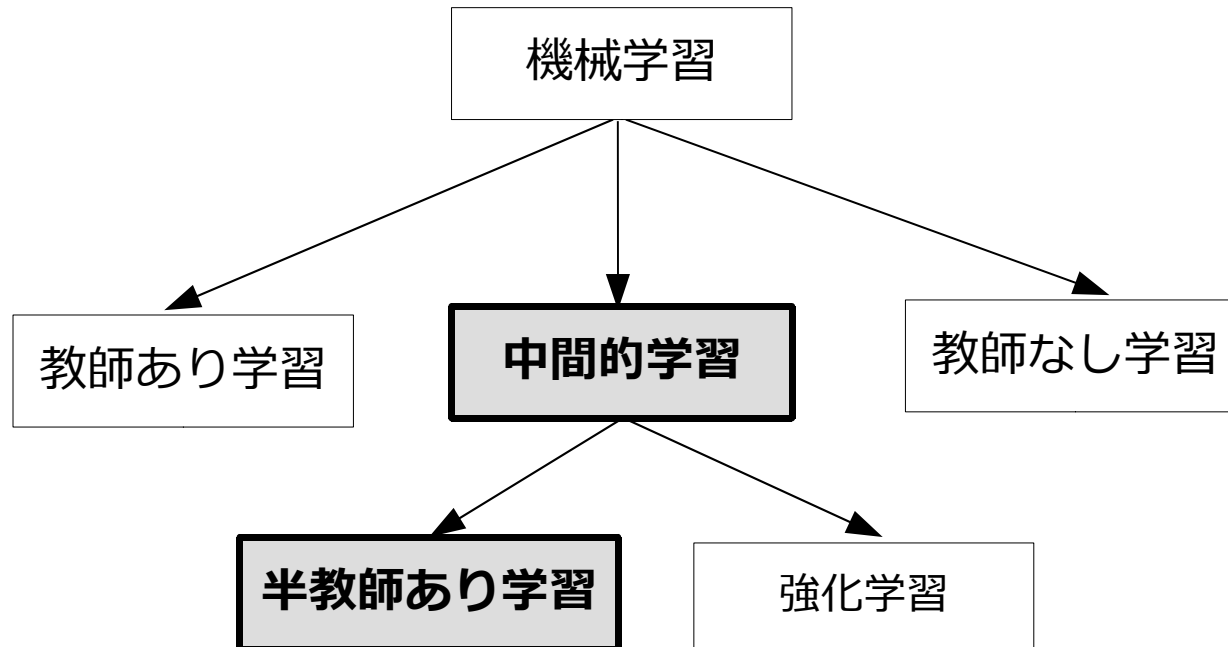


14. 半教師あり学習

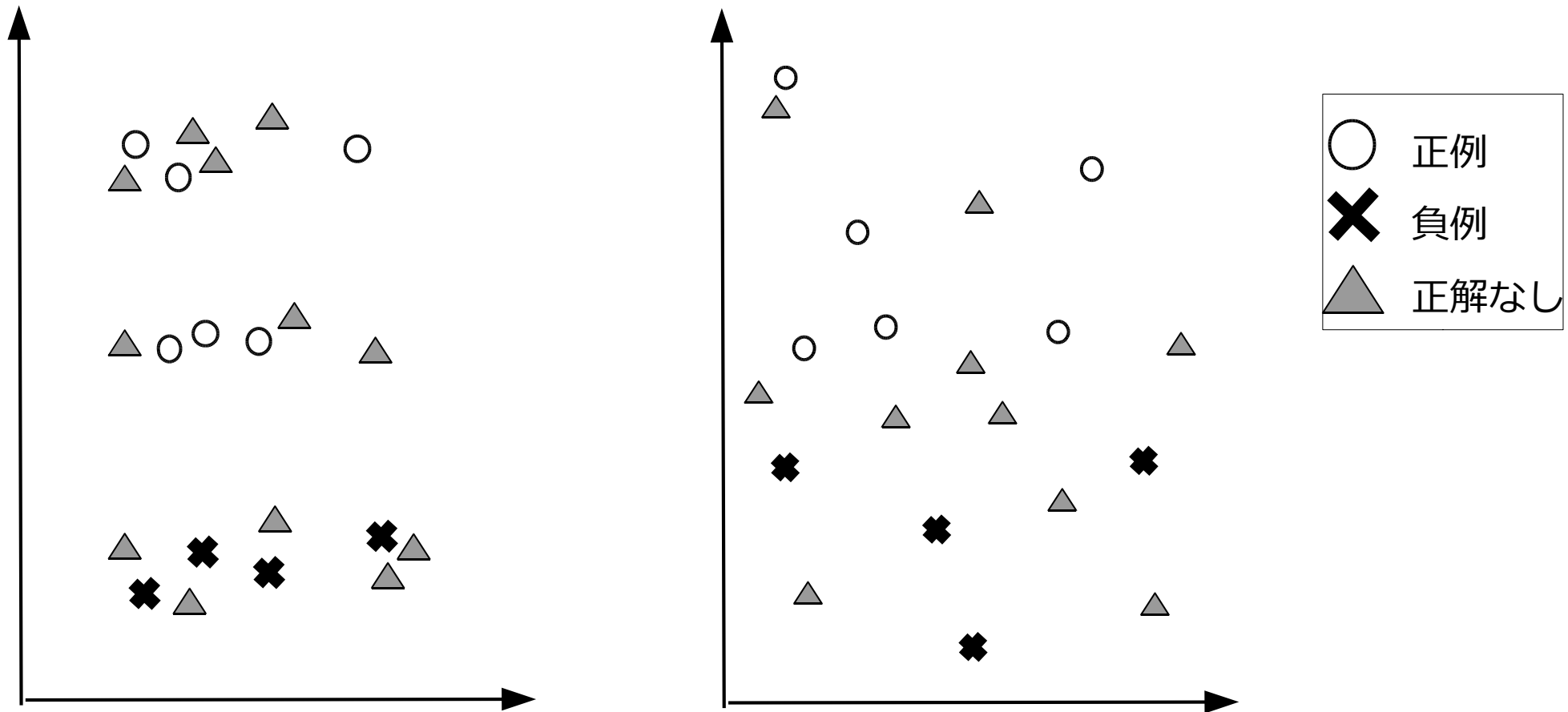
- 正解情報が学習データの一部にのみ与えられている状況での学習



