Uczenie się częściowo nadzorowane

Semi-supervised learning

Część 1 – wprowadzenie, modele generatywne, self learning

Wykład spec. projekt eksploracji danych

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska
Poznań

Poznań, 2020

Plan wykładu

Dzisiaj cz. 1

1. Motywacje do uczenia częściowo nadzorowanego (SSL)
2. Kiedy SSL może pomóc
3. Wybrane metody
   - Cluster-and-label
   - Metody generatywne (np. mieszaniny gausowskie)
   - Self training
   - Co-training and multi-view learning
   - Podejścia grafowe z propagacją etykiet
   - Rozszerzanie SVM
   - Inne
4. Inne zagadnienia i wyzwania

Różne podejścia do uczenia się z przykładów

- Uczenie w pełni nadzorowane (supervised) – etykietowane przykłady

$$L = \left\{ (x_i, y_i) \right\}_{i=1}^{n}$$

- Uczenie nienadzorowane (unsupervised) – nieetykietowane

$$U = \left\{ (x_i) \right\}_{i=1}^{m}$$

- Uczenie częściowo etykietowane (semi-supervised)

  Etykietowane dane L oraz (część) nieetykietowanych U, na ogół $m>>n$

  Nie możemy pozyskać etykiet dla choć wybranych przykładów z U

  Cel: skonstruowanie lepszego klasyfikatora z L i U niż z samych przykładów etykietowanych L

Różne podejścia do uczenia się z przykładów

Motywacje

- Uczenie nadzorowane wymaga wielu etykietowanych przykładów
  - Etykietowanie = kosztowne, czasochłonne i męczące, nierealne dla wielkich kolekcji przykładów
  - Ograniczony dostęp do osób adnotujących
  - Wymóg wiarygodności (ekspert lepszy niż ochotnicy; niektóre dane trudne do etykietyzacji, obciążenie – bias osób)
- Nieetykietowane przykłady
  - Potencjalnie łatwo dostępne w dużej ilości
  - Często możliwe do automatycznej rejestracji, np.
    - Web, pomiary sensorów, pozyskiwane z repozytoriów danych
  - „Tańsze” w przygotowaniu

Przykłady *L* i *U*

Strony WWW, obrazy, dokumenty

Etykietowanie:

- Konieczność czytania lub oglądu
- Wymaga namysłu
- Czasami wspomagane dodatkowymi źródłami

Nieetykietowane:

Potencjalne dostępne w dużej ilości automatycznie pozyskiwane z minimalnym kosztem

Dane bio-medyczne

np. predykcja struktur genetycznych

Adnotacja wymaga wysiłku eksperta, dodatkowej wiedzy, niekiedy potwierdzenia (po długim czasie)

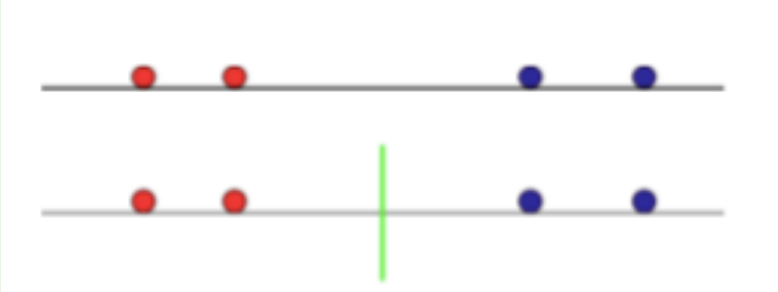
Nowe urządzenia diagn. (np. sekwencjonowanie DNA) może generować dużo danych eksperymentalnych

Związek z ludzkim poznawaniem świata

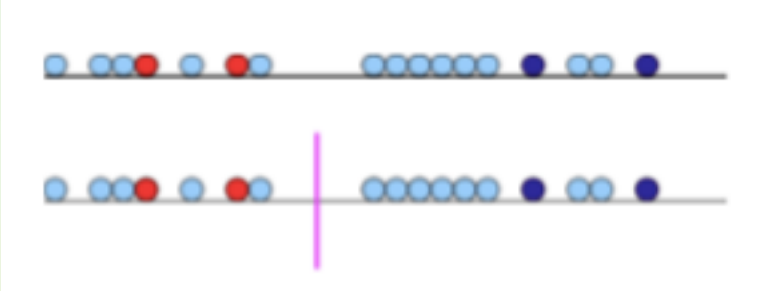
- Ludzie często poznają obiekty ze świata z ograniczonym wskazywaniem kategorii, a wiele sami klasyfikując, np.
- Przykład uczenia dzieci pojęć
  - X=zwierze (cechy) y = gatunek, np. pies
  - Nauczyciel-rodzic wskazuje i wyjaśnia “to jest pies”
  - Dzieci obserwuja inne zwierzęta (nie objaśnine) i samodzielnie klasyfikują

Kiedy SSL może pomóc?

