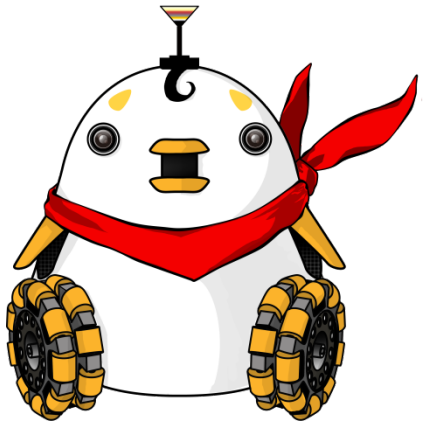
人工知能

第11回学習と認識(2)
パターン認識

立命館大学 情報理工学部 知能情報学科
萩原良信



STORY 学習と認識(2)

- ホイールダック2号はクラスタリングによって，目で見た物体をいくつかのグループに分けることに成功した．これで，新しい物体を見たときにもその物体がどのグループに属するかがわかるだろう．そうすれば，ホイールダック2号は目の前にあるものが何かわかるに違いない．例えば，目の前の対象が宝箱なのかゴールなのかがわかるに違いない．
- しかし，ホイールダック2号は宝箱を五つほど開けたところで気づいた．「どうやら，宝箱には財宝が入っているものと，罠が入っているものがあるらしい．」その2種類はどうも宝箱の見た目が少し違うようなのだが，他のゴールや普通の道に比べると，よく似ていたために，教師なし学習のクラスタリングの結果としては，同じクラスタになっていた．
- これではホイールダック2号にとっては区別がつかない．しかし，この「財宝が入っていた」宝箱の画像と「罠が入っていた」宝箱の画像を集めれば，その違いを学習することができるのではないだろうか．

仮定　　学習と認識(2)

- ホイールダック2号は適切な画像特徴量を有限次元ベクトルで取得できるものとする.
- ホイールダック2号は分類のための教師信号を認識することができるものとする.

Contents

11.1.1 機械学習の分類

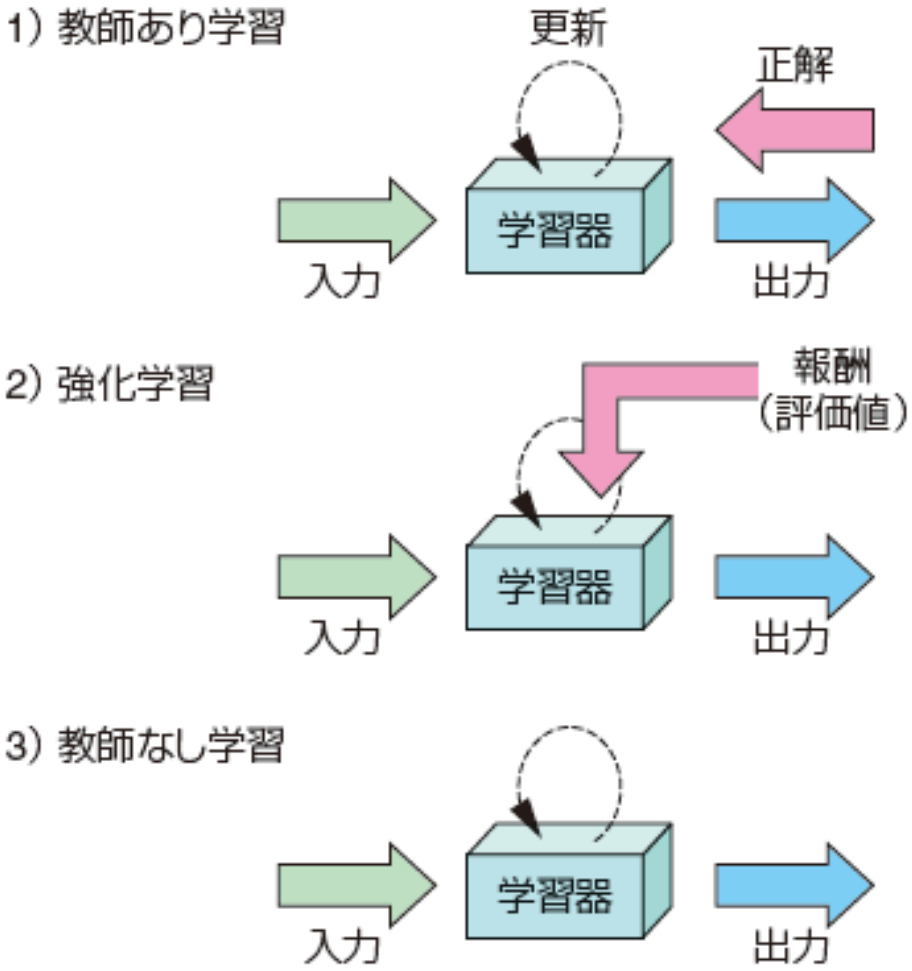


図 11.2 機械学習の分類

- 内部モデル学習
- 時系列データ学習
- 回帰問題

- 強化学習
- 最適化問題
- Genetic Algorithm

- クラスタリング
- 低次元化
- データマイニング

11.1.5 機械学習の共通問題

- 結局は関数fの最適化
  - 学習器は結局は入力から出力への変換を学習する数学的存在としてモデル化される.
  - より具体的に言うと，学習器は何らかの関数fを持ち，これを関数fの内部パラメータθを変化させることで学習する.
  - このθはニューラルネットワークの結合重みであったり，強化学習器のQ 値であったりする.
- **訓練データ(training data)とテストデータ(test data)**
  - 機械学習においては学習用データとテスト用データを区別することが重要である.
  - 特に教師あり学習では学習用データに対しては教師信号として「答え」が与えられるため，正しい「答え」を出力できるようになるのは当たり前である.
  - 学習用データで学習した学習器が，テスト用データに対して正しい答えを返せるようになるのが大切である.