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

?

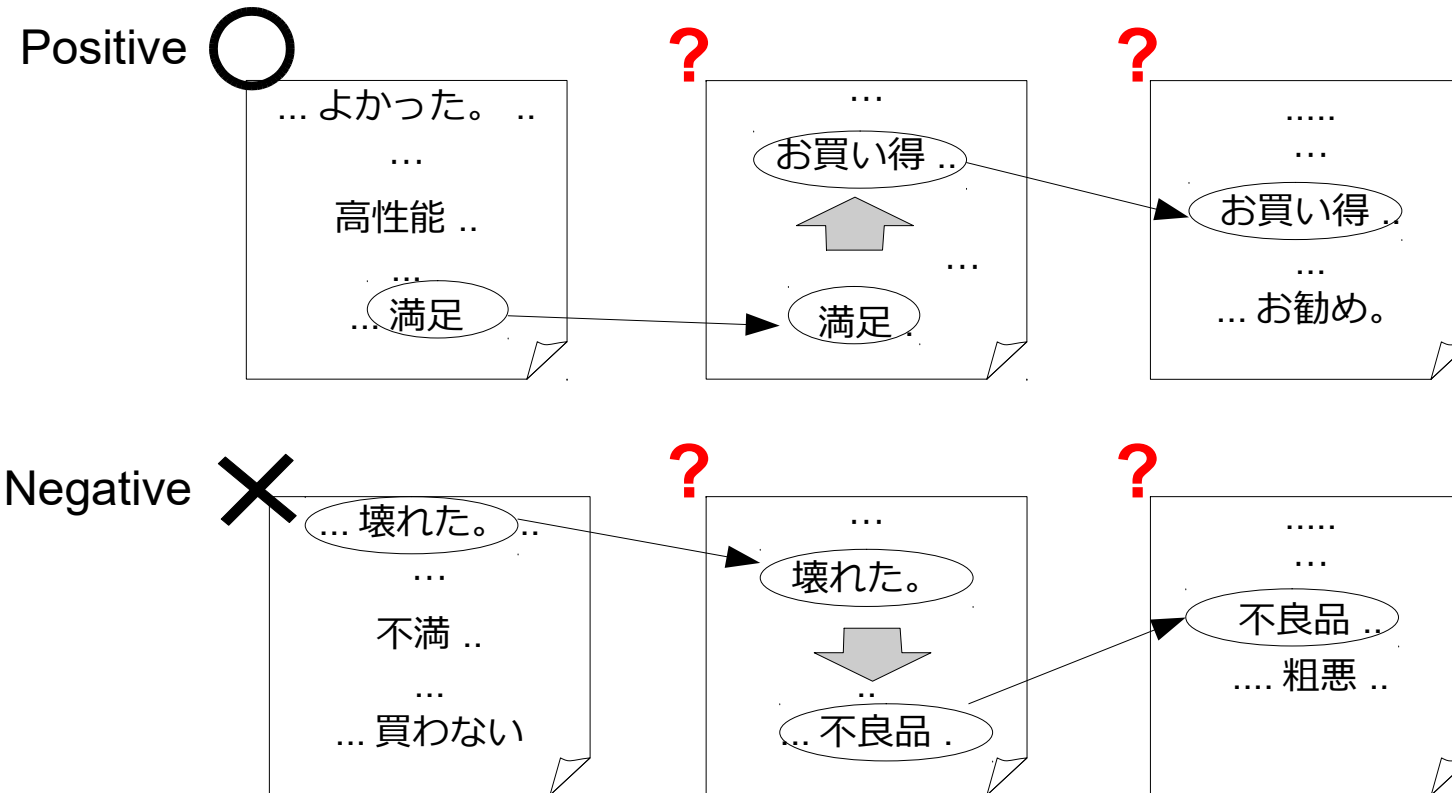
...
...
壊れた。 ..
... 買わない ..
....