Etykiety : Czerwone klasa 1, niebieskie klasa -1



Dodajmy rozkład przykładów nieetykietowanych (szare)

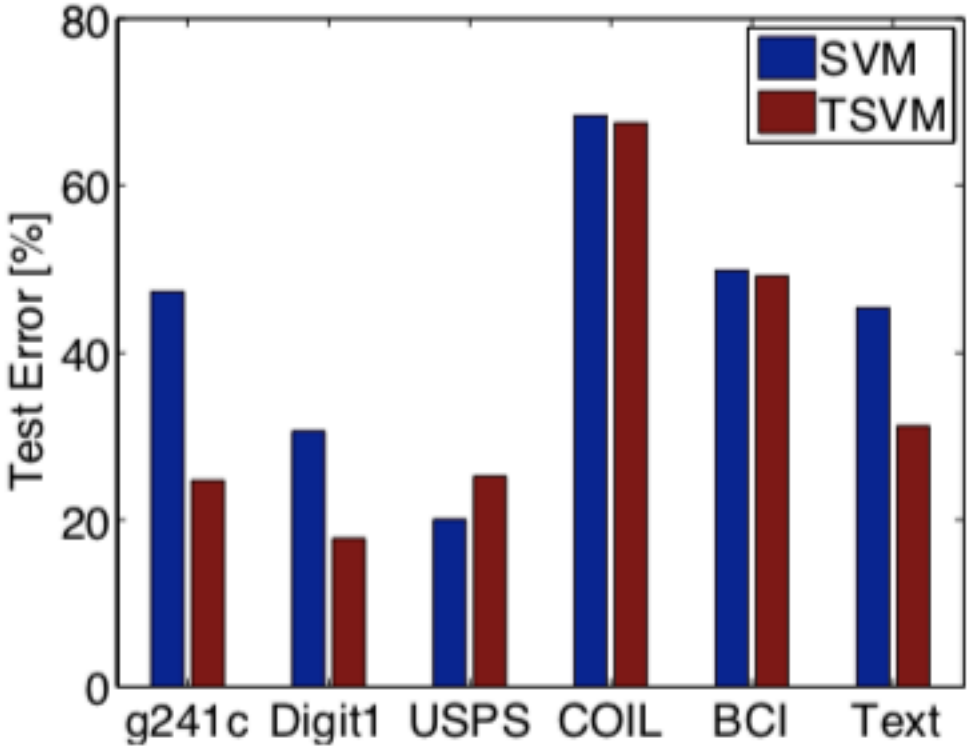


Zmiana granicy decyzyjnej na dokladmiejszą!

