RITSUMEIKAN

機械学習 第１４回 半教師あり学習

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義スケジュール

担当教員１：福森（第1回～第15回）

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	系列データの識別
14	**半教師あり学習**
15	強化学習

担当教員２：叶昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- 半教師あり学習　正解が一部の学習データにのみ与えられている状況での 学習方法
- 自己学習　正解付きデータで作成した識別器の出力の内、確信度の高い結果を信じて、そのデータを正解付きデータに取り込み、 自分を再度学習させることを繰り返す手法
- 共訓練（きょうくんれん）　異なった特徴を用いて識別器を2つ作成し、相手の識別 結果を利用して、それぞれの識別器を学習させる手法
- YATSIアルゴリズム　正解付きデータで一度だけ識別器を学習し、その識別器で全ての正解なしデータを識別し、その結果を 重み付きで利用してk - NN法による識別器を作る手法
- ラベル伝搬法（でんぱん）　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手法
- 演習問題

取り扱う問題の定義：半教師あり学習

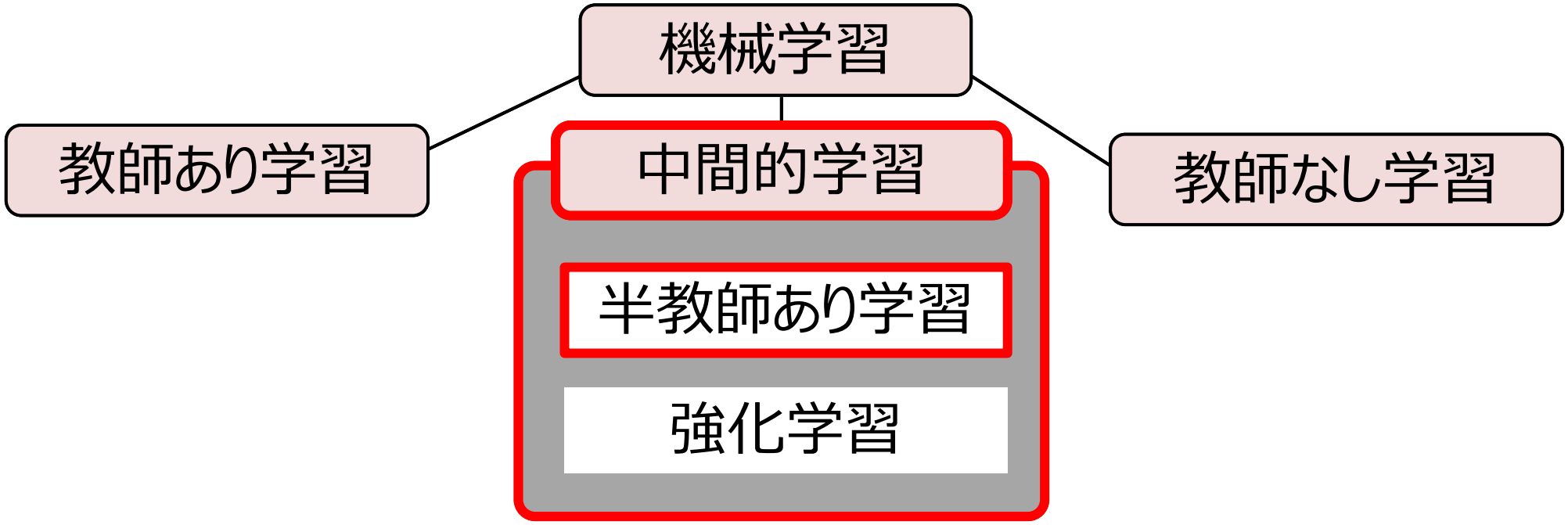
- **カテゴリ形式、または数値形式**の特徴ベクトルの**教師あり・教師なしデータが混在**する状況での識別学習