演習11–1 機械学習の分類

- 以下の機械学習はそれぞれ「教師あり学習」「教師なし学習」「強化学習」のいずれにあたるか？
  1. 問題を解くと得点だけがしめされて，「後のことは自分で考えなさい！」と言われる試験
  2. 問題を解くとそれぞれの解答が示されて「後のことは自分で考えなさい！」と言われる試験
  3. 100人のマンガのキャラの絵を見せられて「キャラの類似性にもとづいて10グループにわけよ」と言われる課題
  4. 100件のワンルーム不動産の物件に対して，駅からの距離，床面積，風呂トイレの有無，賃料を収集し，駅からの距離，床面積，風呂トイレの有無から賃料を予測出来るようにするタスク．

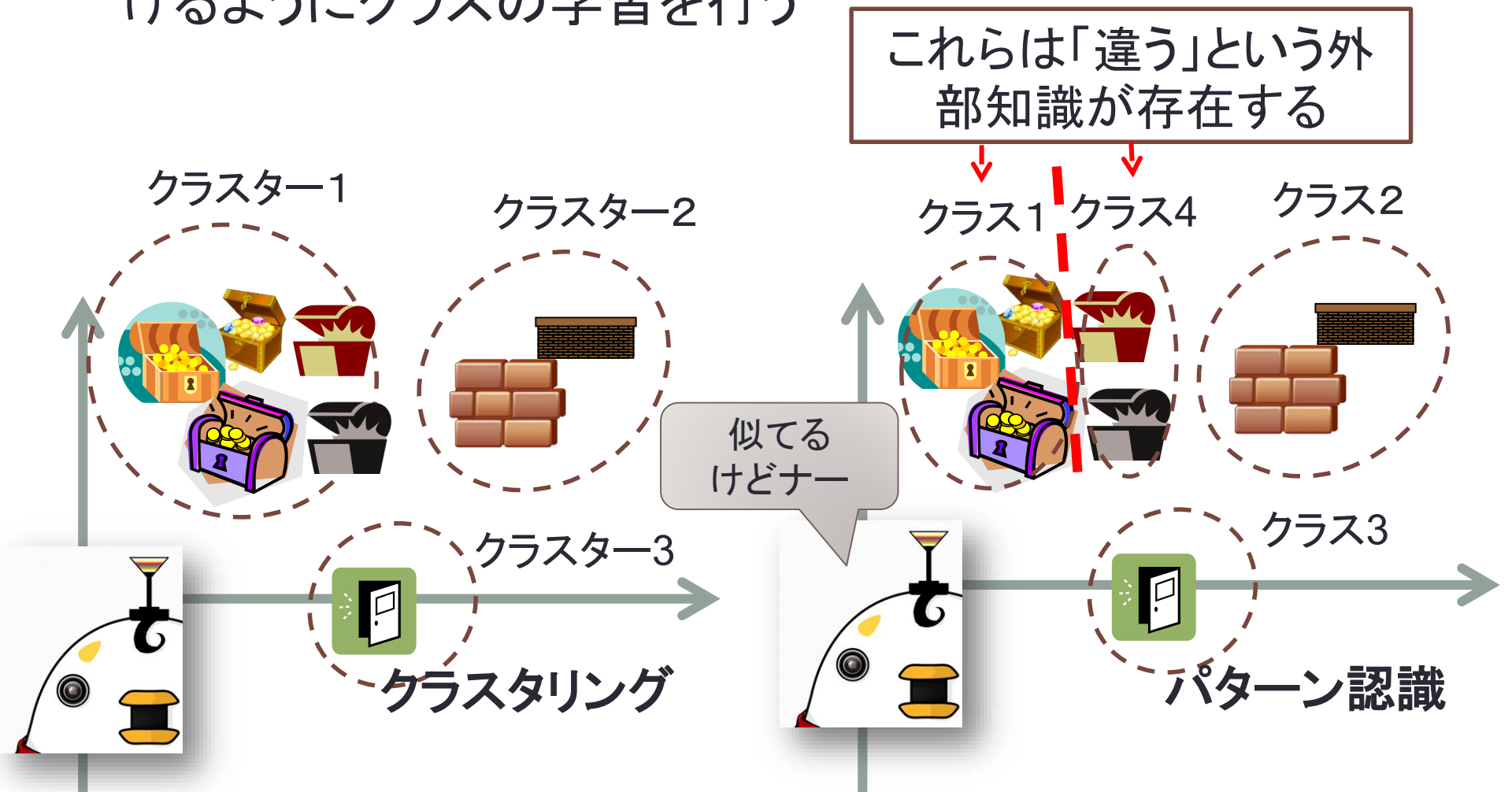
Contents

11.2.1 パターン認識と応用

- パターン認識とは画像や音声などデータに対して行う情報処理で，観測されたデータを予め定められた複数の概念のうちの一つに対応させる処理である．この概念は**クラス**と呼ばれる．
- 文字認識(character recognition)
  - 画像データを認識して文字の種類を認識する
  - タッチペン入力の書き文字認識など
- 音声認識(speech recognition)
  - 人間の声を認識して文字列として解釈する．
  - モバイルデバイスでの音声情報検索など
- 画像認識(image recognition)
  - カメラ画像に写った物体が何の物体であるか認識する一般物体認識，表情認識などがある．