Założenie: przyklady z różnych klas tworzą spójne rozkłady

Za wykładem Zhu

Inny przykład użycia Transductive SVM



10 labeled points
~1400 unlabeled points

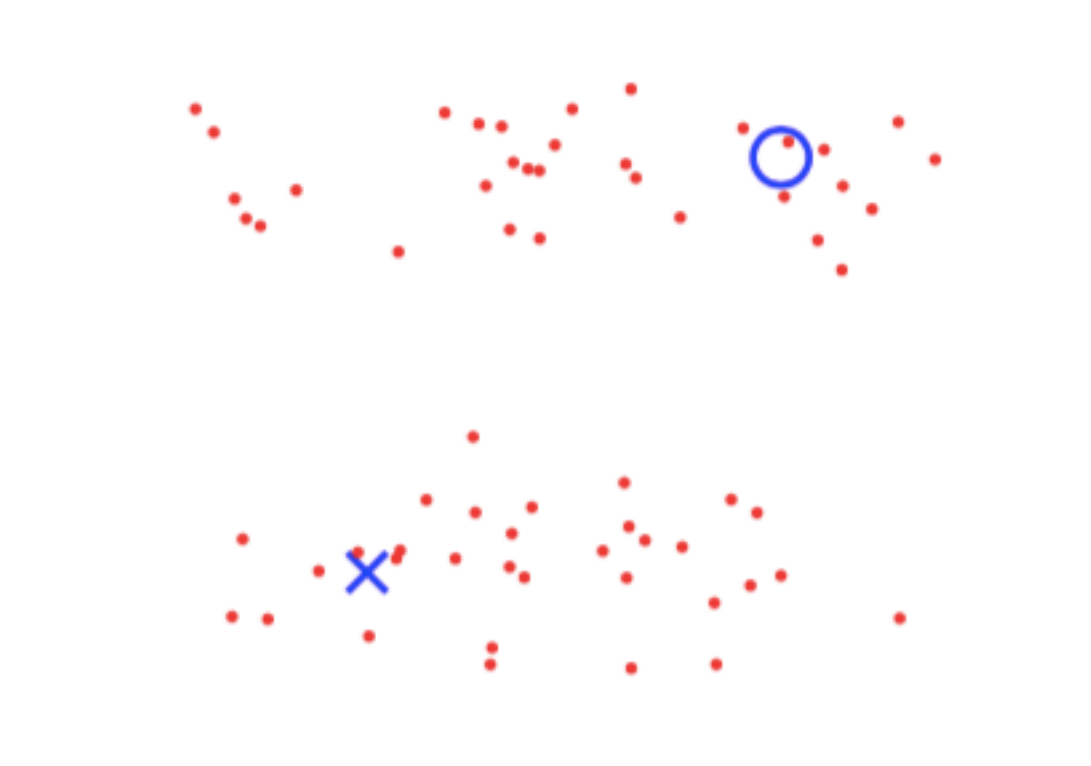
SVM: supervised
TSVM: semi-supervised

Założenia bliskości punktów

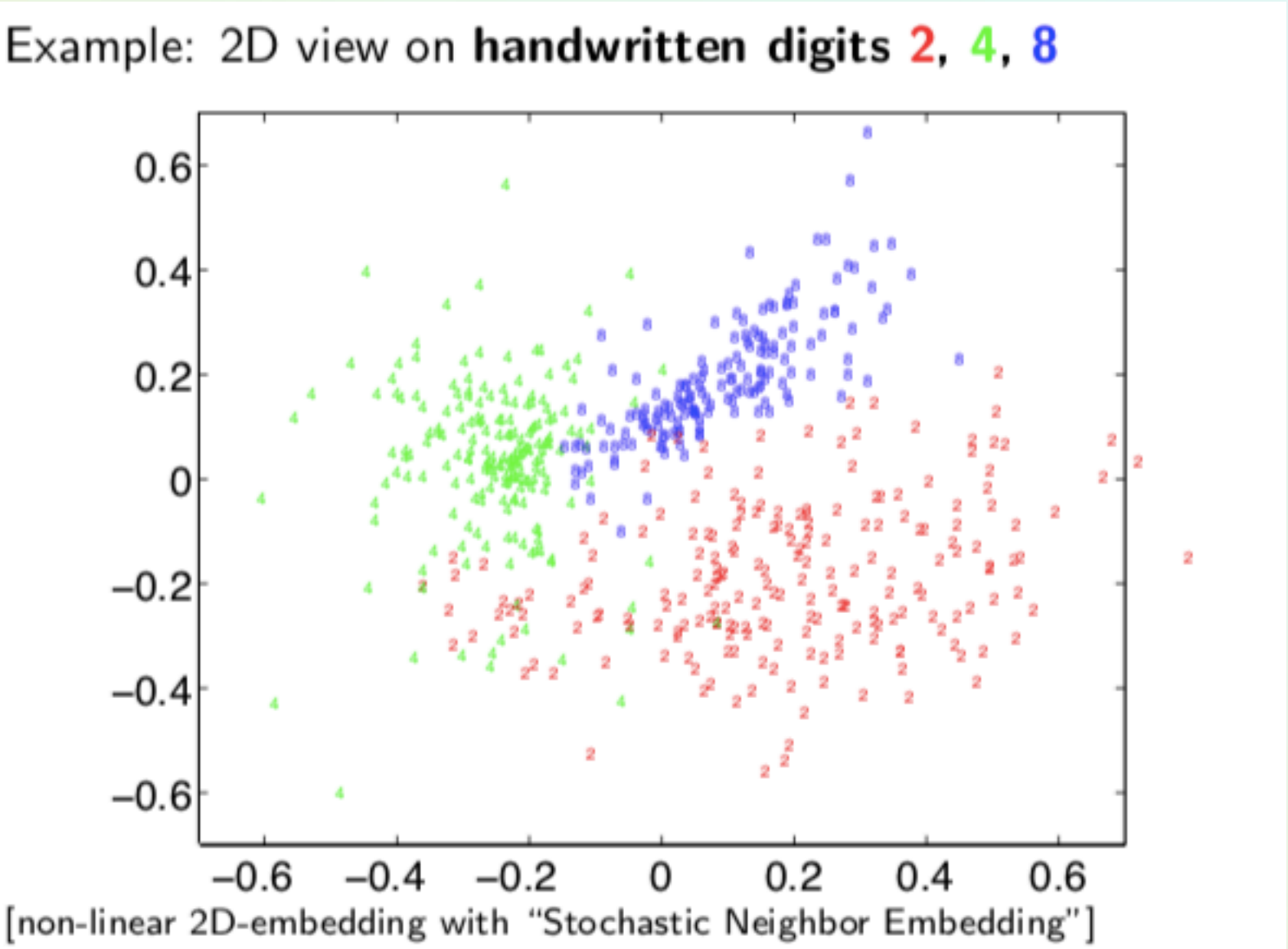
Punkty położone blisko siebie mają podobne etykiety

