

# 13. 系列データの識別

## 13.1 ラベル系列に対する識別

- ラベル系列に対する識別問題の分類
  - 入力の系列長と出力の系列長が等しい
    - 例) 形態素解析、固有表現抽出
    - 系列ラベリング問題  $\Rightarrow$  CRF
  - 入力の系列長に関わらず出力の系列長が 1
    - 例) 動画像の分類、話者認識
    - 系列識別問題  $\Rightarrow$  HMM
  - 入力の系列長と出力の系列長に対応関係がない
    - 例) 連続音声認識
    - 系列変換問題  $\Rightarrow$  RNN

## 13.2 系列ラベリング問題— CRF—

- 系列ラベリング問題とは
  - 入力系列の個々の要素に対してラベルを付与する
  - 系列の要素の出現確率は前後の要素に依存
    - 1 入力 1 出力の識別器を連続的に適用する方法では性能が上がらない
      - ⇒ 入力や出力の系列としての特徴を使う
  - 可能な出力系列の組合せは膨大な数になるので、単純な事後確率最大法は使えない
    - ⇒ 探索によって最適解を求める

## 13.2 系列ラベリング問題— CRF—

- 系列ラベリング問題の事例

- 形態素解析

入力	系列	で	入力	さ	れる	各	要素
出力	名詞	助詞	名詞	動詞	接尾辞	接頭辞	名詞

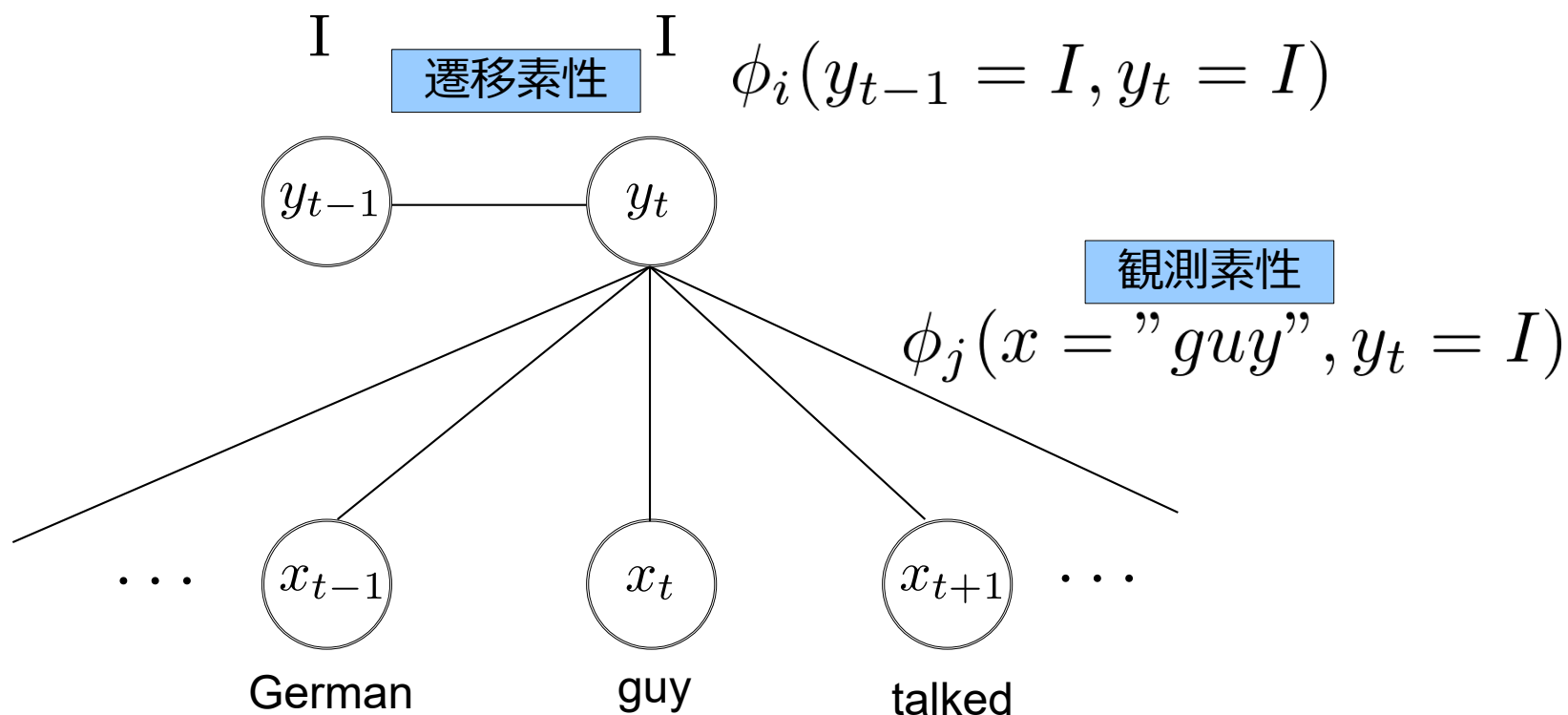
- 固有表現抽出（例：人を指す表現の抽出）

入力	Suddenly,	the	tall	German	guy	talked	to	me
出力	O	B	I	I	I	O	O	B

B: begin  
I: inside  
O: outside

## 13.2 系列ラベリング問題— CRF—

- 対数線型モデルによる系列ラベリング
  - 素性関数の導入



## 13.2 系列ラベリング問題— CRF—

- 対数線型モデル (softmax)

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x},\mathbf{w}}} \exp(\mathbf{w} \cdot \phi(\mathbf{x}, \mathbf{y}))$$

$$Z_{\mathbf{x},\mathbf{w}} = \sum_{\mathbf{y}} \exp(\mathbf{w} \cdot \phi(\mathbf{x}, \mathbf{y}))$$

識別モデル

あるクラスの事後確率が上がれば、他のクラスは下がる

- 出力の決定

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$$

$$= \arg \max_{\mathbf{y}} \frac{1}{Z_{\mathbf{x},\mathbf{w}}} \exp(\mathbf{w} \cdot \phi(\mathbf{x}, \mathbf{y}))$$

$$= \arg \max_{\mathbf{y}} \mathbf{w} \cdot \phi(\mathbf{x}, \mathbf{y})$$

## 13.2 系列ラベリング問題— CRF—

- 素性関数の制限

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} \sum_t \mathbf{w} \cdot \phi(\mathbf{x}, y_t, y_{t-1})$$