クラスタリングとパターン認識の違い

- 画像の異なり具合を基準にしてクラスの境界を引くよりも，外部的な知識である「ルール」に基いて，その違いを見分けるようにクラスの学習を行う

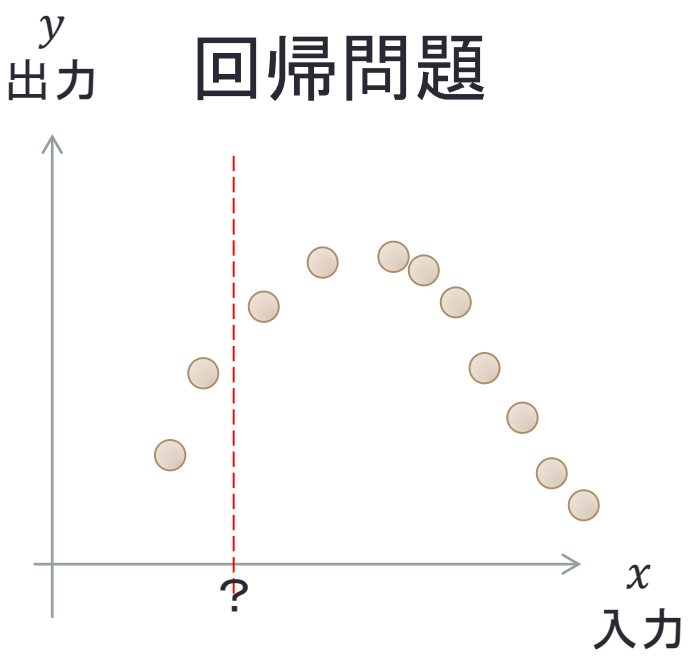


11.2.2 回帰問題と分類問題

- 目的
  - 入力ベクトルxに対して正しい出力ベクトルyを出力出来るようになること.
- 問題の分類
  - 回帰問題 (regression)
    - 入力ベクトルに対して通常**実数値の値**を返し，未知入力に対する出力の予測を行う.
    - 学習データとしては(x, y)の値の組が渡される.
  - 分類問題 (classification)
    - 入力ベクトルに対して正事例であるか負事例であるかの**二値$\{1, 0\}$の値**を返すことで分類を行う.（多値のものもあり）
    - 学習データとしては正負のラベルの付けられたデータセットを用いる.

11.2.2 回帰問題(regression)

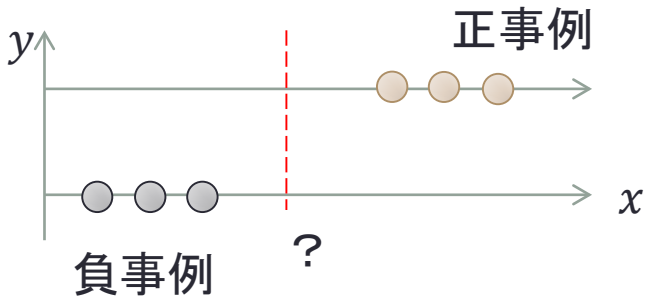
- 回帰問題は入力ベクトルに対して実数値の値を返す連続的な関数関係を学習する問題である．学習後は未知入力に対する出力値の予測を行う．学習データとしては入力ベクトルxと出力ベクトル（もしくは出力値）yの組み合わせ(x, y)の集合が学習器に渡される．
- 様々な(x, y)上の点が与えられた時に未知の入力，たとえば，？マークの位置の入力に対する出力yを答えるのが回帰問題である．

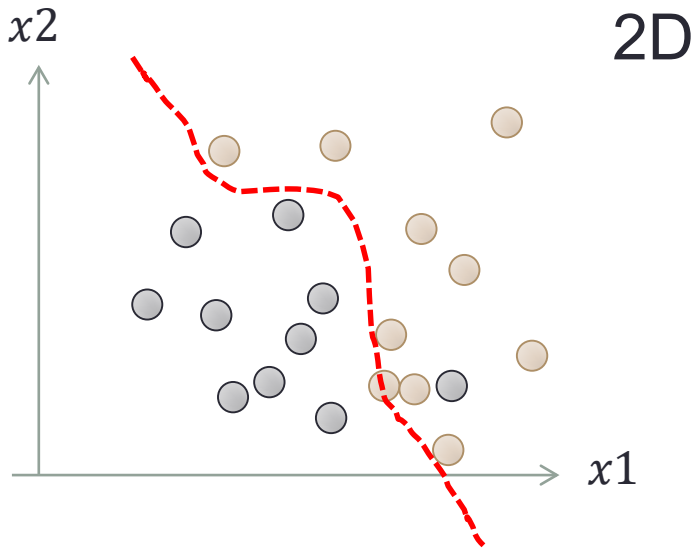


結局は$y = f(x)$のfの推定問題となる場合が多い

11.2.3 分類問題(classification)

- 分類問題は入力ベクトルに対して正事例(true)か負事例(false)かを返す法則を学習する問題である.
- もしくは，有限個のクラスのどれに属するかを学習する問題.
- 様々な(x, y)上の点が与えられた時に未知の入力，たとえば，？マークの位置の入力に対する出力yを答えるのが分類問題である.