Założenie nt. skupień w danych

- Dane naturalnie tworzą skupiska
- Punkty w jednym skupisku powinny mieć tą samą etykietę



Przykład rozpoznawania liter



Cluster and label

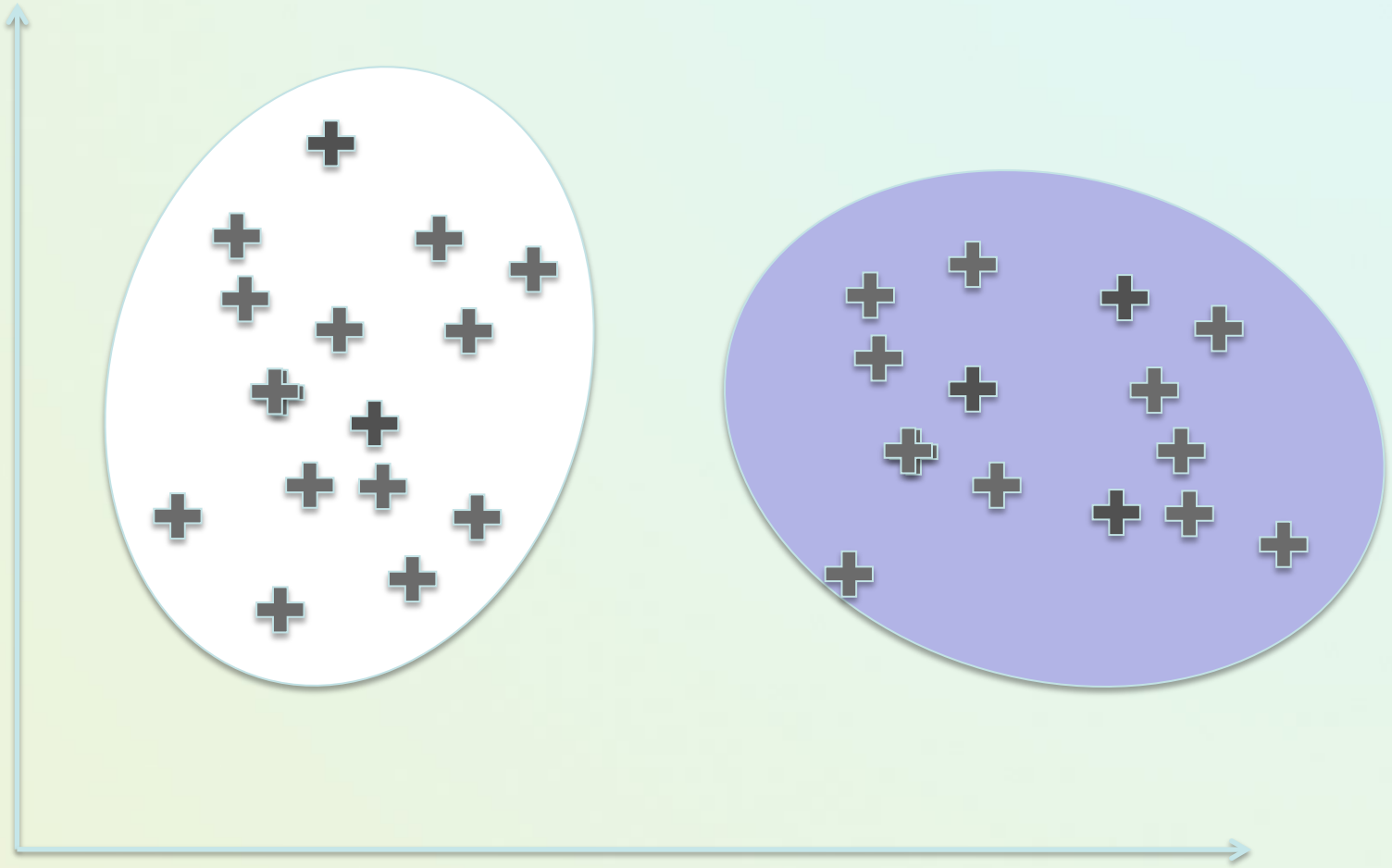
Input: etykietowane przykłady L={x,y} /n/ oraz nieetykietowane U={x} /m/ oraz algorytmy A grupowania i uczenia klasyfikatora K

1. Znajdź skupienia w zbiorze nienadzorowanych x (m+n)
2. Dla każdego ze skupień, znajdź zbiór wszystkich etykietowanych przykładów S
3. Naucz algorytmem K klasyfikator f na podstawie zbioru S
4. Zastosuj f do predykcji etykiet na pozostałych nieetykietowanych przykładów w skupieniu