?

.....
...
不満 ..
...
... 買わない

14.1.2 カテゴリ特徴の場合

- 特徴の伝播



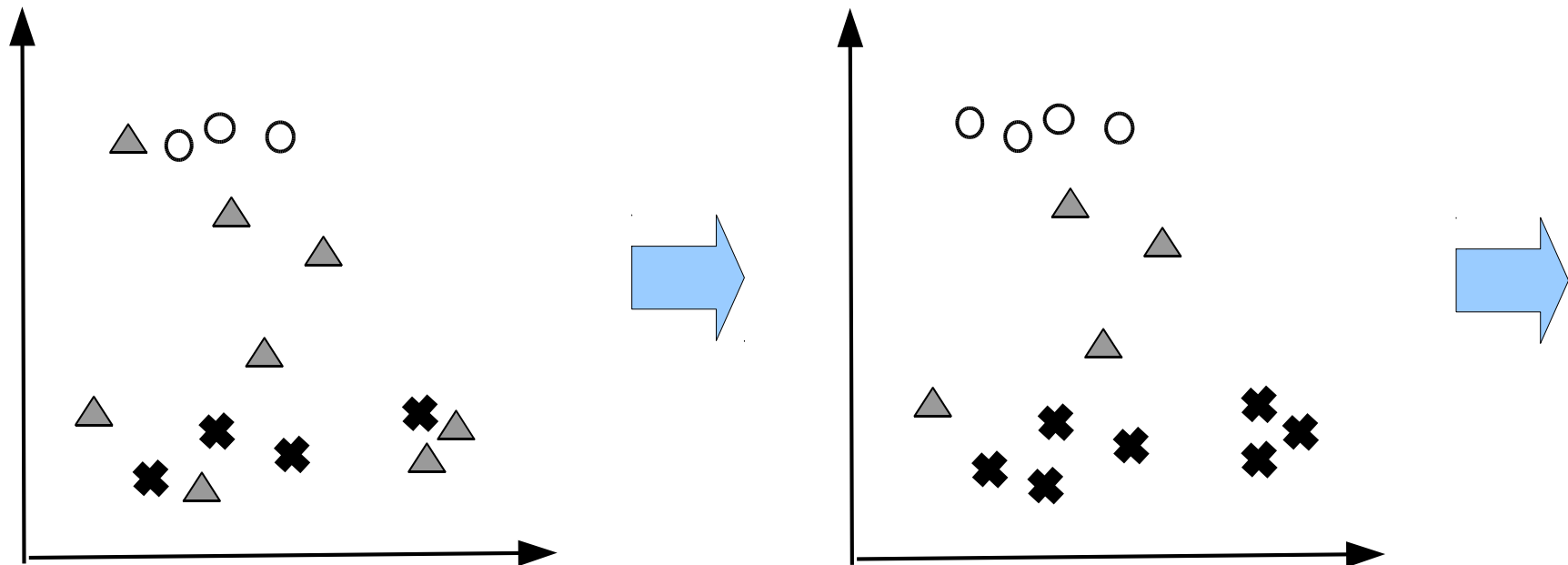
14.1.3 半教師あり学習のアルゴリズム

- 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