結局は$y = f(x)$のfの推定問題となる場合が多い

主要な手法

回帰問題

- 線形回帰
- 一般線形モデル
- ニューラルネットワーク
- カーネル回帰
- ガウス過程回帰(GP)
- その他

分類問題

- パーセプトロン
- ニューラルネットワーク
- SVM(サポートベクターマシン)
- ランダムフォレスト
- 混合ガウス分布
- ナイーブベイズフィルタ
- その他

演習11-2 教師あり学習の分類

- 以下の学習はそれぞれ「分類問題」「回帰問題」のいずれにあたるか？
  1. カピバラの写真10枚を「これがカピバラだ」と見せられた後に，デグーの写真10枚を「これがデグーだ」と見せられる．その後にどちらかの写真を見せられて，それが何かを当てる課題．
  2. 自分一人でペットボトルに入れるビー玉の数を変えては，風呂に投げ入れ，沈むかどうかを判定し，何個入れれば風呂の水に沈むかというルールを学習すること．
  3. 100件のワンルーム不動産の物件に対して，駅からの距離，床面積，風呂トイレの有無，賃料を収集し，駅からの距離，床面積，風呂トイレの有無から賃料を予測出来るようにするタスク．
  4. 初速度をV [m/s]して弾丸を射出し，その落下点x[m]を多数計測することで V-xの関係を学習し一般法則を導き出そうとすること．

カピバラとデグーの分類

カピバラ

カピバラ

テストデータ：これはどっち？

Contents

11.3.1 予測誤差最小化による学習

- 回帰問題を解くための最も基本的な方法は，入力xと出力yの関係が$y = f(x;\theta)$という関係にあると考え，予測誤差を最小化するように学習器の最適なパラメータθ^*を求める方法である．

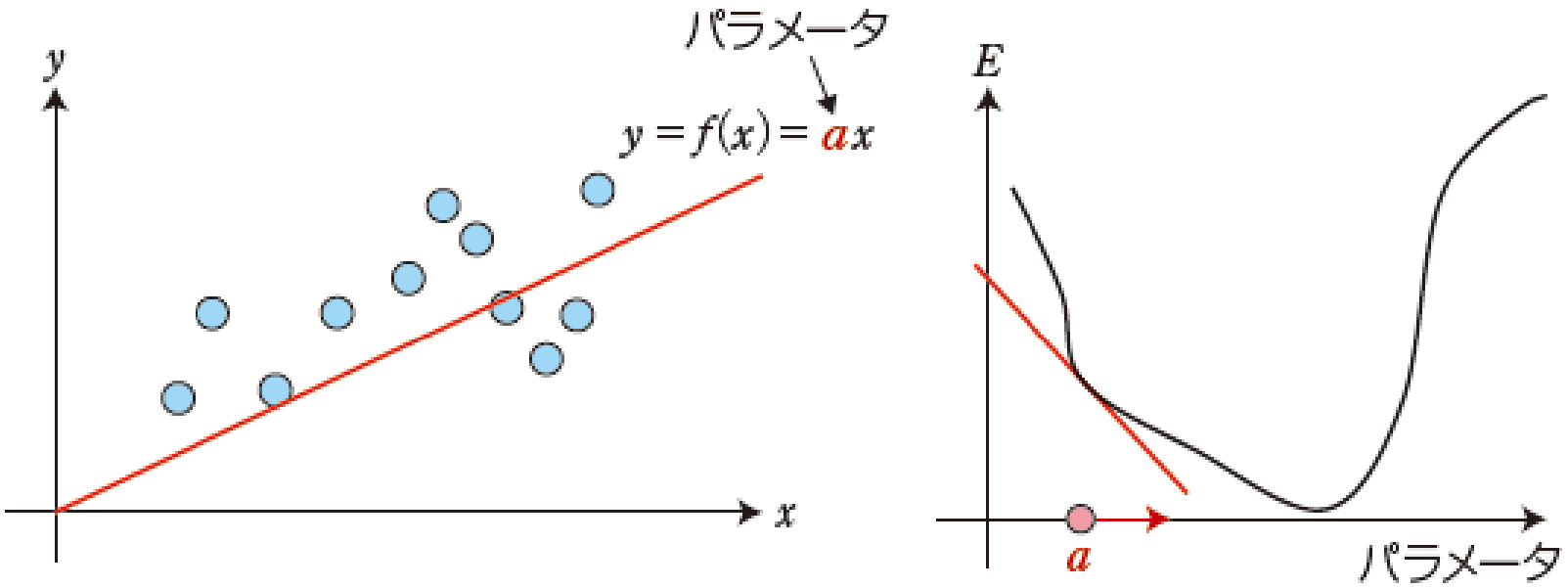


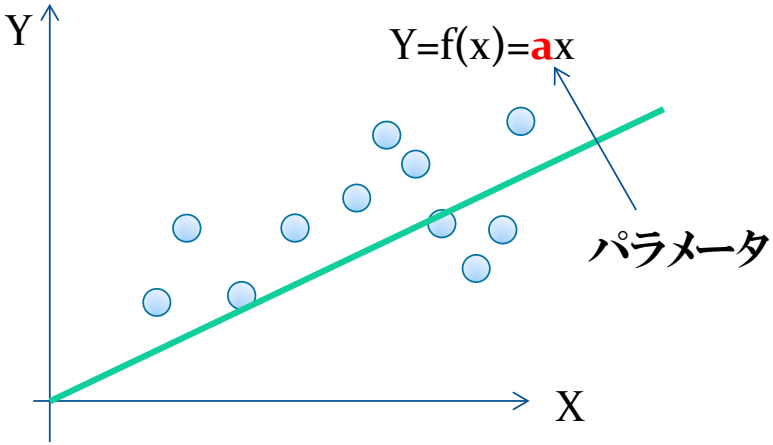
図 11.4 回帰問題と最急降下法

予測誤差最小化

- 与えられたデータに対して近似誤差が最小になるように関数$f(x)$のパラメータを調整する.
- 最急降下法（又は勾配法）
  - 誤差が徐々に小さくなるように，誤差の偏微分を計算して逆方向にパラメータを修正
- 最小二乗法
  - $f(x)$が線形関数の場合は解析的に（閉形式で）解ける

