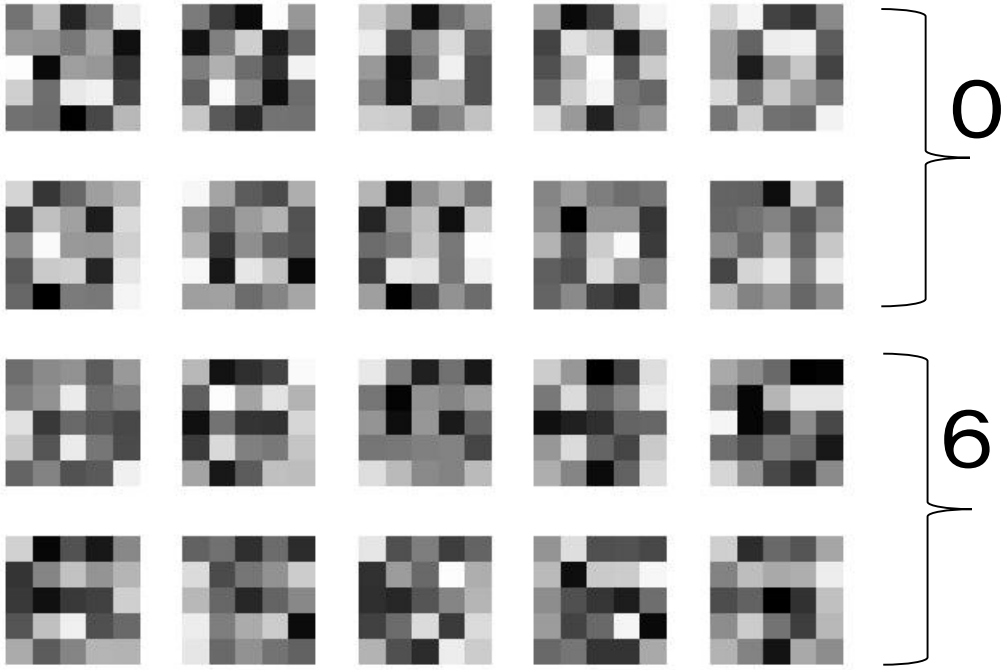
学習理論は何をしたいのか

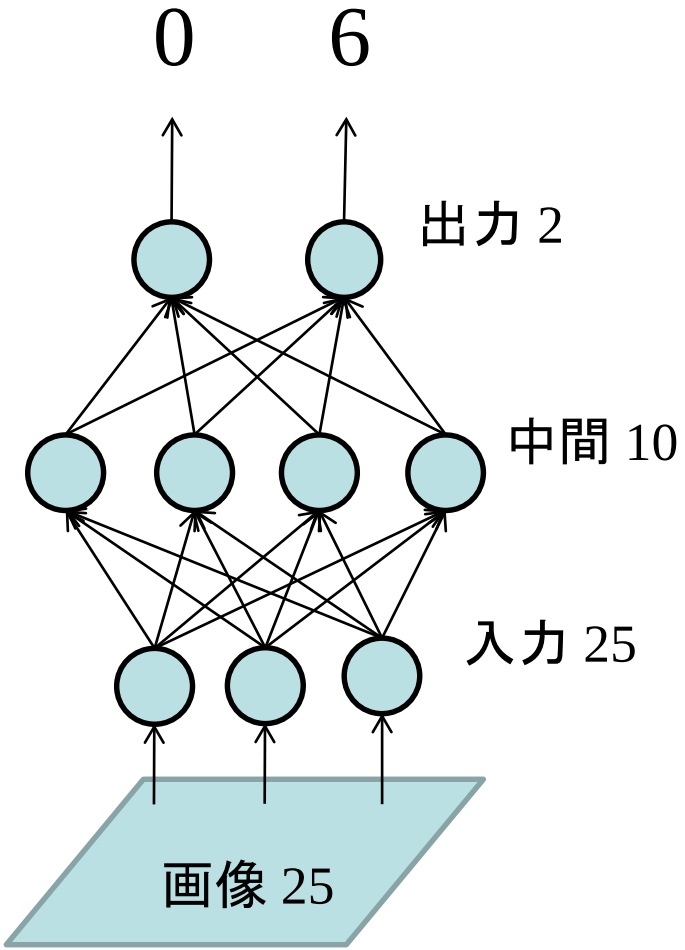
- 真の分布がわからなくても、尤度関数はデータとモデルだけで記述できる。
- 尤度関数を大きくすると「真とモデルのKL情報量」がほぼ小さくなる。つまり未知の真がほぼ推定できる。
- 「尤度関数最大⇔真とモデルのKL情報量が小さい」が等価であるためには $n=\infty$ が必要。
- n が有限なら「対数尤度関数≒真とモデルのKL情報量」のずれは、関数が複雑であるほど大きくなる。深層学習では、そのずれはベイズ法以外では解明されていない。
- 対数尤度関数とKL情報量のずれがわかると、データに対して最適なモデルやハイパーパラメータを決めることや統計的検定を作ることができるようになる。