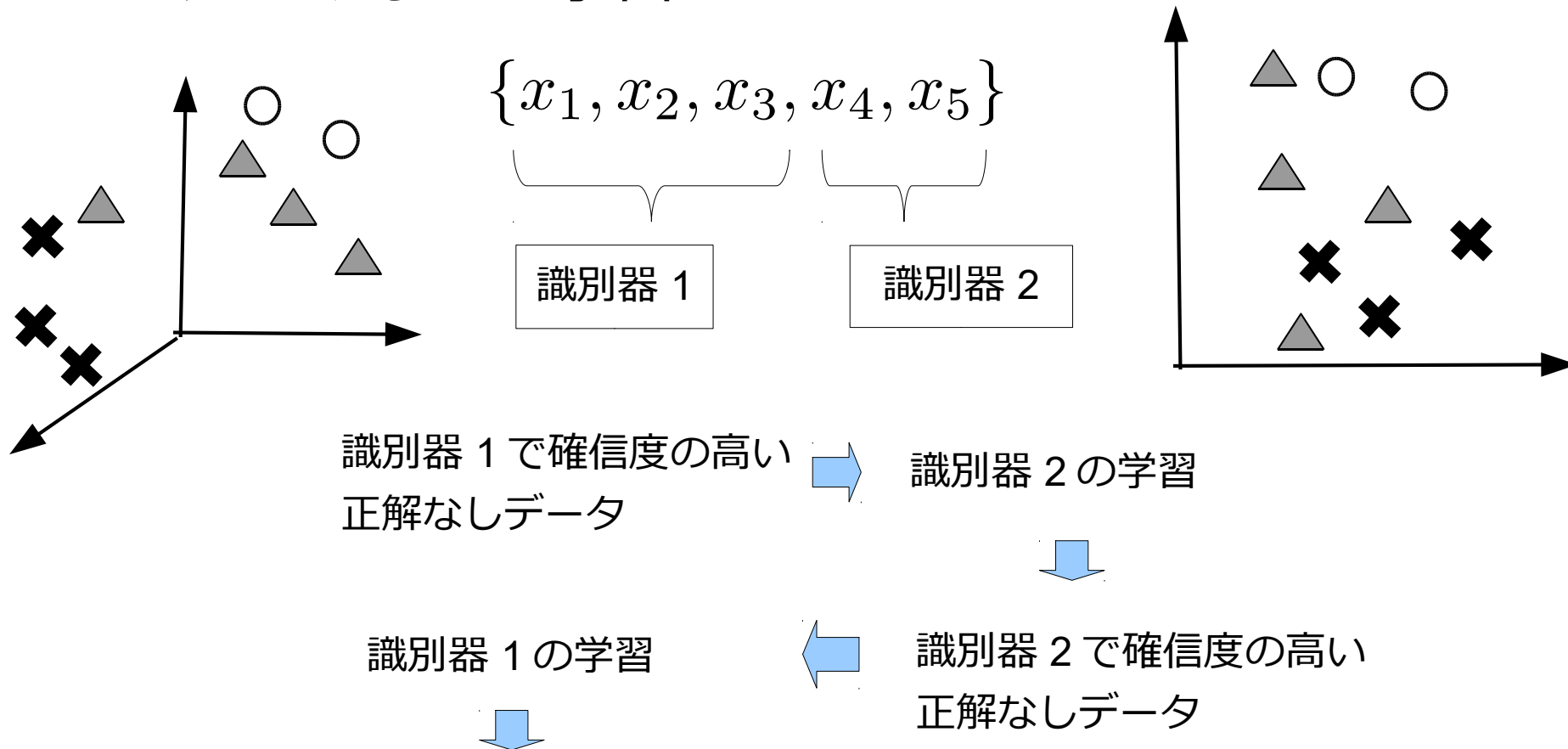


14.2 自己学習

- 自己学習の性質
 - クラスタ仮定や低密度分離が満たされるデータに対しては、高い性能が期待できる
 - 低密度分離が満たされていない場合、初期識別器の誤りが拡大してゆく可能性がある

