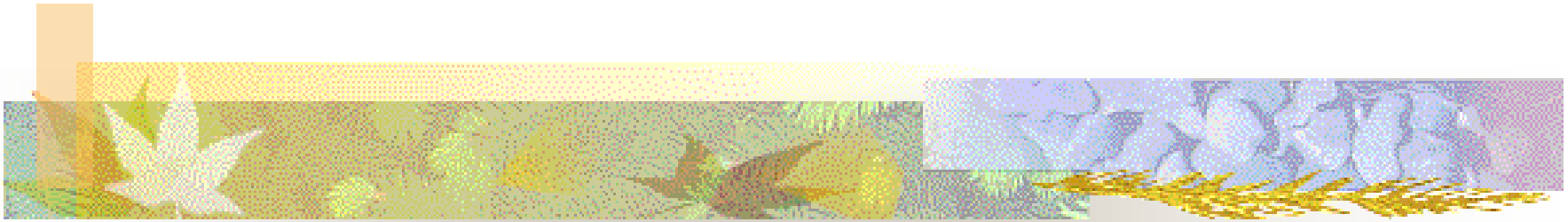
情報学習理論



渡辺澄夫

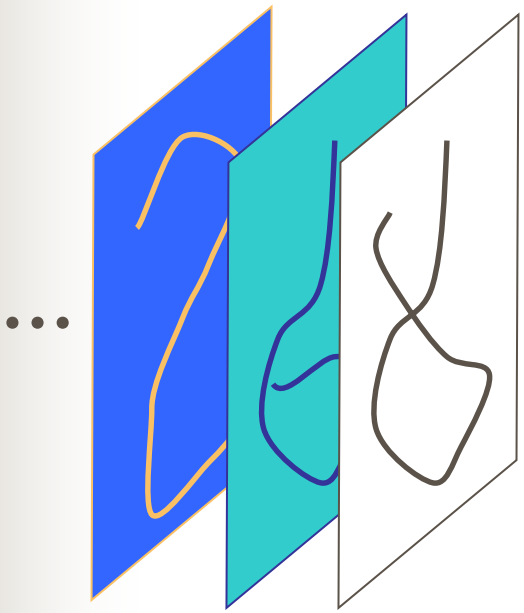
いろいろな学習

(1) 教師あり学習
(2) 教師なし学習
(3) そのほか

まず教師あり学習について学びます。
教師あり学習は実世界で重要なので
多くの方法があります。

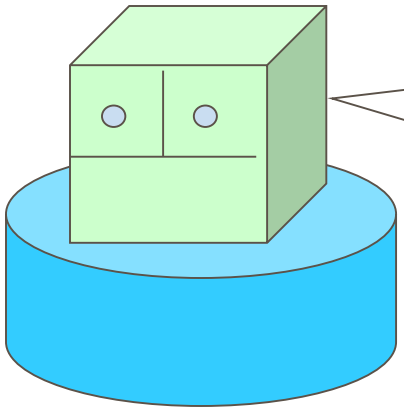
（線形回帰）　今日
（神経回路網）　来週以降
（SVM）　来週以降

教師あり学習とは



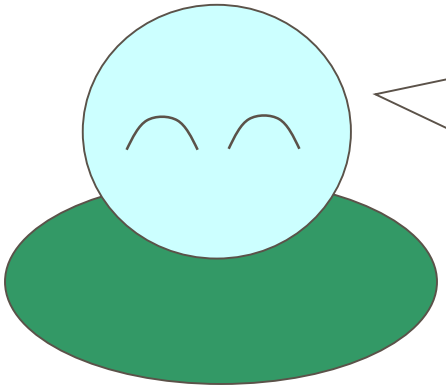
…

文字の例



先生

8, 6, 2…
でしょう

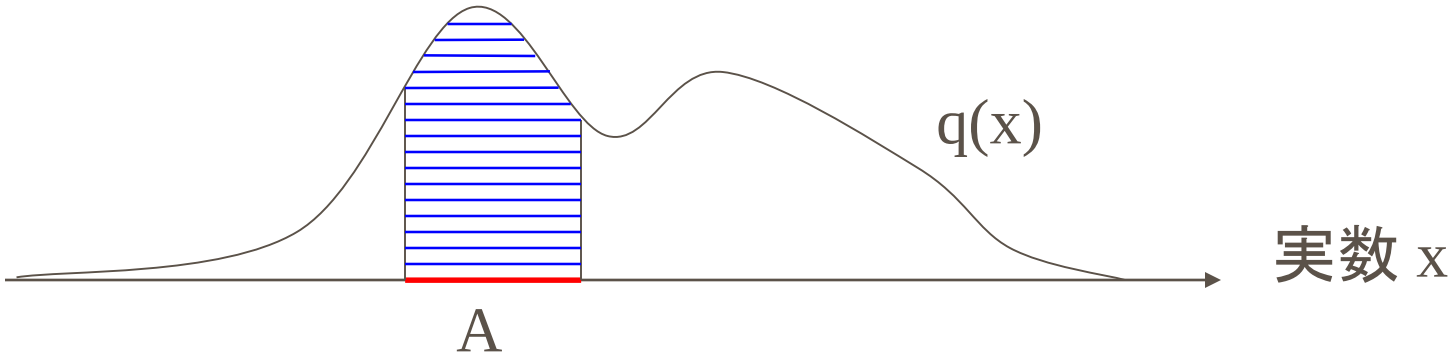


生徒

文字を読めるように学習します

復習：確率密度関数

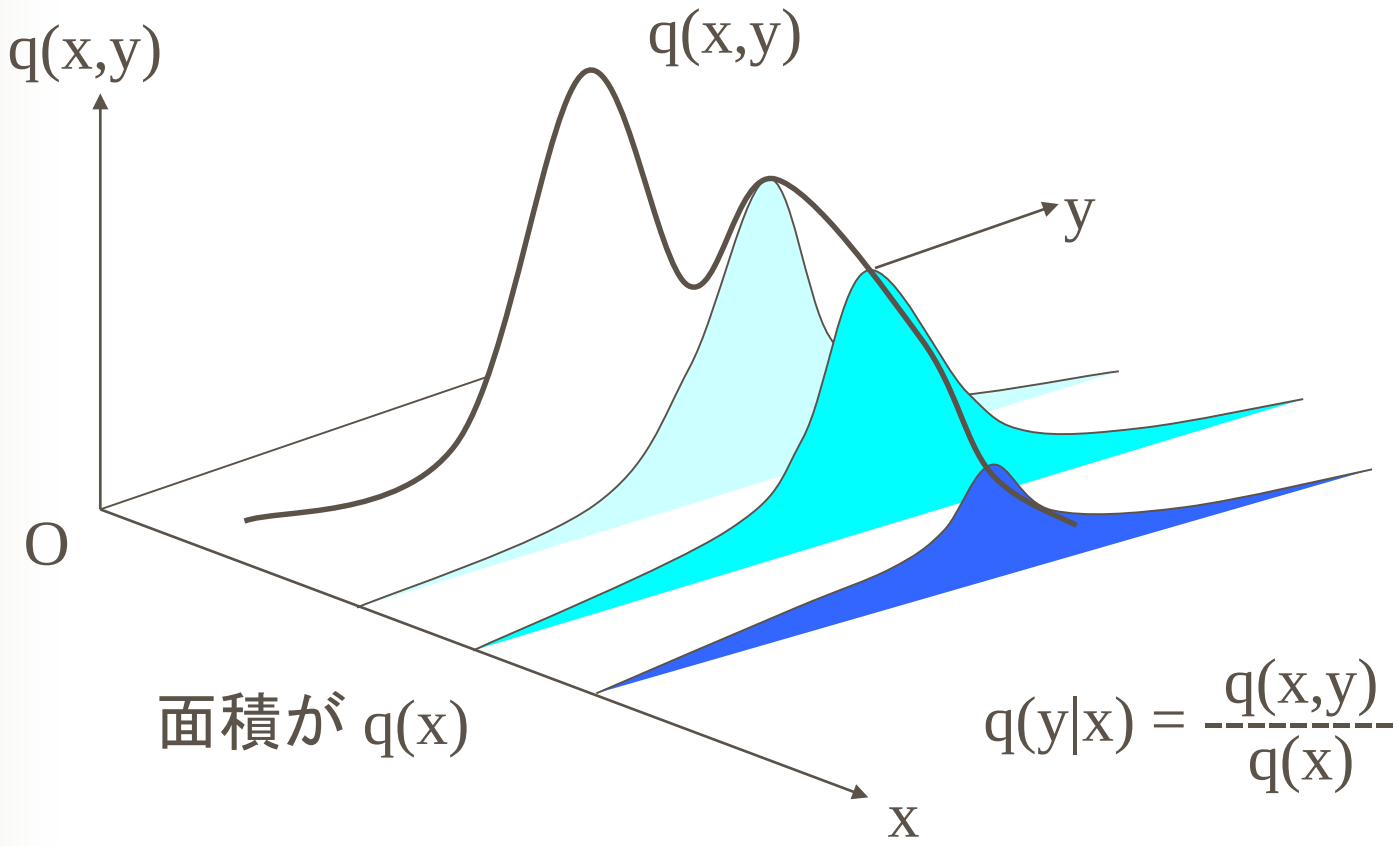
確率変数 X = 確率的にばらつく変数



確率変数 X が確率密度関数 q(x) を持つとき

$$「X \in A となる確率」 = \int_A q(x)\,dx$$

復習：条件つき確率密度関数



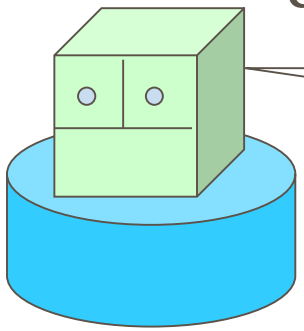