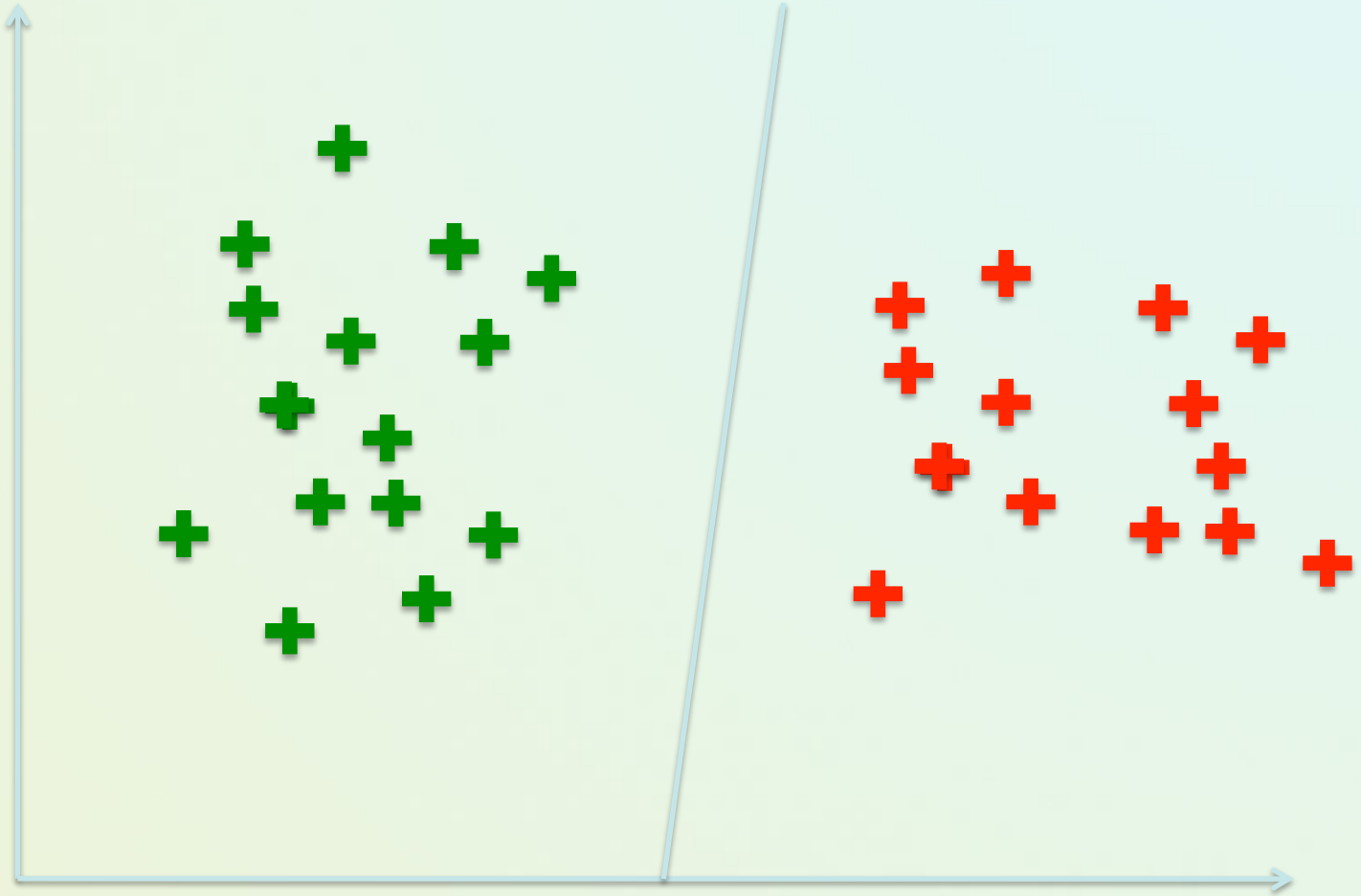
Na koniec nauczyć klasyfikator na całym zbiorze przykładów (obejmującym zaetykietowane przykłady ze skupisk) -> **output**