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



14.3 共訓練

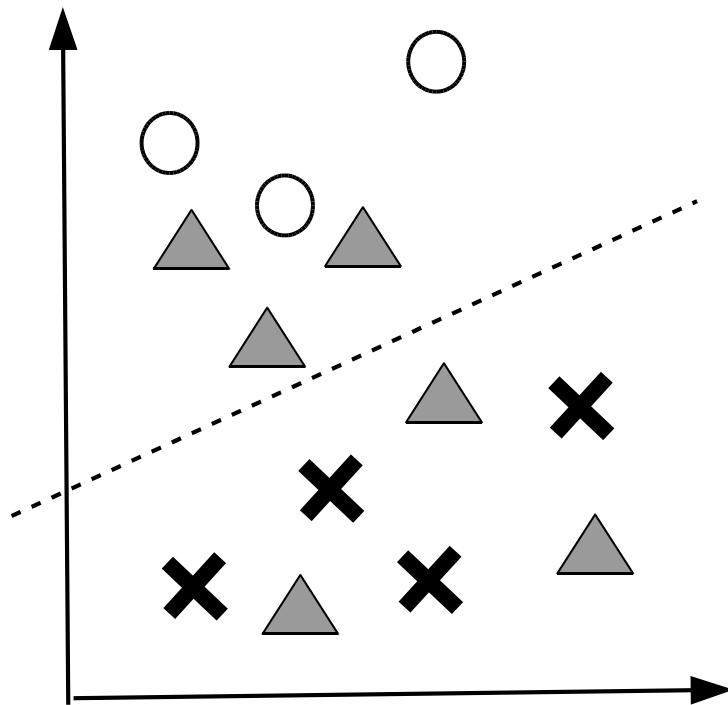
- 共訓練の特徴
 - 学習初期の誤りに対して頑健
- 共訓練の問題点
 - それぞれが識別空間として機能する特徴集合を、どのようにして作成するか
 - 全ての特徴を用いる識別器よりも高性能な識別器が作成できるか

14.4 YATSI アルゴリズム

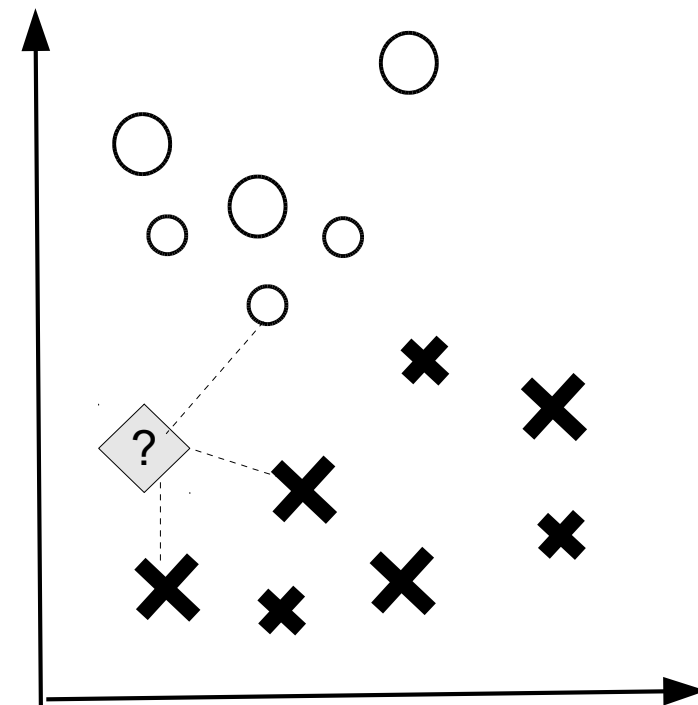
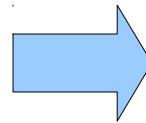
- YATSI(Yet Another Two-Stage Idea)