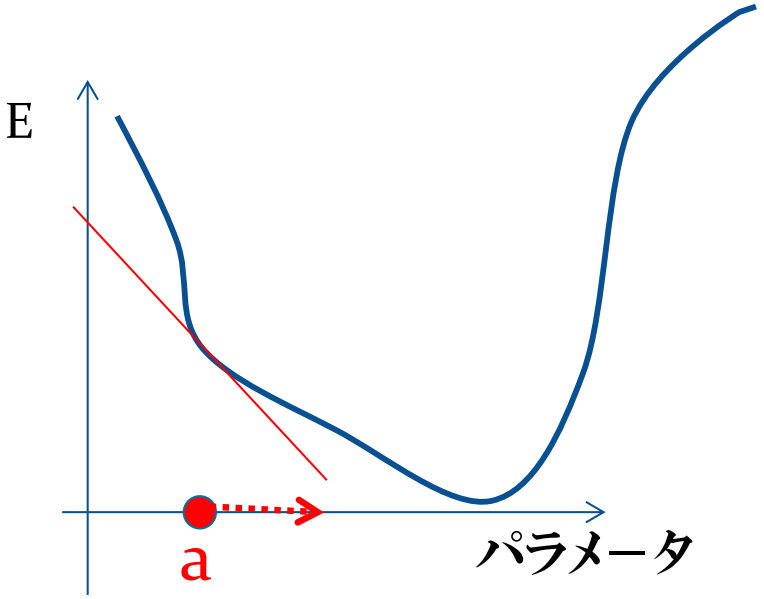
$$E = \sum_i ||y_i - f(x_i)||^2 \longrightarrow \text{最小化}$$

$$(x_i, y_i) \in D$$



最急降下法(勾配法)

- 誤差Eが徐々に小さくなるように，誤差の偏微分を計算して逆方向にパラメータを修正
- 具体的には$\theta^{(k)} = a$において勾配(gradient)$\partial E / \partial \theta$を計算し，その逆方向に更新する．$\theta$を誤差最小の方向に降下させる．



f のパラメータを θ とすると，k ステップにおいて

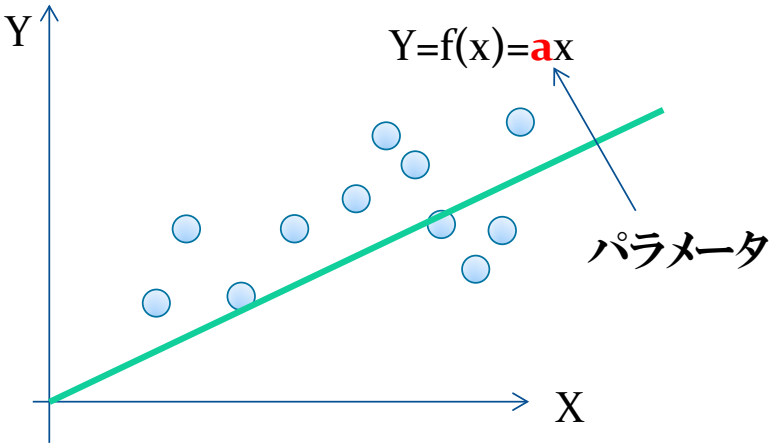
$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \frac{\partial E}{\partial \theta}|_{\theta=\theta^{(k)}}$$

線形回帰:最小二乗法

- 最小二乗法
  - $f(x)$が線形関数の場合は解析的に(閉形式で)解ける

$$E = \sum_i ||y_i - f(x_i)||^2 \longrightarrow \text{最小化}$$

$$(x_i, y_i) \in D$$



演習11–3 最小二乗法

- xとyは本質的には線形関係を持っている($y = ax + b$). しかし，xに対するyの値を計測する時に必ず誤差が生じる.
- (x, y) =(1,2),(2,4),(3,5),(4,7)の観測が得られた際に，最小二乗法にもとづいてa, bを求めよ.

- 誤差EとEを最小化する極値の式を以下とする

$$\mathrm{E}(a, b) = \sum_{i \in \{1,2,3,4\}} (y_i - ax_i - b)^2$$

$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$$

11.3.3 一般線形モデル

- 線形回帰では，線形な関数，つまりグラフにプロットしたときに直線や平面になる関数関係しかモデル化できない．
- 線形回帰の枠組みをそのまま拡張し，非線形関数に対応する簡便な方法として一般線形モデルが存在する．
- 基底関数$b_i(x)$の線形和でモデル化する．
- ω_iは重みである．