- ビタビアルゴリズムによって探索が可能

---

**Algorithm 12.1** ビタビアルゴリズム

---

for  $t = 2$  to  $|\mathbf{x}|$  do

  for all  $y_t$  do

$$\alpha(t, y_t) = \max_{y_{t-1}} \{ \mathbf{w} \cdot \phi(\mathbf{x}, y_t, y_{t-1}) + \alpha(t-1, y_{t-1}) \}$$

$$B(t, y_t) = \arg \max_{y_{t-1}} \{ \mathbf{w} \cdot \phi(\mathbf{x}, y_t, y_{t-1}) + \alpha(t-1, y_{t-1}) \}$$

  end for

end for

$\mathbf{y}^* = \alpha$  の最大値に対応する  $B$  を逆に辿る

---

## 13.3 系列識別問題— HMM—

- 例題

- PC 操作系列による熟練度の判定

- k: キーボード、 g: マウス、 e: エラー
- 初心者の入力系列例

k e k g k e k g g k g k k e g e e k e e e g e

- 熟練者の入力系列例

k k e k g k k k e k g k g g g e g k g

- 判定したい入力系列

k g e k g k k g e k g e k e e k e g e k

## 13.3 系列識別問題— HMM—

- 生成モデルによるアプローチ
  - 系列識別問題ではクラスの事前確率を得られることが多い

$$\begin{aligned} y^* &= \arg \max_y P(y|\mathbf{x}) \\ &= \arg \max_y \frac{P(\mathbf{x}, y)}{P(\mathbf{x})} \\ &= \arg \max_y \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \\ &= \arg \max_y P(\mathbf{x}|y)P(y) \end{aligned}$$