Obserwacja: zakłada się, że granice skupisk są zgodne z granicami decyzyjnymi

SSL – grupowanie = założenia nt. rozkładów

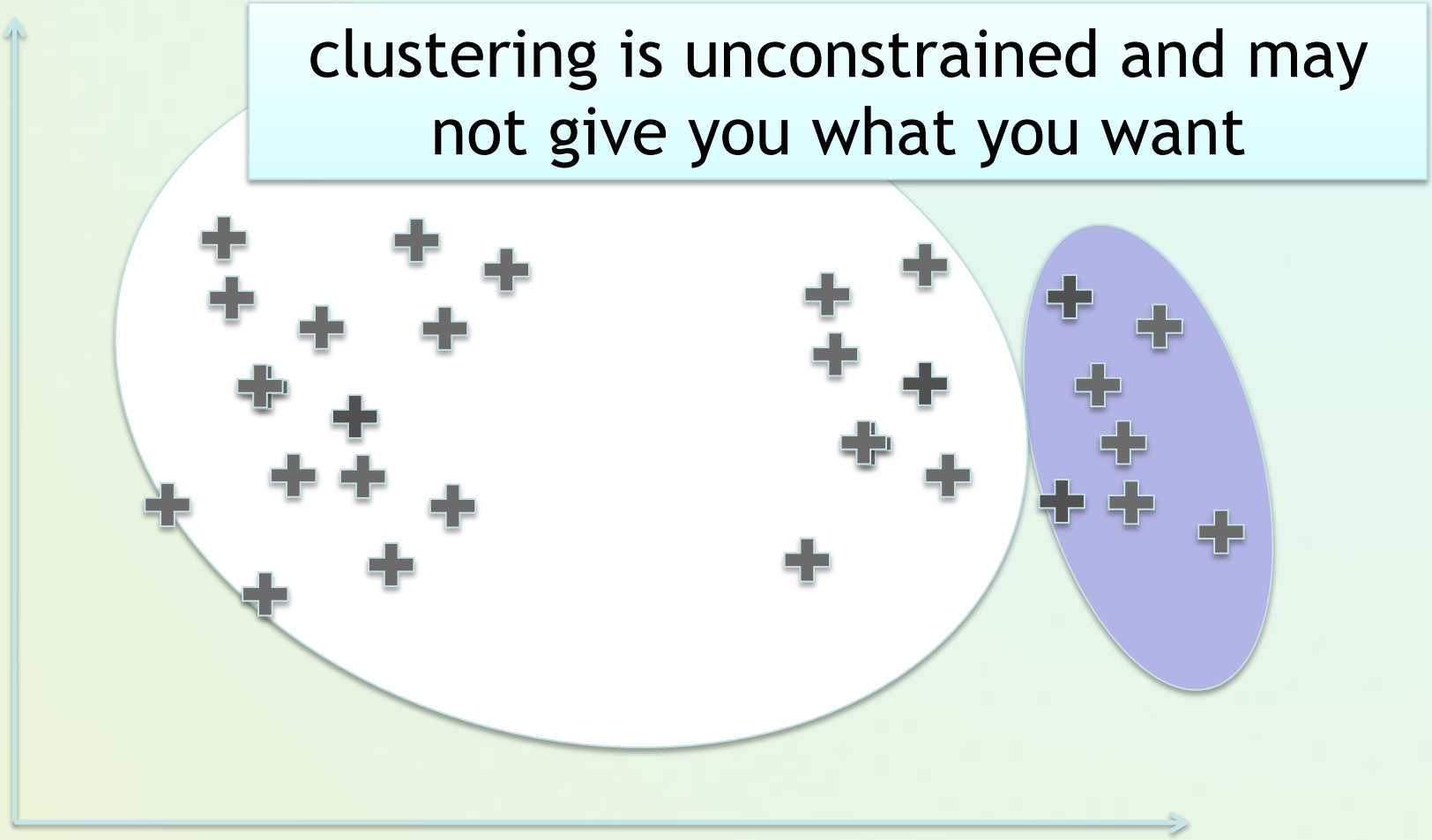


SSL is Between Clustering and SL



Za

SSL is Between Clustering and SL



Dobór algorytmu grupowania?

Modele generatywne

Modelują rozkład prawdopodobieństwa p(xy)

“Generatywne” – mogą poprzez próbkowanie pozyskiwać dodatkowe punkty lub generować sztuczne

Przykłady:

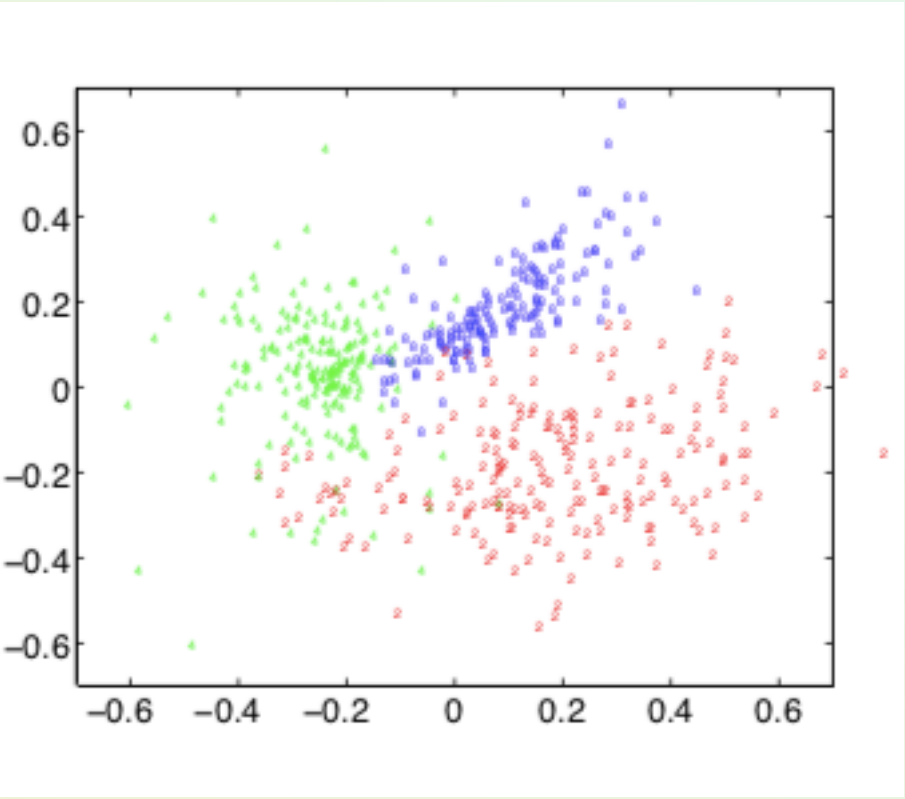
Podejścia Gausowskie, odmiany Naïve Bayes, mieszaniny rozkładów, ukryte modele Markova (HMM), sieci Bayesowskie, “Markov random fields”