$$f(x) = \sum_i w_i b_i(x) = w^T b(x)$$

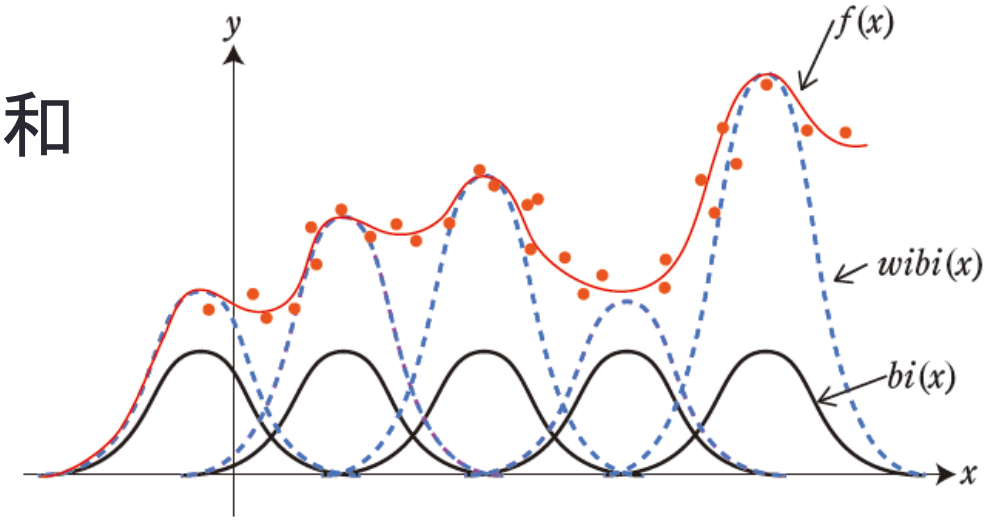


図 11.5 一般線形モデルによる回帰

11.3.4 ニューラルネットワーク

- 人の脳で行われている情報処理を**模倣した**情報処理モデルである.
- 回帰問題にも分類問題にも用いることができる.
- BP法(誤差逆伝搬法:勾配法の効率的計算方法)によって学習する.

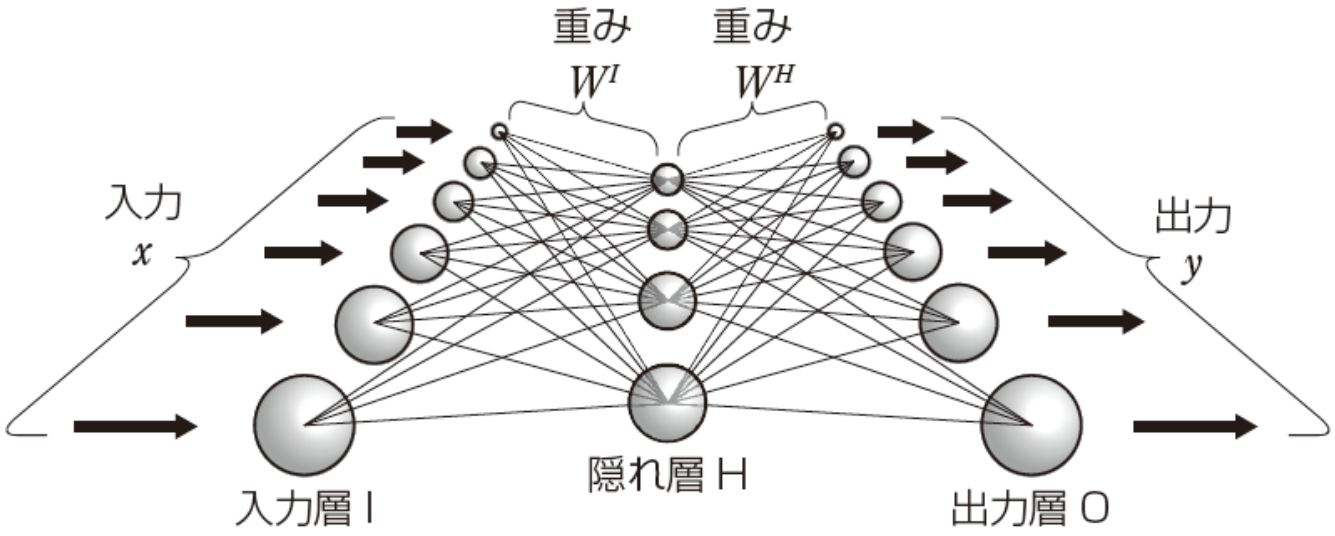


図 11.6 ニューラルネットワーク

Contents

11.4.1 識別モデルと生成モデル

- 識別モデル
  - 正事例と負事例を区別するための境界線を訓練データから直接的に求めようとする. どう分けるか？を考える
- 生成モデル
  - 分類対象となるデータがどのような確率モデルから生成されたかをモデル化し，そのモデルに基づいて分類を行う. 生成された過程を考える