生成モデル

尤度は、あるクラスの確率モデルを、他のクラスとは無関係に求めている

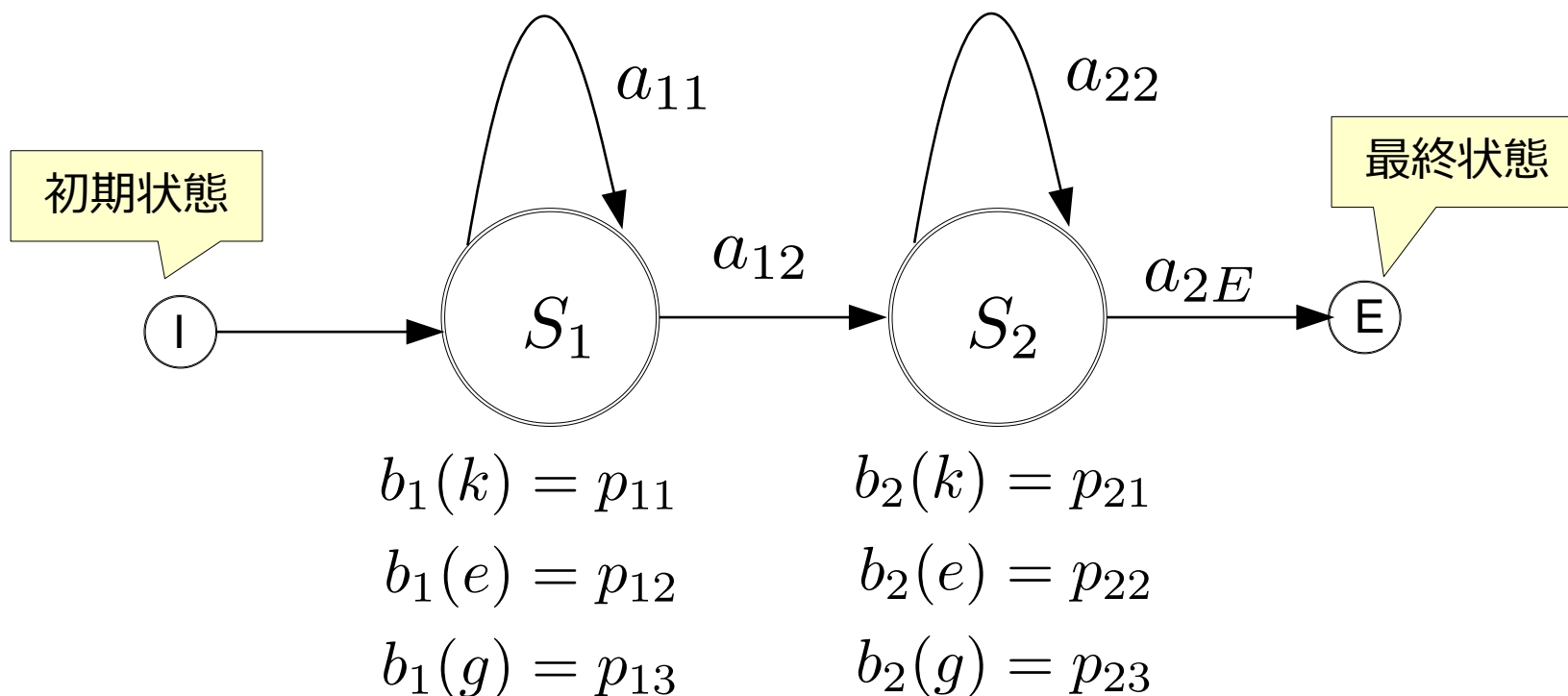
尤度

事前確率



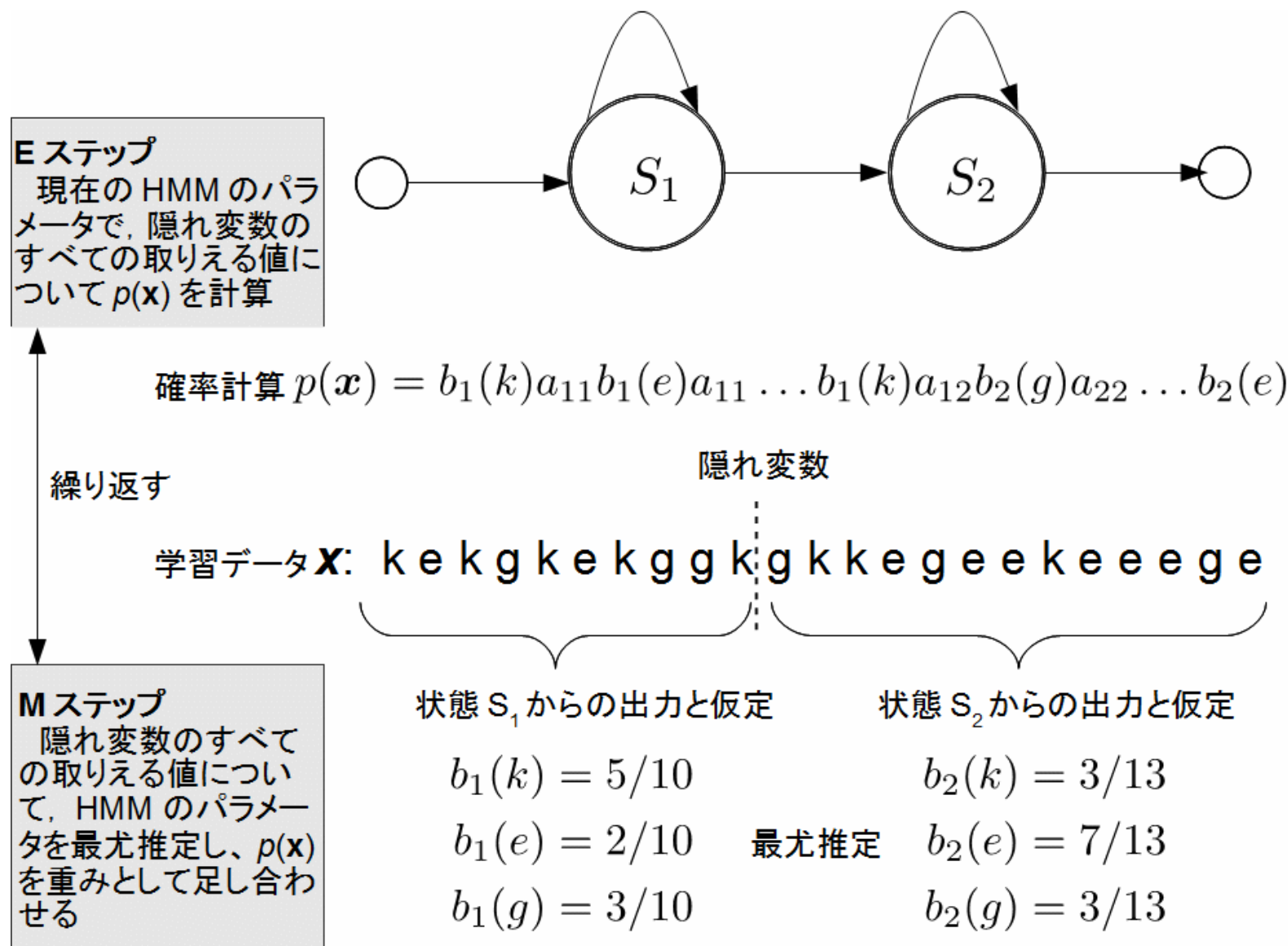
## 13.3 系列識別問題—HMM—

- 不定長入力に対する尤度計算法
  - 自己遷移を持つ確率オートマトンを用いる



## 13.3 系列識別問題— HMM—

### • HMM の学習 : EM アルゴリズム



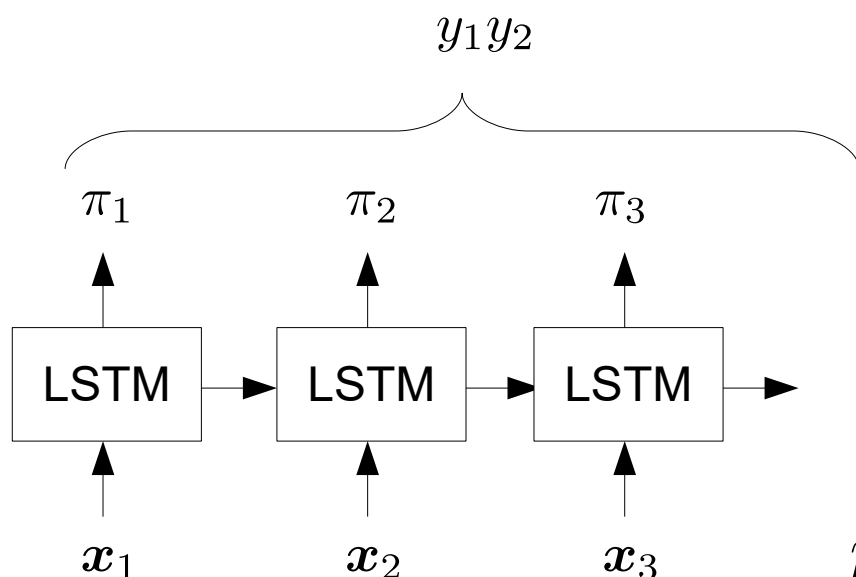
# 系列変換問題 — RNN—

- 系列変換問題の定式化
  - 入力系列  $\mathbf{x} = \mathbf{x}_1, \dots, \mathbf{x}_T$
  - 出力系列  $\mathbf{y} = y_1, \dots, y_L$
  - 一般に  $T > L$
  - 系列識別と探索を組み合わせた複雑な処理が必要
- End-to-End アプローチ
  - 入力から出力への変換をニューラルネットワークで学習
    - Connectionist temporal classification
    - Attention モデル