教師なしデータ　$\{x_i\}$　特徴ベクトル（カテゴリ or 数値）

教師ありデータ　$\{x_i, y_i\}$　正解情報（カテゴリ）

※ 半教師あり学習は、教師あり・教師なしデータが混在

$i = 1,2, \ldots, N$ ⟶ 学習データの総数



半教師あり学習とは

□ 半教師あり学習

- 正解が一部の学習データにのみ与えられている状況での学習方法
- 一般的に正解付きデータで識別器を作成して、正解なしデータを識別器の性能向上に役立てる

通过有正确答案的数据构建判别器，没有正确答案的数据提高判别器的性能

- 例：ツイートなどのweb文書を用いたある商品のPN判定

crawler program爬虫程序

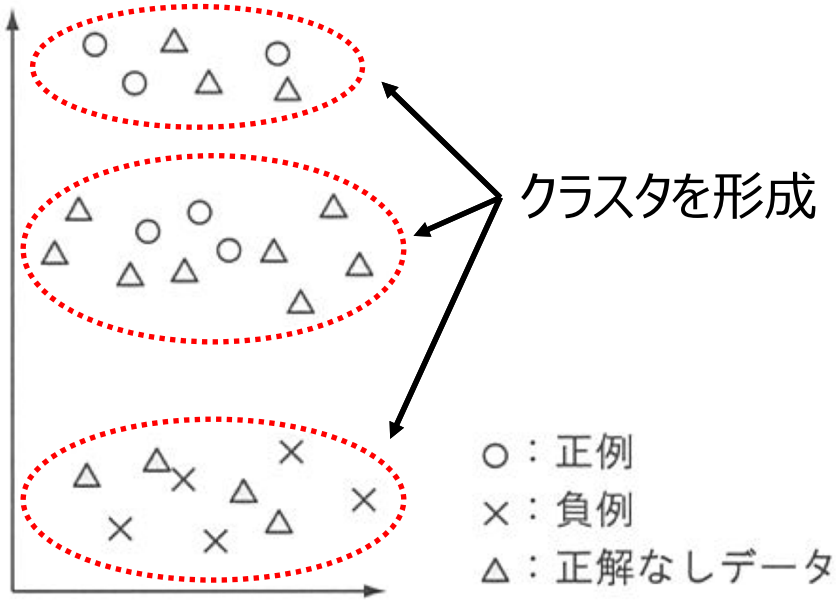
  - その商品名を含むweb文書をクローラプログラムで多数入手可能
    - それらに正解情報（肯定的[P]/否定的[N]）を付与するのは負担大
    - 現実的には、集めた文書の一部にしか正解を与えることができない
  - 「**少量**の正解付きデータ」と「**大量**の正解情報なしデータ」がある状況で識別器を構成

半教師あり学習とは

数値特徴の場合

半教師あり学習に**適する**データ 适合半监督学习的数据

- データがクラスタ(cluster)を形成しているとみなせる



半教師あり学習に**適する**データ

同一(どういつ)クラスタ内に異なるクラスのデータが混在していない状況

正解付きデータの分布から
正解なしデータの正解情報を
比較的、高精度に推測(すいそく)できる（だろう）

識別器の性能向上が期待できる

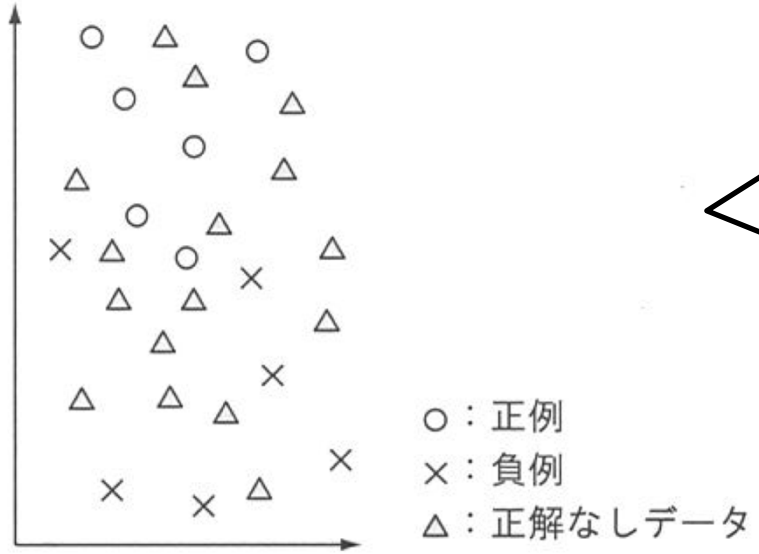
可以通过有正确答案的数据分布，按比例准确估计没有正确答案数据的正确信息

半教師あり学習とは

□ 数値特徴の場合（つづき）

■ 半教師あり学習に**適さない**データ 不适合半监督学习的数据

- どのデータに正解情報が付いているのかによって
  推定される識別境界の位置が大きく異なりそうなデータ
  根据具有正解信息的数据（的位置），预测的识别边界的位置存在很大不同