Wykorzystanie do analizy przykladów L i U – modyfikacje podejść mieszanin Gausowskich EM

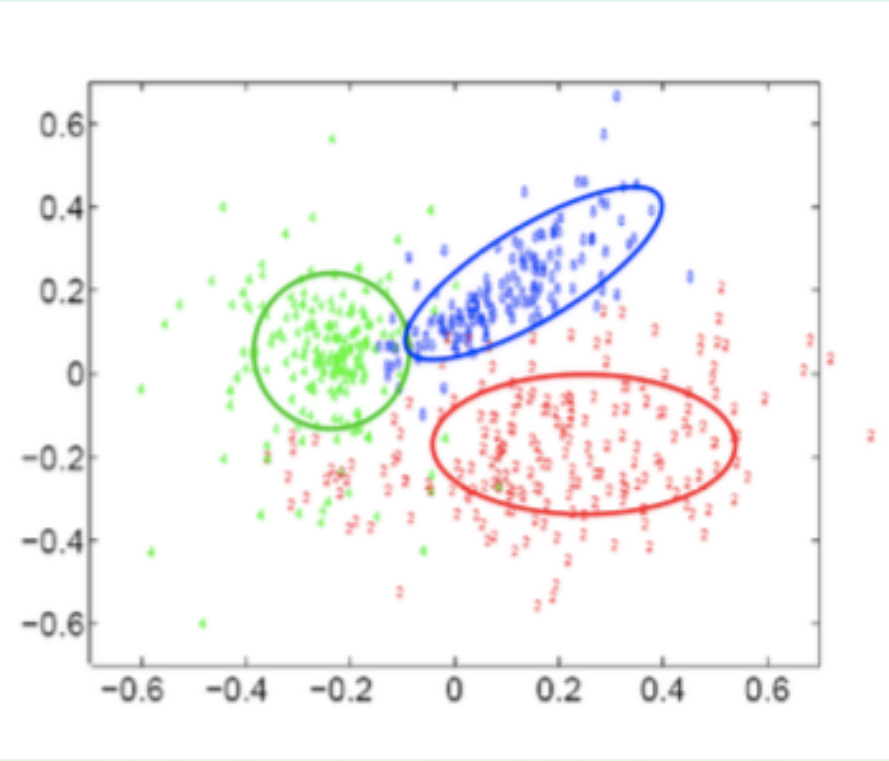
Gaussian mixture model

Skupisko przykładów modelowane rozkładem Gausowskim

$$Pr(\mathbf{x}, y) = Pr(\mathbf{x}|y)\,Pr(y) = \mathcal{N}(\mathbf{x}|\mu_y, \mathbf{\Sigma}_y)\,Pr(y)$$



Oryginalny rozkład przykładów



Każde skupisko – klasa przybliżona rozkładem Gausowskim

Mieszaniny rozkładów dla danych etykietowanych

Model parameters: $\theta = \{w_1, w_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$
The GMM:

Estimate the parameters from the labeled data

$$\begin{aligned} p(x, y|\theta) &= p(y|\theta)p(x|y, \theta) \\ &= w_y \mathcal{N}(x; \mu_y, \Sigma_y) \end{aligned}$$

Classification: $p(y|x, \theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)}$

Decision for any test point not in the labeled dataset

Za wykładem Barnabas Poczos SSL

Estymacja wiarygodności

Likelihood (assuming independently drawn data points)