４　実験してみる

画像認識の例

学習用データ, n=20, 40, …, 2000
テスト用データ, m=2000.
Lasso 学習





学習誤差と汎化誤差

各データについて500回Lasso 学習を繰り返してして止める。
パラメータ w* が得られる。

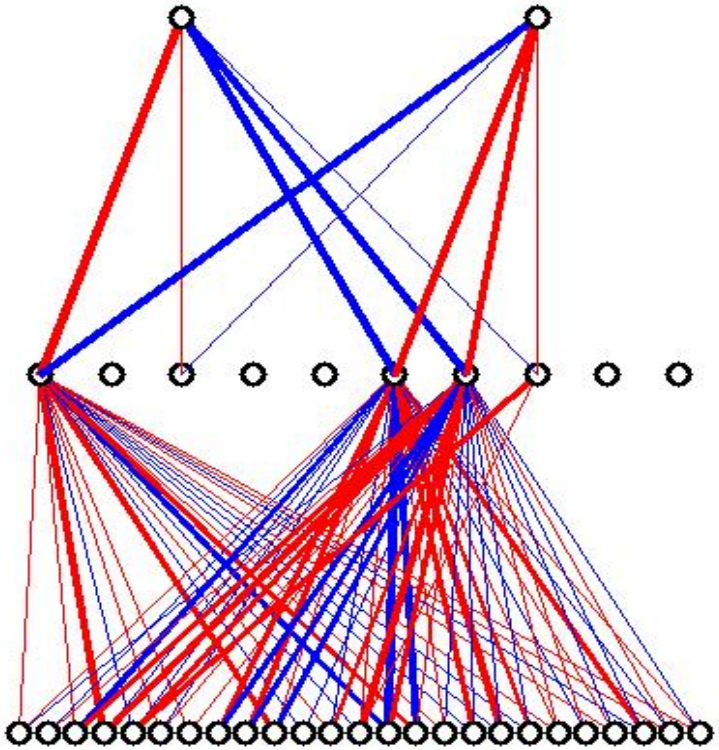
学習誤差関数　$E(w) = (1/n) \sum_{i=1}^{n} ||Y_i - f(X_i, w)||^2$

汎化誤差関数　$G(w) = (1/m) \sum_{i=1}^{m} ||Y_i - f(X_i, w)||^2$

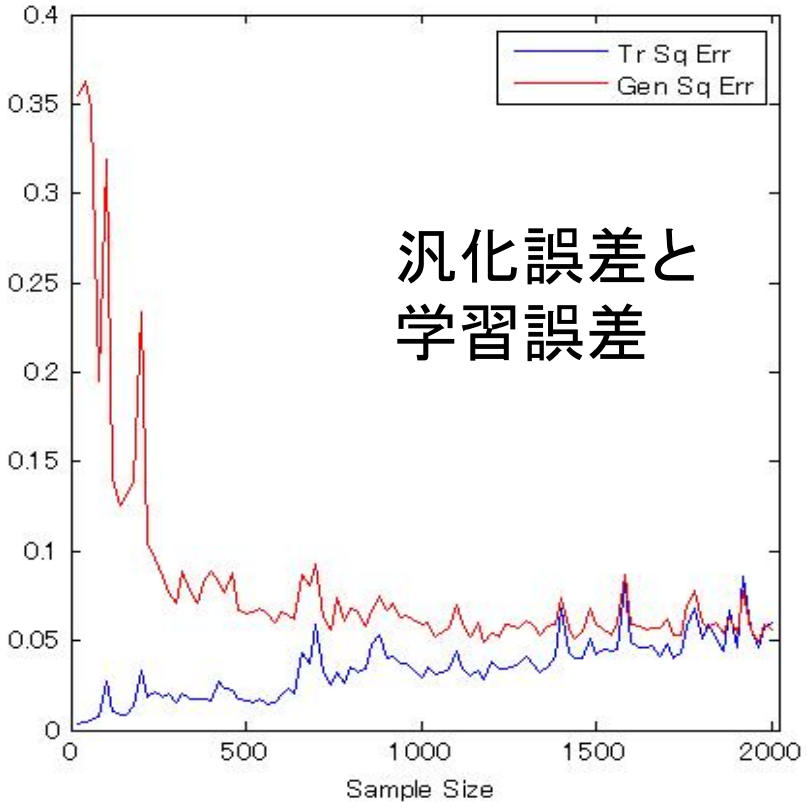
学習誤差 E(w*), 汎化誤差 G(w*) をプロットする。

☆　繰り返し数500は十分多いとはいえない。
　　学習を途中で止めたことになる。

実験の例



Lasso 学習の結果(n=2000)



学習データの数(n)

学習誤差と汎化誤差

- 平均的には　学習誤差＜汎化誤差　であるが、個々の例ではそうなるとは限らない。データの数 n が大きくなると学習誤差と汎化誤差は近づいていく。

- 学習誤差も汎化誤差も Lasso ハイパーパラメータの影響が大きい。今回は人間力でハイパーパラメータを決めている。

- まったく同じデータを使ってもパラメータの初期値に依存して学習結果は異なる。