半教師あり学習に**適さない**データ

明確なクラスタが確認されない状況
正解なしデータが
間違ったクラスに分類される可能性がある

逆に識別器の性能を低下させる

没有正解的数据可能被分类到错误的类中，
导致判别器的性能下降。

半教師あり学習とは

□ 数値特徴の場合（つづき）

■ 半教師あり学習が可能なデータであるための仮定

1. **半教師あり平滑性仮定**
   - 2つの入力x_1とx_2が高密度領域で近ければ、出力y_1とy_2も関連している

     所属するクラスや正解情報の有無に関わらず
     一定の超立方体中のデータが多い領域
     ちょうりっぽうたい
2. **クラスタ仮定**
   - 同じクラスタに属する入力は同じクラスになりやすい
3. **低密度分離**
   - 識別境界は低密度領域にある 识别边界位于低密度区域
4. **多様性仮定** 高维数据可以映射在低维数据
   - 高次元データは、低次元の多様体（たようたい）上に写像できる
     - » 高次元でも「次元の呪い」にかかっていない

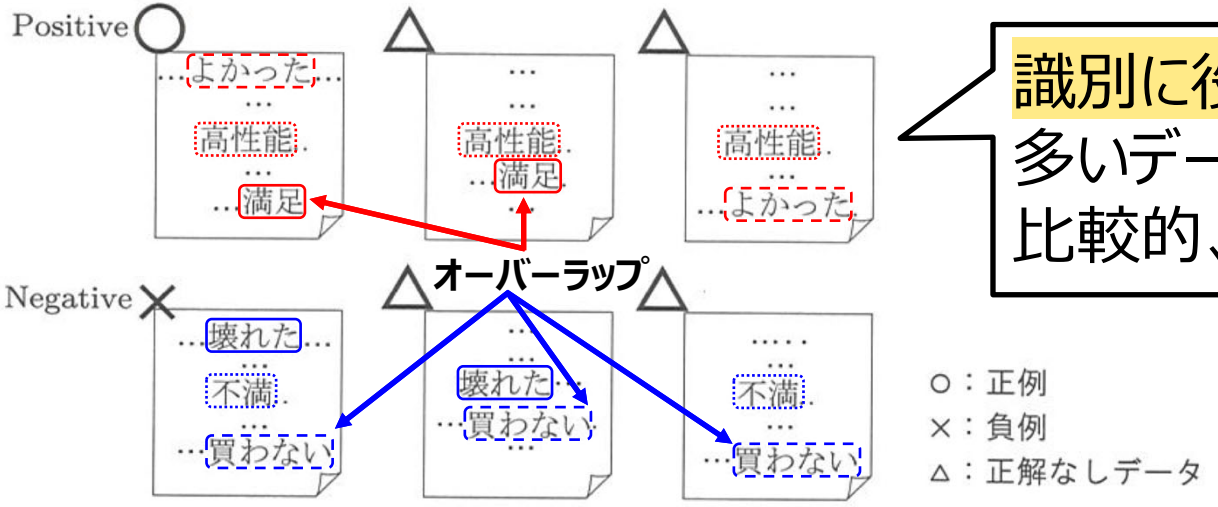
半教師あり学習とは

類別特征

□ カテゴリ特徴の場合

类别特征的学习数据能够大量输入的问题基本上集中于输入语言数据的方法

- カテゴリ特徴の学習データが大量に入手可能な問題はほぼ言語データを入力とするものに絞(しぼ)られる
- 半教師あり学習に**適する**データ
  - 正解付きデータで抽出された特徴語の多くが、正解なしデータにも含まれる