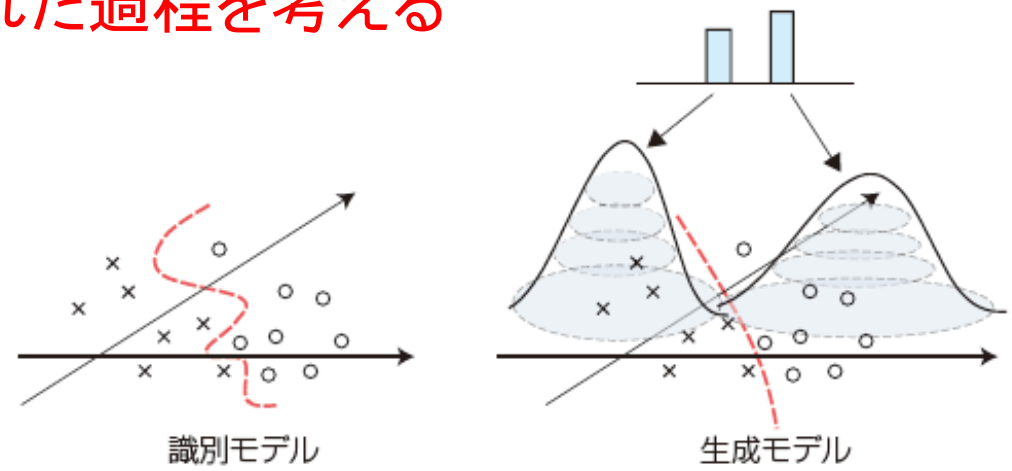


図 11.7 識別モデルと生成モデル

11.4.2 サポートベクトルマシン

- SVM は線形分類器であるパーセプトロンにカーネル法(kernel method)を組み合わせることによって実現される

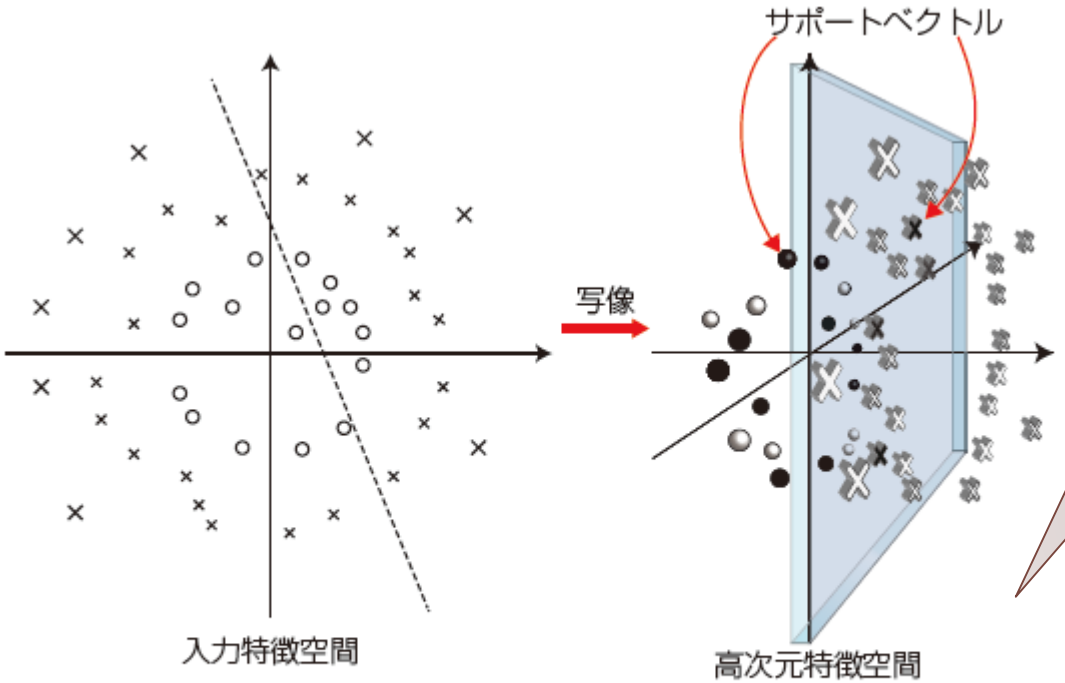


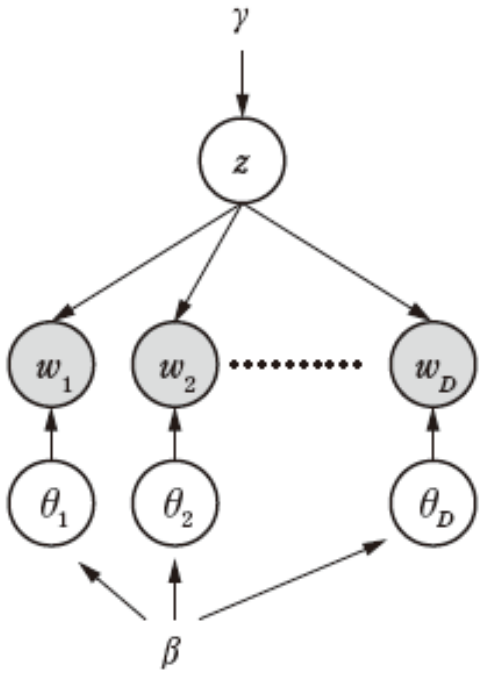
図 11.8 サポートベクトルマシンのイメージ図

複雑な分離面も表現可能．汎化性能が大変高い！便利！

識別モデル

11.4.3 ナイーブベイズモデル

- ナイーブベイズモデル(naive Bayes model) は生成モデルに基づき分類を行うために用いられる最も単純なモデルの一つである.