# Connectionist temporal classification

- アイディア

- 出力記号に blank 記号 \_ を加えて、入力長と出力長を合わせる
- 正解系列に変換可能な出力系列の確率の和を求める



- $h a i$  という正解系列に対して

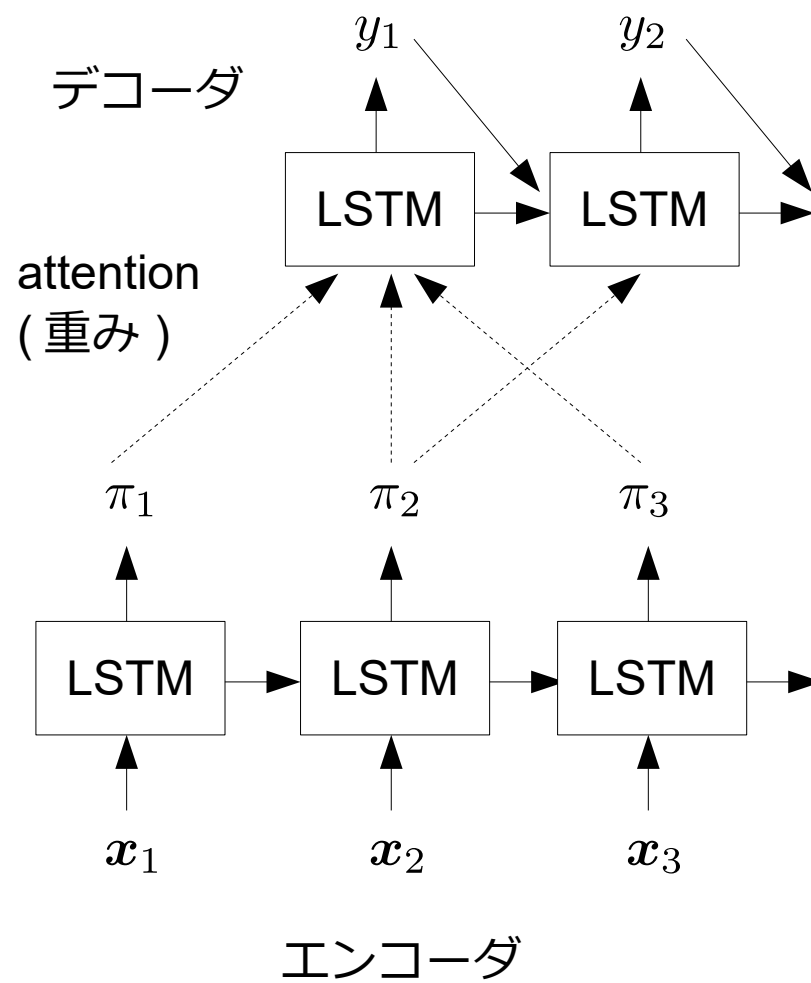
\_h\_\_a\_\_i  
hhh\_aaaa\_i  
hh\_\_\_\_\_ai\_\_

などの出力系列を正解とみなす

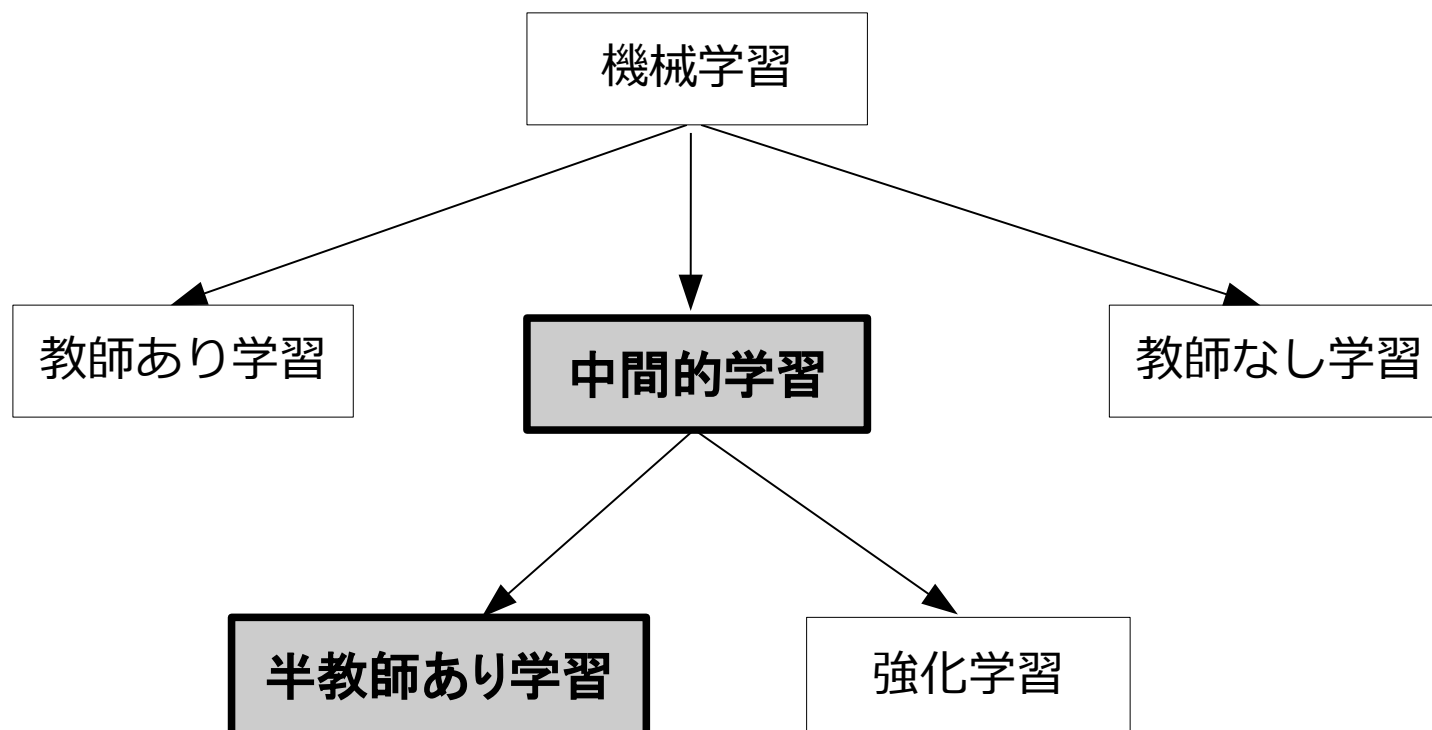
$$p(\mathbf{y}|\mathbf{x}) = \sum_{\boldsymbol{\pi} \rightarrow \mathbf{y}} p(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\boldsymbol{\pi} \rightarrow \mathbf{y}} \prod_{i=1}^T p(\pi_i|\mathbf{x}_i)$$

# Attention モデル

- アイディア
  - 特徴ベクトルを分散表現に変換するエンコーダと、分散表現から出力を求めるデコーダの組み合わせ
  - 一定範囲の分散表現から出力を計算するために注意機構 (attention) を用いる