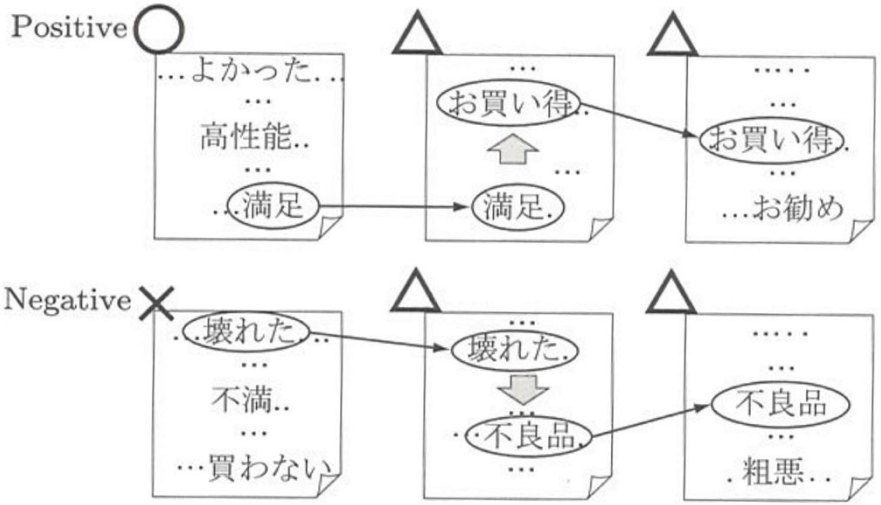


識別に役立つ特徴語のオーバーラップが多いデータは、正解なしデータに対しても比較的、高精度に正解情報を付けられる

半教師あり学習に適したカテゴリ特徴データ

半教師あり学習とは

- カテゴリ特徴の場合（つづき）
  - オーバーラップの伝播
    1. 正解付きデータに含まれる特徴語とオーバーラップが多い
       正解なしデータを新たに正解付きデータとする
    2. 新たに加わったデータに含まれる新たな特徴語を
       元の特徴語集合に加える



オーバーラップの伝播

正解なしデータの一部にオーバーラップがあれば
その一部の正解なしデータが、他の正解なし
データを徐々に巻き込んでいく可能性がある

自然言語で書かれたデータは、
この仮定を満たすことが多いので、半教師あり
学習は文書分類問題によく適用される

半教師あり学習とは

- 半教師あり学習のアルゴリズム
  - 基本的な考え方は、正解付きデータで作成した識別器のパラメータを、正解なしデータを用いて調整する
  - 識別器の作成アルゴリズム
    - これまで紹介してきた手法をタスクに応じて用いれば良い
    - 正解無しデータの識別結果を次回(じかい)の識別器作成に取(と)り込(こ)むには識別結果に確信度を伴うものが適する
  - 繰り返しアルゴリズムは、様々な設定が可能
    - 繰り返しを終了するための閾値をチェックする
    - 識別器のパラメータを繰り返しのたびに変化させる
    - 識別器で使う特徴に制限をかける　など

演習問題14-1（10分間）

- 人間の成長過程は、半教師あり学習に似ていると言われている。その理由を、実際の事例を挙(あ)げながら述べなさい。

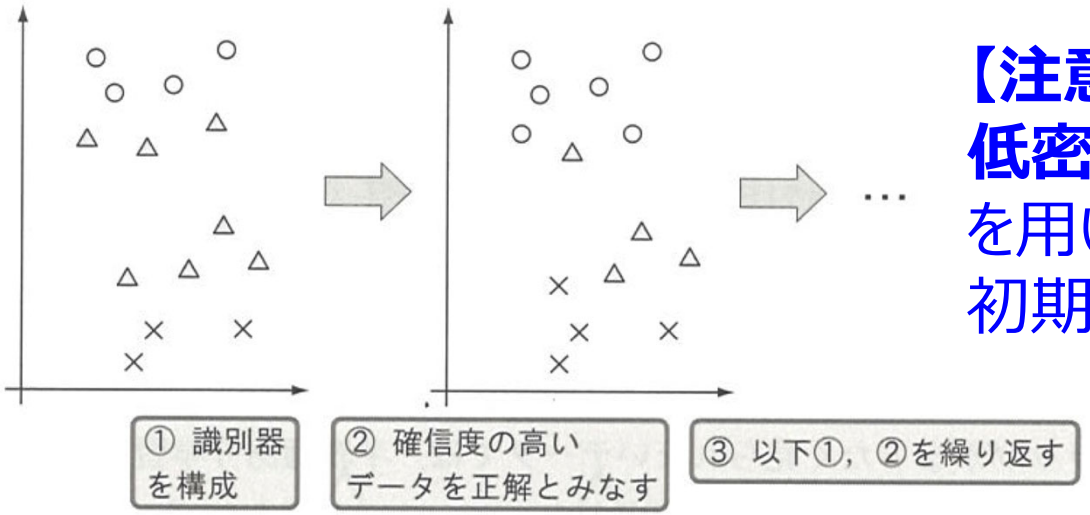
自己学習

□ 自己学習 (self-training)