$$q(y|x) = \frac{q(x,y)}{q(x)}$$

q(y|x) は「x が与えられたときの y の確率密度関数」

教師あり学習は条件つき確率を推定する

真の推論

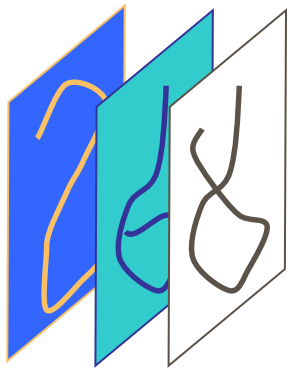
$q(y|x)$



答例 $y_1, y_2, y_3, \cdots$

真の分布

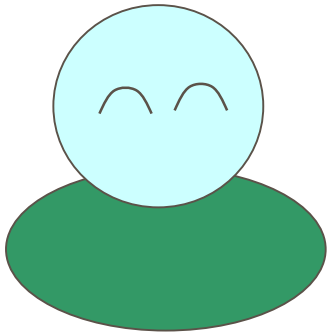
$q(x)$



入力例 $x_1, x_2, x_3, \cdots$

学習モデル

$p(y|x,w)$



学習結果 $p^*(y|x)$

教師あり学習の例

教師あり学習

画像認識、音声認識、時系列予測、
ロボット制御など応用は極めて多数。
「予測の正確さ」が必要になることが多い。

教師あり学習でないもの

データのクラスタリング
データの構造化
データの次元縮約など


牧場の5分後を予想するのが私の仕事です

教師あり学習の例： 回帰問題

入力　X

教師・真　　$Y = g(X)$ +雑音

生徒・モデル　$Y = f(X,w)$ +雑音

(1) 真の関数 g をモデル f で推測する。
(2) g(x) を真の回帰関数という。
(3) 例 { (Xi,Yi) ; i=1,2,…,n} から真を推測。
(4) モデルは真を知らない人間が仮に定めたもの。

線形回帰

データ番号 $i = 1,2,\ldots, n$

入力（d次元） $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$

真（1次元） $Y_i = w_0 \cdot X_i +$ 雑音

モデル（1次元） $Y_i = w \cdot X_i +$ 雑音

データ (X_i, Y_i) が得られたとき
$w_0 = (w_{01}, w_{02}, \ldots, w_{0d})$ を推測する

二乗誤差

データ　$\{(X_i, Y_i) ; i=1,2,\ldots,n\}$

二乗誤差を最小にする w を見つける

$$E(w)=\sum_{i=1}^{n} (Y_i - w \cdot X_i)^2$$

二乗誤差をパラメータで偏微分したときOになるようなパラメータを見つければよい。

$$E(w) = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{d} w_j X_{ij})^2$$

k=1,2,…,d について

$$\partial E/\partial w_k = 2 \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{d} w_j X_{ij})(-X_{ik}) = 0$$