# 14. 半教師あり学習

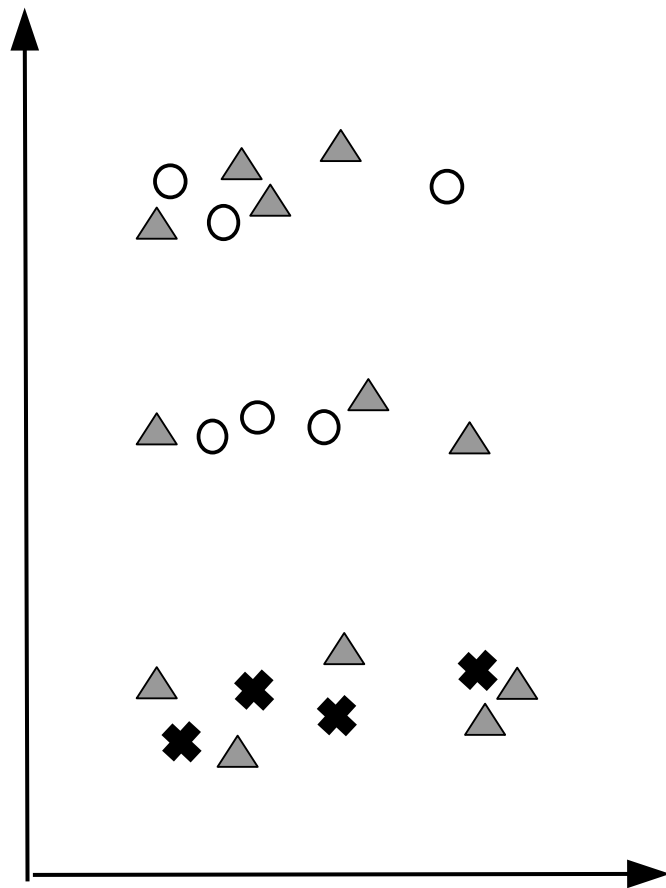


少数の正解付きデータ  
大量の正解なしデータ

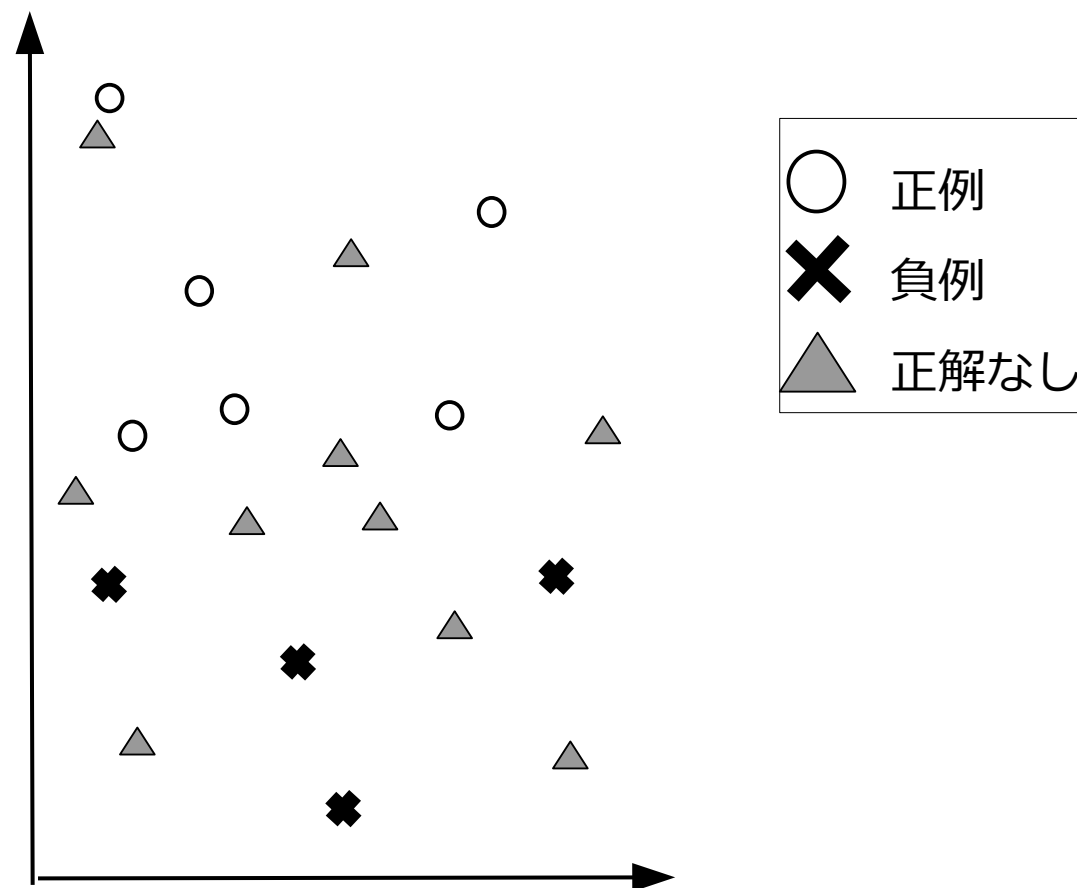
# 14.1 半教師あり学習とは

## 14.1.1 数値特徴の場合

- 半教師あり学習に適した数値特徴データの性質



半教師あり学習に適するデータ



半教師あり学習に適さないデータ

## 14.1.1 数値特徴の場合

- 半教師あり学習が可能なデータ
  - 半教師あり平滑性仮定
    - 二つの入力が高密度領域で近ければ、出力も関連している
  - クラスタ仮定
    - もし入力と同じクラスタに属するなら、それらは同じクラスになりやすい
  - 低密度分離
    - 識別境界は低密度領域にある
  - 多様体仮定
    - 高次元のデータは、低次元の多様体上に写像できる
      - 多様体：局所的に線形空間と見なせる空間



## 14.1.2 カテゴリ特徴の場合

- オーバーラップ
  - 文書からの評判分析の例

Positive ○

... よかった。 ..  
...  
高性能 ..  
... 満足

?

...  
...  
高性能 ..  
... 満足 .  
....

?

.....  
...  
高性能 ..  
...  
... よかった。

Negative ✕

... 壊れた。 ..  
...  
不満 ..  
...  
... 買わない

?