在用有正确答案的数据所做的判别器的输出中，相信有高置信度的结果，并重复将数据纳入有正确答案的数据和重新训练自己的方法。

■ 正解付きデータで作成した識別器の出力の内、確信度の高い結果を信じて、そのデータを正解付きデータに取り込み、自分を再度学習させることを繰り返す手法

- 自分が出した結果を信じて、再度自分を学習させるという考え



【注意】
低密度分離が満たされていないデータ
を用いると、正解付きデータによって作成した初期識別器の誤りが、影響を及ぼし続ける

繰り返しによって、学習データが増加し、
より信頼性の高い識別器を作るのが狙い

共訓練

□ 共訓練 (co-training)

一种创建两个具有不同特征的分类器，并利用彼此的识别结果来训练每个分类器的方法。

- 異なった特徴を用いて識別器を2つ作成し、相手の識別結果を利用して、それぞれの識別器を学習させる手法
  - 判断基準が異なる識別器を2つ用意して、お互いが教え合うという考え方で半教師あり学習を実現する
    半監督(たが)学习是通过使用两个具有不同判断标准的分类器相互学习的思想来实现的。
- 利点
  - 学習初期の誤りに強(つよ)い 在学习的早期阶段对错误有很强的抵抗力
- 欠点(けってん)
  - それぞれが識別器として機能するような異なる特徴集合を見つけるのが難しい
  - 全ての特徴を用いた自己学習の方が性能が良い場合がある
    使用所有特征的自学习方法可能有更好的性能

共訓練

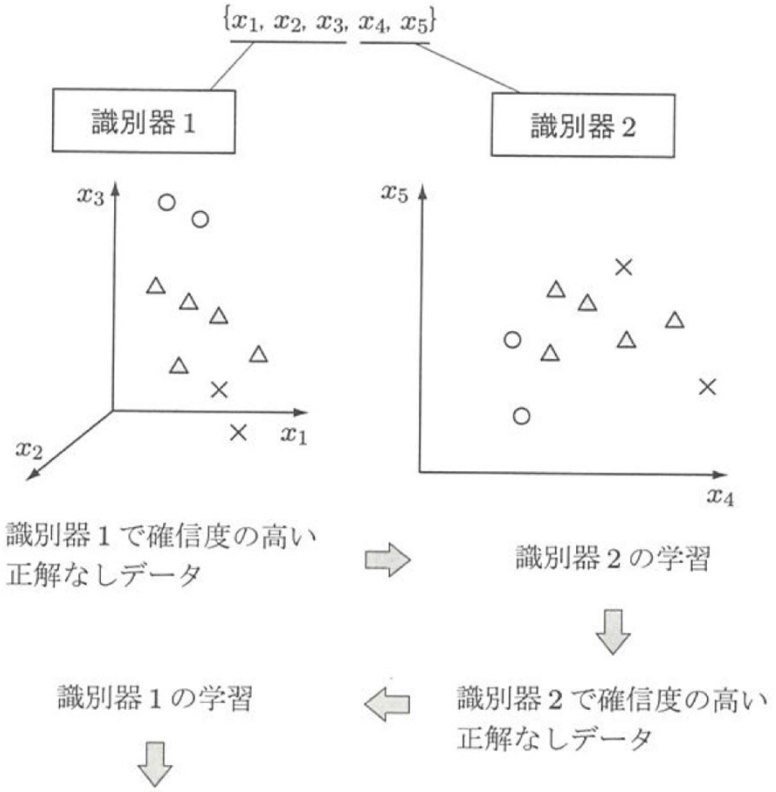
□ 共訓練 （つづき）

■ 共訓練のアルゴリズム

1. 正解付きデータの分割した特徴から識別器1と2を作成し、それぞれの識別器で正解無しデータを識別

2. 識別器1で確信度の高い上位k個のデータを正解付きデータとみなして識別器2を学習する

3. 識別器1と2の役割を入れ替えて精度の変化が小さくなるまで繰り返す



共訓練

YATSIアルゴリズム

YATSIアルゴリズム (Yet Another Two-Stage Idea)

- 正解付きデータで一度(いちど)だけ識別器を学習し、
  その識別器で全ての正解なしデータを識別し、その結果を
  重み付きで利用してk-NN法による識別器を作る手法
  - 数値特徴に対する半教師あり学習について
    自己学習や共訓練のような繰り返しアルゴリズムでの
    繰り返しによる誤りの増幅を避けることを狙う