$(w_1, w_2, \ldots, w_d)$ についての連立方程式が得られる。

$$\sum_{i=1}^{n} Y_i X_{ik} = \sum_{j=1}^{d} w_j \sum_{i=1}^{n} X_{ij} X_{ik}$$

E(w) を最小にするパラメータ w は計算できる。

ベクトル S と行列 I を
次のように定義する

$$\begin{cases} S_k = \sum_{i=1}^{n} Y_i X_{ik} \\ I_{jk} = \sum_{i=1}^{n} X_{ij} X_{ik} \end{cases}$$

この量はどちらも
データから計算できる

連立方程式は

$$S_k = \sum_{j=1}^{d} w_j I_{jk}$$

行列を使って書くと

$$S = Iw$$

答えは

$$w = I^{-1} S$$

素朴な質問：最小二乗法について

(1) 何のための最小二乗法？

「真のパラメータ w_0 を知りたい」
「未来の X についてYの予測誤差を小さくしたい」
「 X の要素で必要なものと不要なものを分けたい」

(2) 最小二乗法って最高?

最小二乗法は最初に学ぶ基本です。
目的に応じてもっと高精度な方法があります。

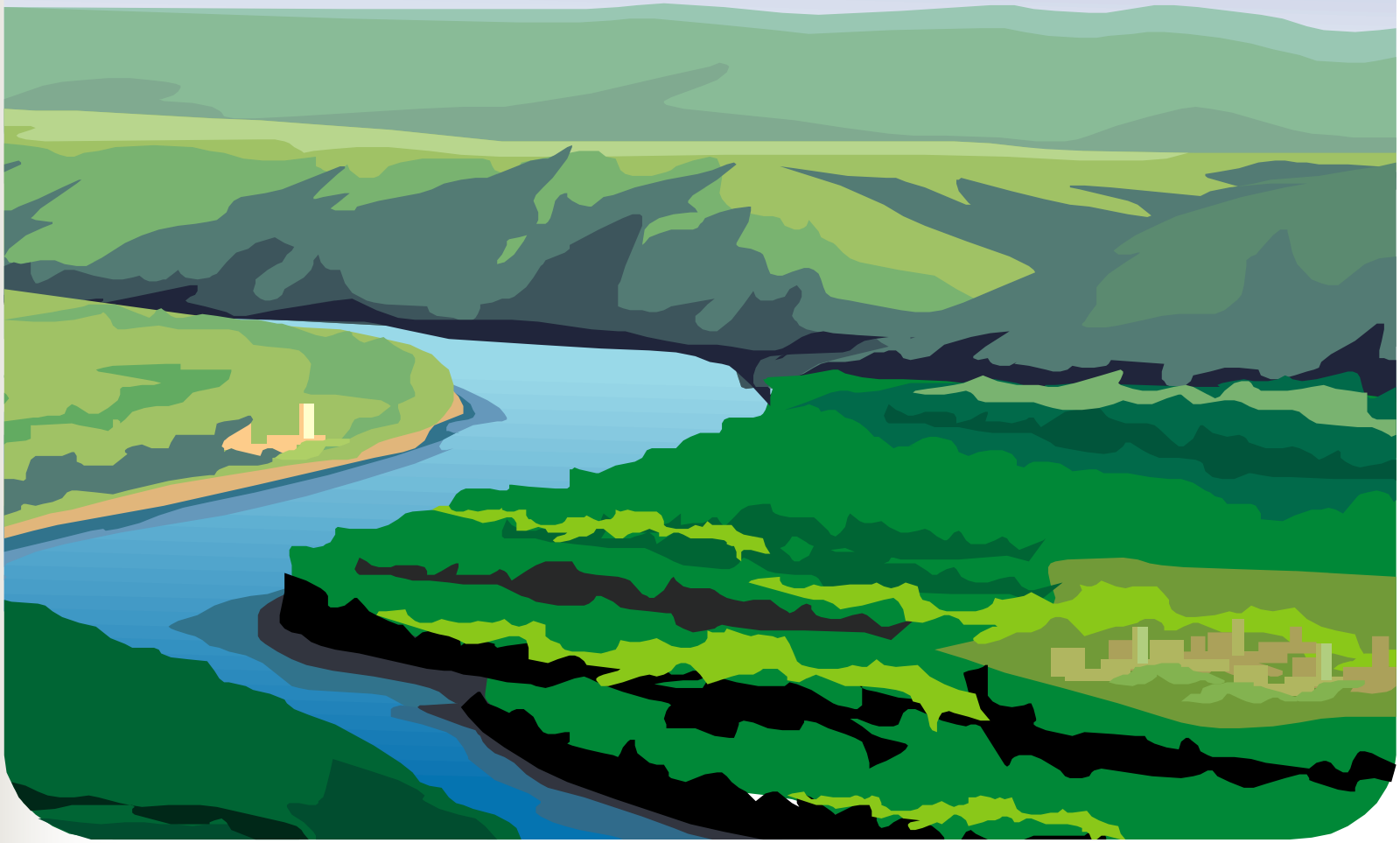
問1　データ　$\{ (X_i , Y_i) ; i=1,2,\ldots,n \}$