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 $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$

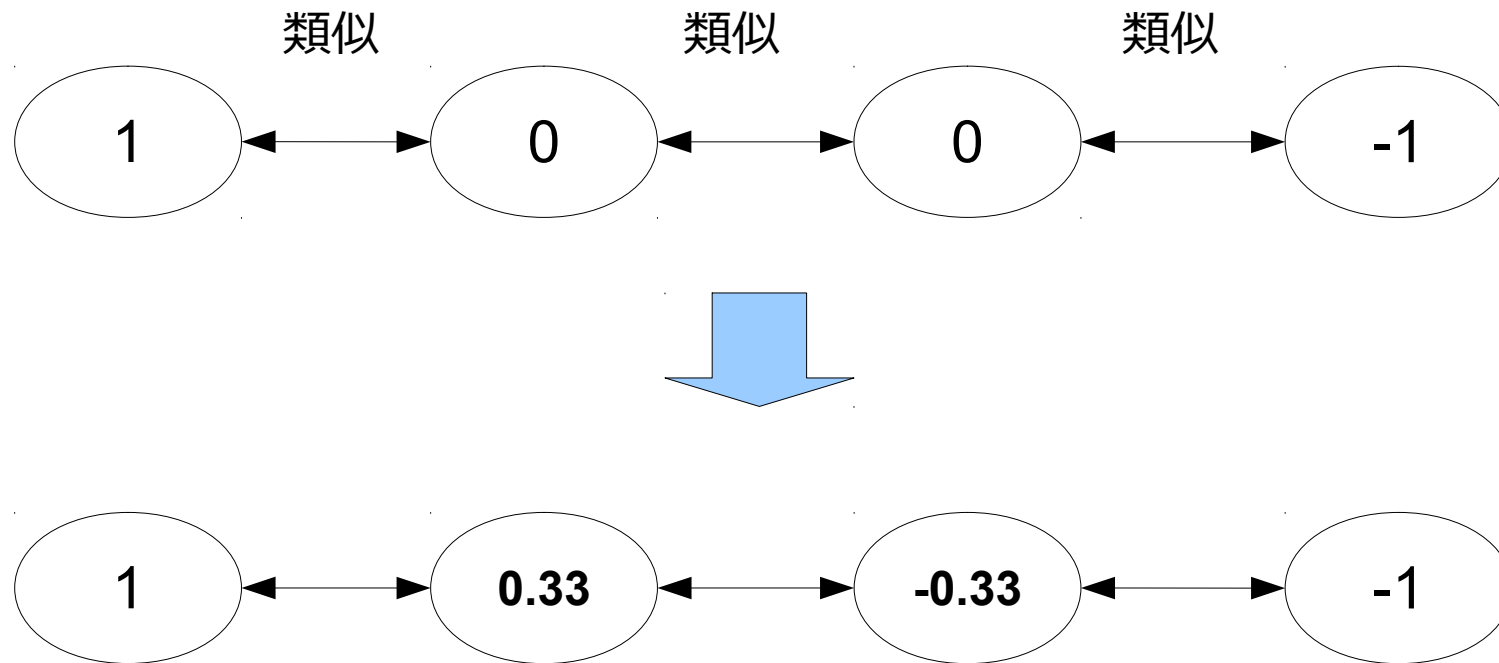
- 全ノードが結合
 - 連続値の類似度が与えられる

- 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



Multi-instance learning

- Multi-instance learning の問題設定

- 学習データ

- データの集まり (bag) に対して 1 つのラベルが付いている

- 例)

- A さん、B さんの Tweet : 面白い

- C さん、D さんの Tweet : 面白くない

全 Tweet がこのラベルに
当てはまるわけではない

- 問題

- 未知のデータの集まり (bag) (たとえば E さんの一定期間の Tweet) が与えられたとき、面白い or 面白くないを判定

Multi-instance learning

- 入力を集約する学習手法
 - bag の特徴ベクトルの集約情報（平均値・中央値・最大値・最小値など）を新たに特徴とする
 - 上記データに対して通常の教師あり学習
- 出力を集約する学習手法
 - bag の各データにその bag のクラスラベルを与えて、通常の教師あり学習
 - 判定したい bag に対して、個々のデータのクラスを判定し、多数決などの投票