β, γ：ハイパーパラメータ

z：スパムメールかどうかを表す変数

w_i：単語が観測されたかどうかを表す変数

θ_i：単語の発生確率

生成モデル

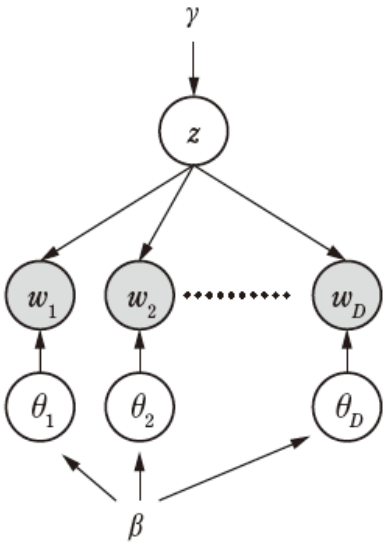
図 11.9 ナイーブベイズモデルのグラフィカルモデル

スパムメールのナイーブベイズフィルタ

- メールがスパムメールかどうかを判定する分類問題を考える.

$$P(z|W) = P(z|w_1, w_2, \ldots, w_D)$$
$$\propto P(w_1, w_2, \ldots, w_D|z)P(z)$$
$$= \left(\prod_{i=1}^{D} P(w_i|z)\right) P(z)$$

スパムメールが届く確率



メールにワードWが含まれる時にスパムメールである確率$P(z=1|W)$を求める

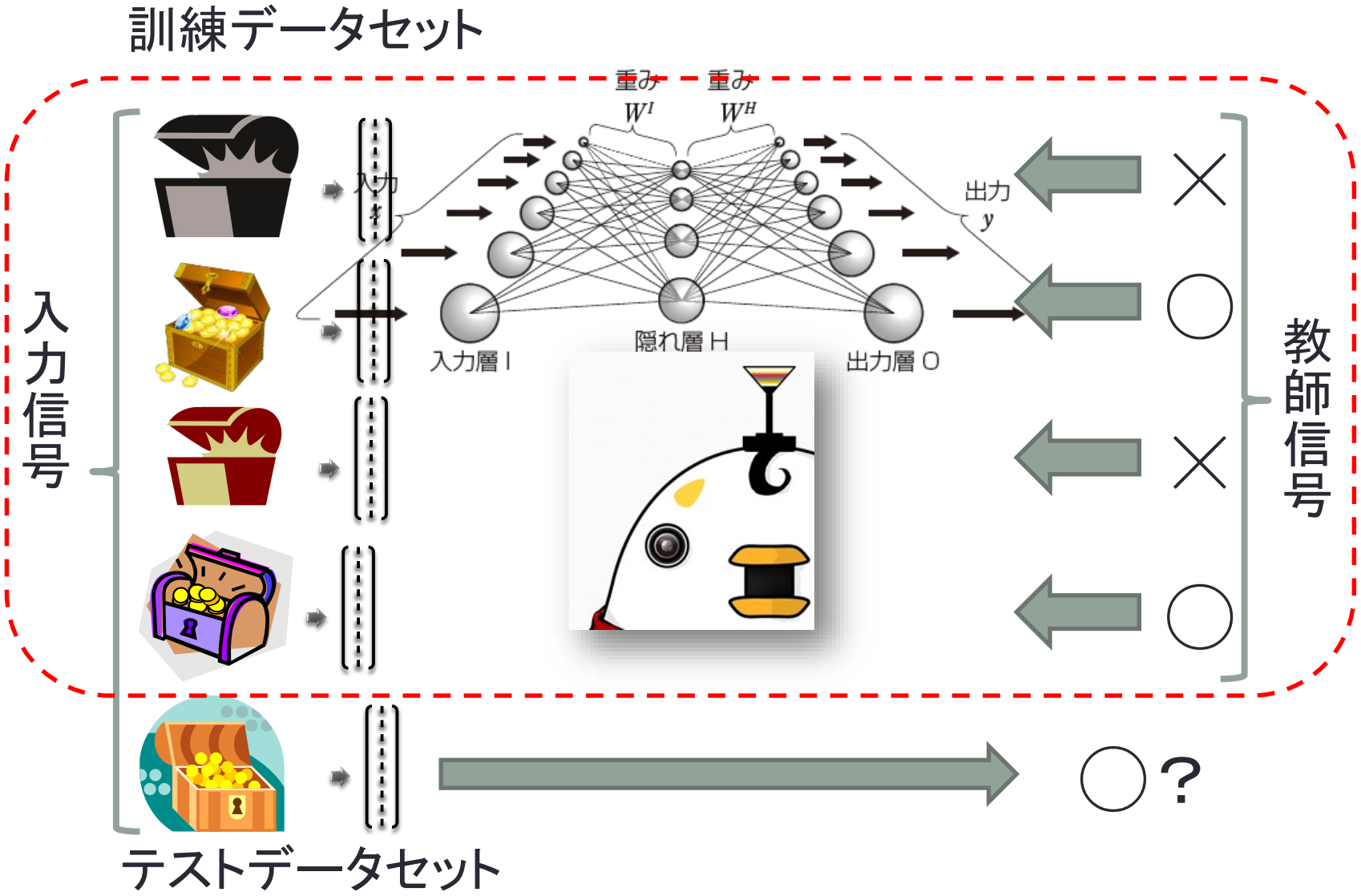
スパムフィルタがつくれます！

表 11.1 スパムメール分類問題における単語出力分布の例

	w_1: “お世話”	w_2: “お得”	w_3: “女子高生”
$P(w_i\|z=1)$: スパム	0.05	0.60	0.30
$P(w_i\|z=0)$: 正常	0.30	0.10	0.01

訓練データから学習可能！

ホイールダック2号の学習



まとめ

- 機械学習の分類法について学んだ.
- パターン認識とその応用事例について概要を学んだ.
- 回帰問題と分類問題の区別について学んだ.
- 線形回帰および一般線形モデルにおける最小二乗法について学んだ.
- ニューラルネットワークとその学習方法について簡単に学んだ.
- 識別モデルと生成モデルの区別について学んだ.
- ナイーブベイズモデルについてスパムメールフィルタの事例を交えて学んだ.