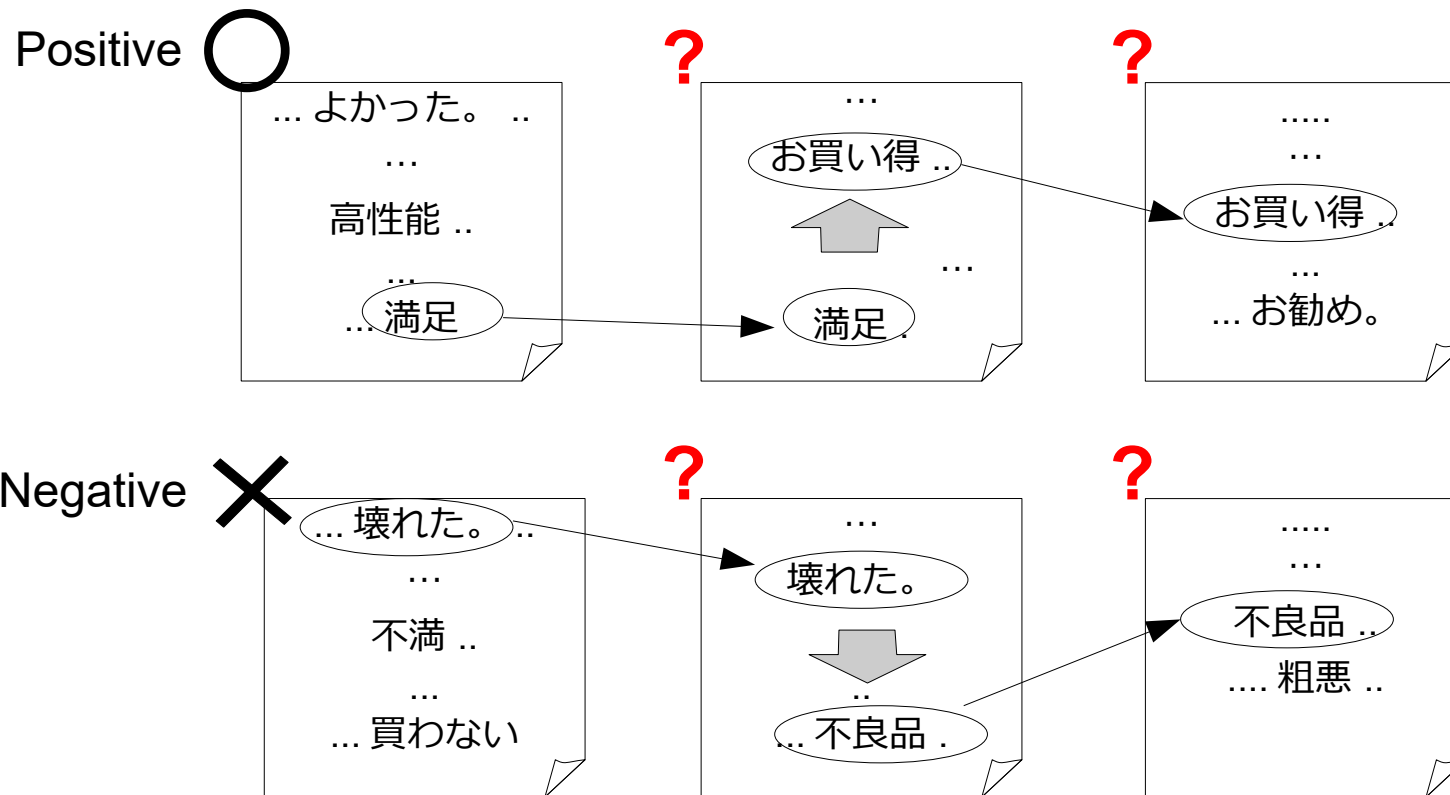
...  
...  
壊れた。 ..  
... 買わない .  
....

?

.....  
...  
不満 ..  
...  
... 買わない

## 14.1.2 カテゴリ特徴の場合

- 特徴の伝播



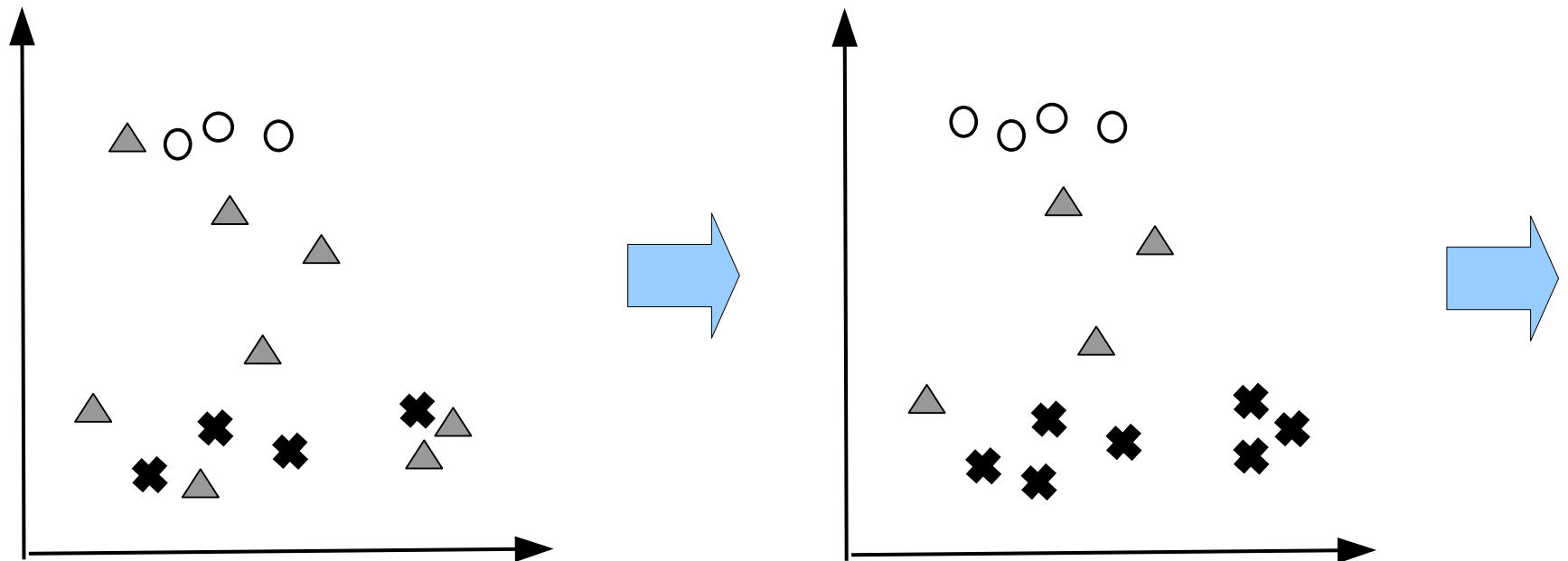
## 14.1.3 半教師あり学習のアルゴリズム

- 半教師あり学習の基本的な考え方
  - 正解付きデータで識別器を作成
  - 正解なしデータで識別器のパラメータを調整
- 識別器に対する要求
  - 確信度の出力：正解なしデータに対する出力を信用するかどうかの判定に必要