二乗誤差

$$E(a)=\sum_{i=1}^{n} (Y_i - aX_i)^2$$

を最小にする a を求めよ。

学習のモデルは毎年毎年よく似ているが
学習理論の応用は毎年毎年移り変わる



学習して何を知りたいのか

入力　X

教師　$Y = g(X) + 雑音$

データ

$\{(X_i, Y_i)\}$

生徒　$Y = f(X,w) + 雑音$

学習後　$Y = f(X,w^*) + 雑音$

学習誤差 =「学習後の二乗誤差」

汎化誤差 =「真g(X) と学習後 f(X,w*) の差」

学習誤差と汎化誤差

二乗誤差を最小にするパラメータを w^* とし、
学習誤差 $E(w^*)$ と汎化誤差 $G(w^*)$ を次式で定義する。

$$E(w^*) = (1/n) \sum_{i=1}^{n} (Y_i - f(X_i, w^*))^2$$

$$G(w^*) = \iint (y - f(x, w^*))^2 \, q(y|x) q(x) dx dy$$

汎化誤差の原因

バイアス（関数近似誤差）

どんなにパラメータ w を工夫しても
f(x,w) は g(x) になれない

バリアンス（統計的推定誤差）

データに雑音が含まれるので
w の推定がばらつく

バイアスとバリアンス

バイアス　　モデルが複雑なほど小さい
　　サンプル数が増えても減らない

バリアンス　　モデルが単純なほど小さい
　　サンプル数が増えると減る

両立しない　→　バイアス・バリアンス問題

現実の問題において必ず現れる

☆　バイアスは分かりやすいがバリアンスは分かりにくい。
実応用で「バイアスを減らすことだけに集中」が頻発する。

（例題） 線形テンソルモデル

入力　$X=(X_1,X_2)$

教師　$Y = g(X_1,X_2) +$ 正規雑音

生徒　$Y = \sum_{0 \leqq j,k \leqq 4} w_{jk} (X_1)^j (X_2)^k +$ 正規雑音

学習例 $\{ (X_{1i},X_{2i},Y_i) ; i=1,2,\ldots,n \}$ が与えられれば
パラメータ w_{jk} は最小二乗法で求められる。

実際に学習してみよう

ケース1

教師：$Y = X_1X_2$ + 正規雑音

テンソルモデルで**実現可能**
バイアス＝0

ケース2

教師：$Y = \exp(-5*(X_1^2+X_2^2))$+ 正規雑音

テンソルモデルで**実現不可能**
バイアスは零ではない。

問2

汎化誤差(Generalization Error)の値を計算して書き入れましょう。

	n=100	n=1000
実現可能		
実現不可能		