    数值特征的半监督学习：旨在避免(避免)由于迭代算法（例如(例如)自学习和协同训练）
    中的重复而导致的错误放大
- YATSIアルゴリズムは多クラス分類にも適用できる
  - 自己学習や共訓練：基本的に2クラス分類問題が対象

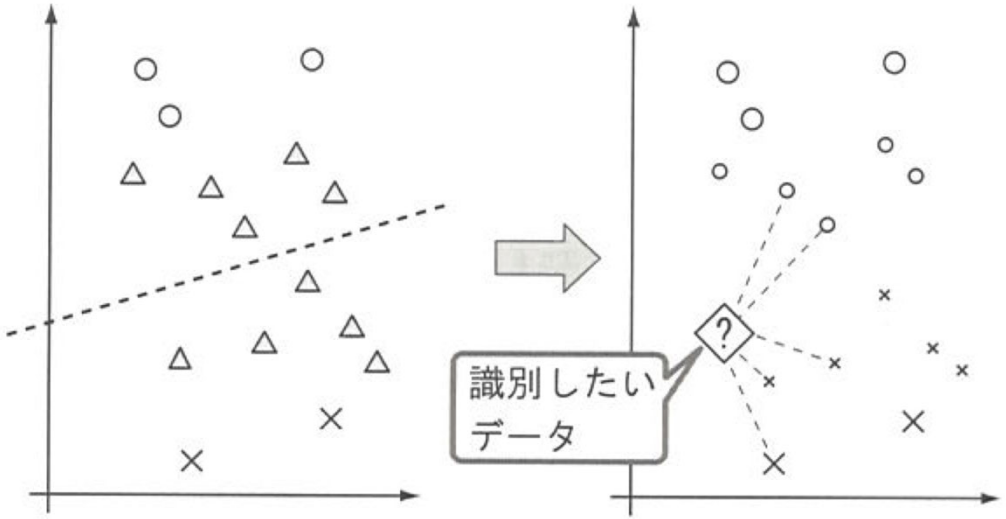
    自学习和协同训练 ：基本用于二分类问题

YATSIアルゴリズム

YATSIアルゴリズム（つづき）

識別器CをD_lで学習
識別器CでD_uを識別
D_lに重み1、D_uに重み $F \times \frac{|D_l|}{|D_u|}$ を付ける

D_l ：正解付きデータ
D_u ：正解なしデータ
F ：正解なしデータの寄与分
（どれだけ正解なしデータの識別結果を信用するか）



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 法

ラベル伝搬法

ロ ラベル伝搬法

- 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手法
  - 近くのノードは同じクラスになりやすいという仮定で
    正解なしデータを予測　假设附近的节点可能属于同一类，预测无标签的数据
- ラベル伝搬法は、評価関数 $J(\boldsymbol{f})$ を最小化する

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

  - y_i ：i番目のノードの正解情報
    - 正例：1、負例：-1、正解なし：0 のいずれか
  - f_i ：i番目のノードの予測値（1または-1のいずれか）
  - w_{ij} ：i番目のノードとj番目のノードの結合度

ラベル伝搬法

ラベル伝搬法の学習手順

1. データ間の類似度に基づいたグラフ構築 基于类间的相似度的图的构建
   - データ間の基準は「ガウシアンカーネル」や「k-NN法」が使われる
     - ガウシアンカーネル（Gaussian kernel）：$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$
       - » 全ノードが結合し、連続値の結合度が与えられる 所有节点都被加入并赋予一个连续的耦合值
     - K-NN法
       - » 近傍のk個のノードのみが結合する省（しょう）メモリの手法で結合度は0または1で表現される 一种只连接附近k个节点的内存节省方法，耦合度表示为0或1。
2. 評価関数の最小化 重复从标记节点向未标记节点传播标签的操作，通过最小化评价函数，优化相邻节点尽可能具有相同的标签。
   - ラベル付きノードからラベルなしノードにラベルを伝搬させる操作を繰り返し、評価関数$J(\boldsymbol{f})$の最小化を通じて、隣接するノードがなるべく同じラベルをもつように最適化

演習問題14-2（10分間）

- 半教師あり学習を活用できる場面（ばめん）を考えなさい