## 14.2 自己学習

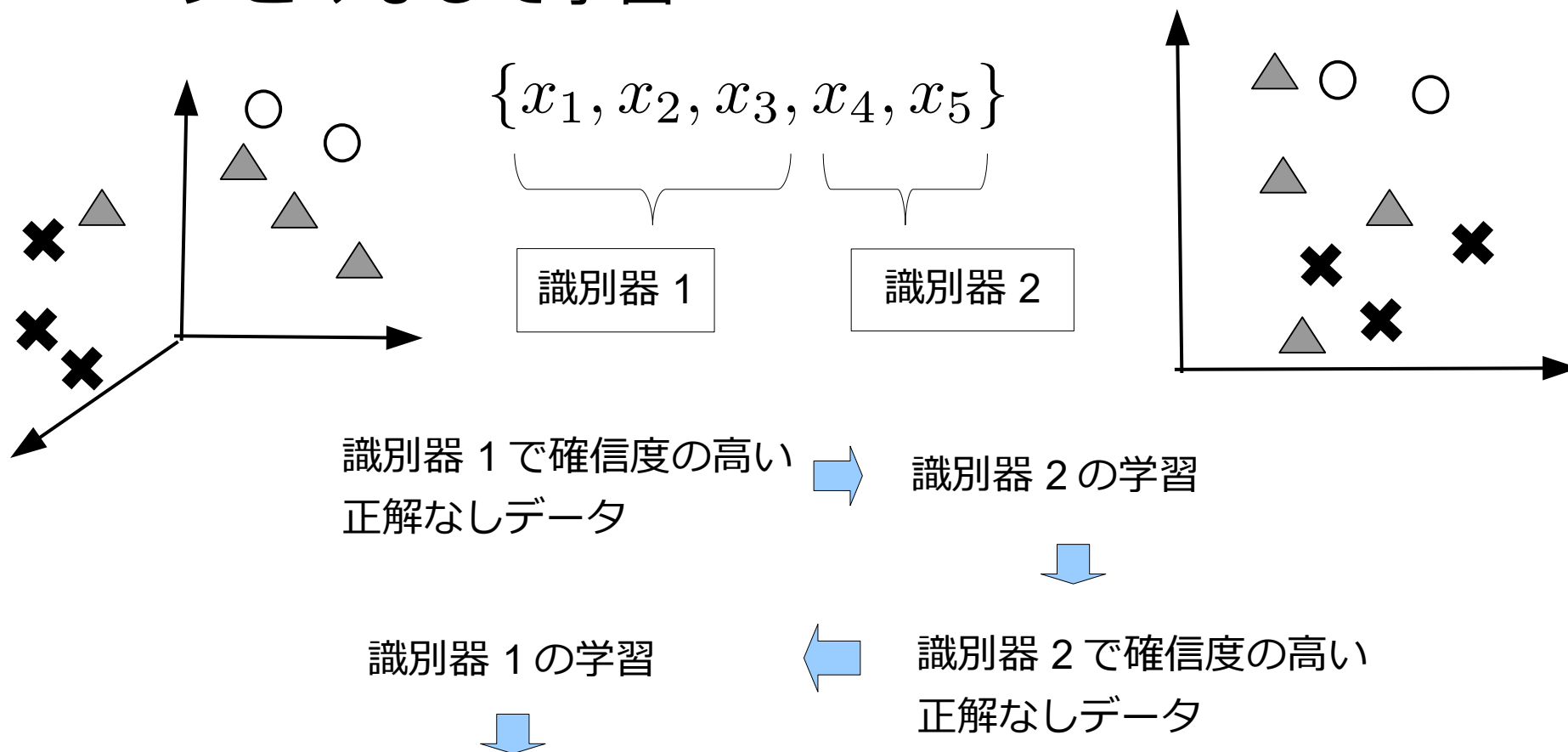
- 自己学習のアルゴリズム

1. 正解付きデータで初期識別器を作成
2. 正解なしデータの識別結果のうち、確信度の高いものを、正解付きデータとみなす
3. 新しい正解付きデータで、識別器を学習
4. 2, 3 を繰り返す



## 14.3 共訓練

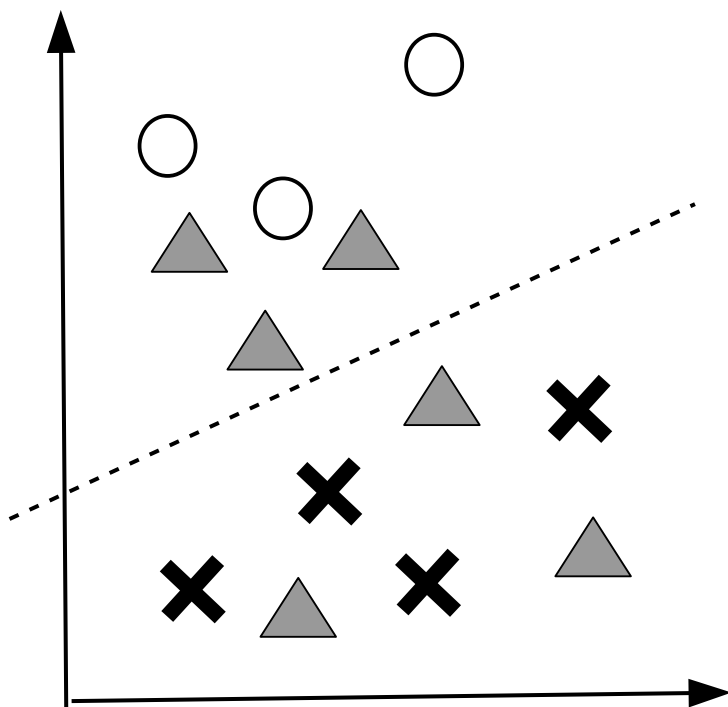
- 共訓練とは
  - 判断基準が異なる識別器を交互に用いる
  - 片方の確信度が高いデータを、相手が正解付きデータとみなして学習



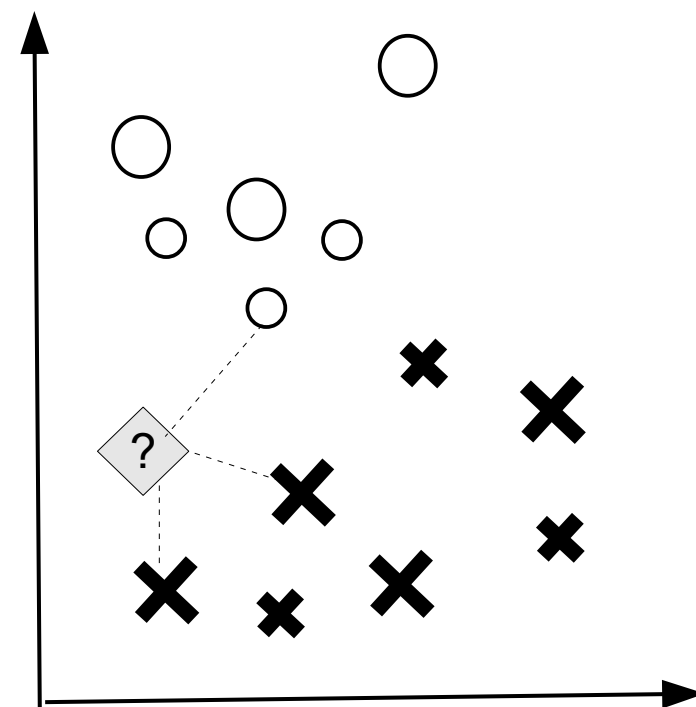
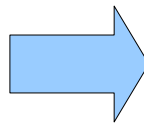
## 14.4 YATSI アルゴリズム

- YATSI(Yet Another Two-Stage Idea)  
アルゴリズムの考え方

- 繰り返し学習による誤りの増幅を避ける



正解付きデータで作った識別器  
で全データを識別



正解付きデータ :1  
識別後の正解なしデータ :0.1  
の重みで k-NN

調整可能

## 14.5 ラベル伝搬法

- ラベル伝搬法の考え方
  - 特徴空間上のデータをノードとみなし、類似度に基づいたグラフ構造を構築する
  - 近くのノードは同じクラスになりやすいという仮定で、正解なしデータの予測を行う
  - 評価関数（最小化）

$$J(\mathbf{f}) = \sum_{i=1}^l (y_i - f_i)^2 + \lambda \sum_{i < j} w_{ij} (f_i - f_j)^2$$

予測値と正解  
ラベルを近づける

隣接ノードの  
予測値を近づける

$f_i$  :  $i$  番目のノードの予測値

$y_i$  :  $i$  番目のノードの正解ラベル  $\{-1, 0, 1\}$

$w_{ij}$  :  $i$  番目のノードと  $j$  番目のノードの結合の有無

## 14.5 ラベル伝搬法

1. データ間の類似度に基づいて、データをノードとしたグラフを構築

- 類似度の基準

- RBF  $K(x, x') = \exp(-\gamma \|x - x'\|^2)$

- 全ノードが結合
    - 連続値の類似度が与えられる

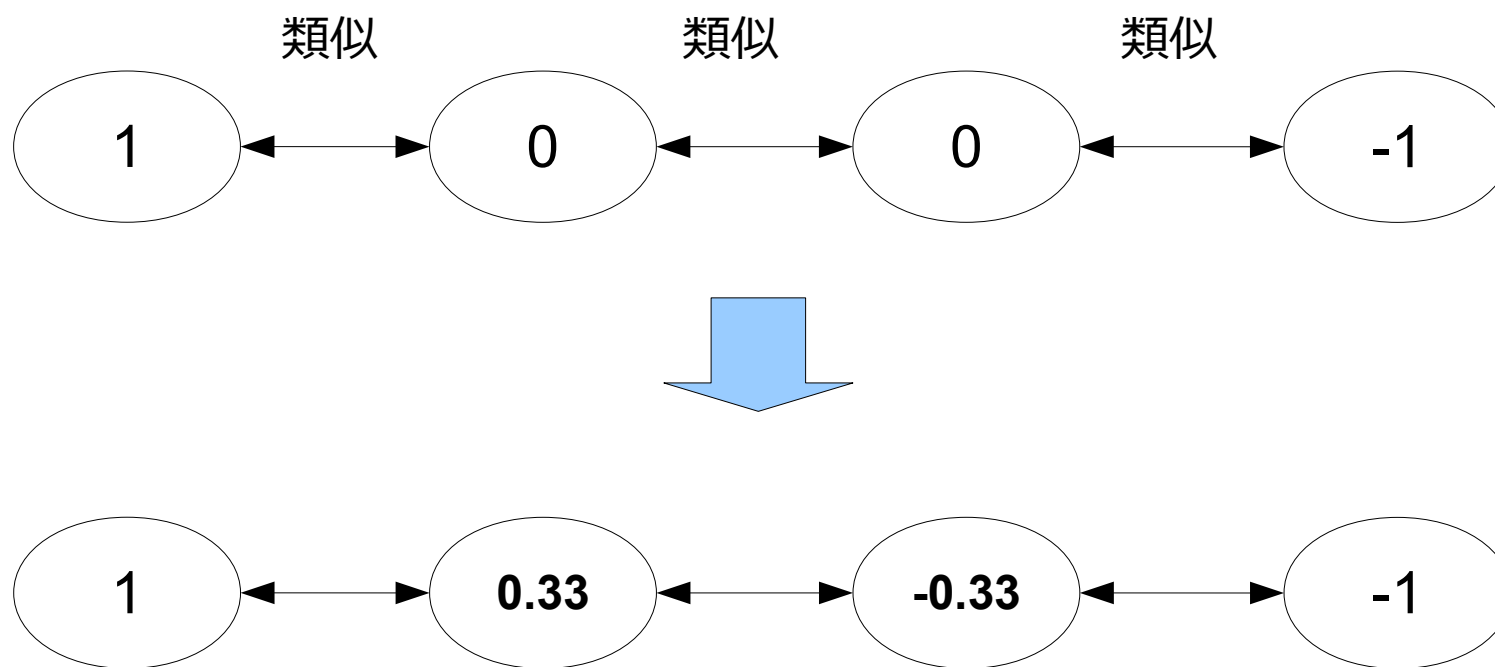
- K-NN

- 近傍の  $k$  個のノードが結合
    - 結合の有無は 0 または 1 で表現
    - 省メモリ



## 14.5 ラベル伝搬法

2. ラベル付きノードからラベルなしノードにラベルを伝播させる操作を繰り返し、隣接するノードがなるべく同じラベルを持つように最適化



# まとめ

- Weka での半教師あり学習
  - Collective パッケージのインストールが必要
  - 大半は 2 クラス問題にのみ適用可能
    - 例) diabetes データ
- ラベル系列に対する識別問題
  - 入力の系列長と出力の系列長との関係で手法を選択
    - CRF, HMM, RNN
- 半教師あり学習
  - 低密度分離など特定の条件を満たすデータには有効