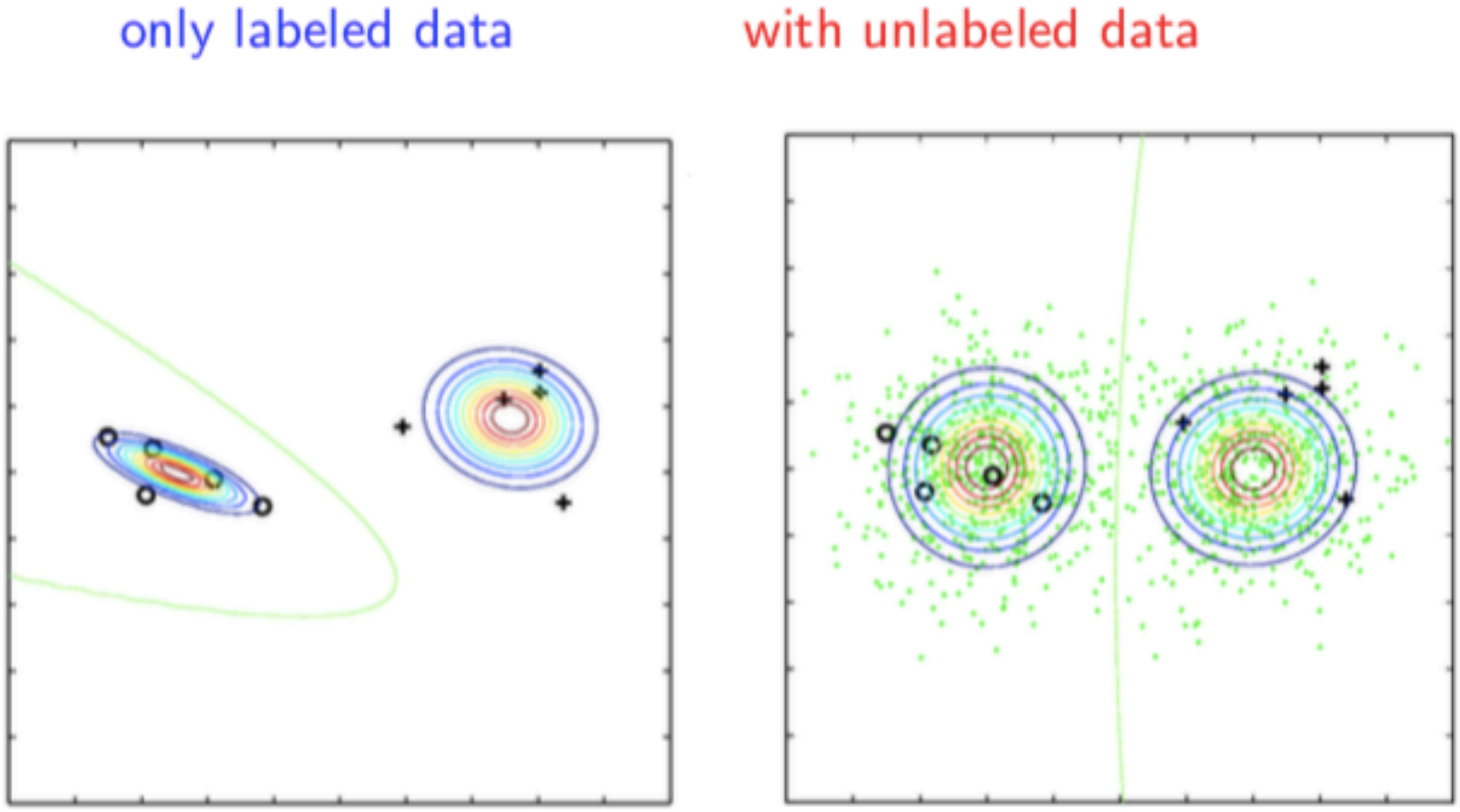
$$\begin{aligned} Pr\left(data\,|\theta\right) &= \prod_i Pr\left(\mathbf{x}_i, y_i\,|\theta\right) \prod_j Pr\left(\mathbf{x}_j\,|\theta\right) \\ &= \prod_i Pr\left(\mathbf{x}_i, y_i\,|\theta\right) \prod_j \sum_y Pr\left(\mathbf{x}_j, y\,|\theta\right) \end{aligned}$$

Minimize negative log likelihood:

$$-\log \mathcal{L}\left(\theta\right) = \underbrace{-\sum_i \log Pr\left(\mathbf{x}_i, y_i\,|\theta\right)}_{\textit{typically convex}} \underbrace{-\sum_j \log \left(\sum_y Pr\left(\mathbf{x}_j, y\,|\theta\right)\right)}_{\textit{typically non-convex}}$$

Standard tool for optimization (=training):
Expectation-Maximization (EM) algorithm

Wykorzystanie w SSL z nieetykietowanymi



from [Semi-Supervised Learning, ICML 2007 Tutorial; Xiaojin Zhu]

Podejścia Self-learning

Ucz się *f* z *L* i klasyfikuj przykłady *U* oceniąc pewność predykcji

Wykorzytaj cześć najpewniejszych predykcji do rozszerzania zbioru uczącego

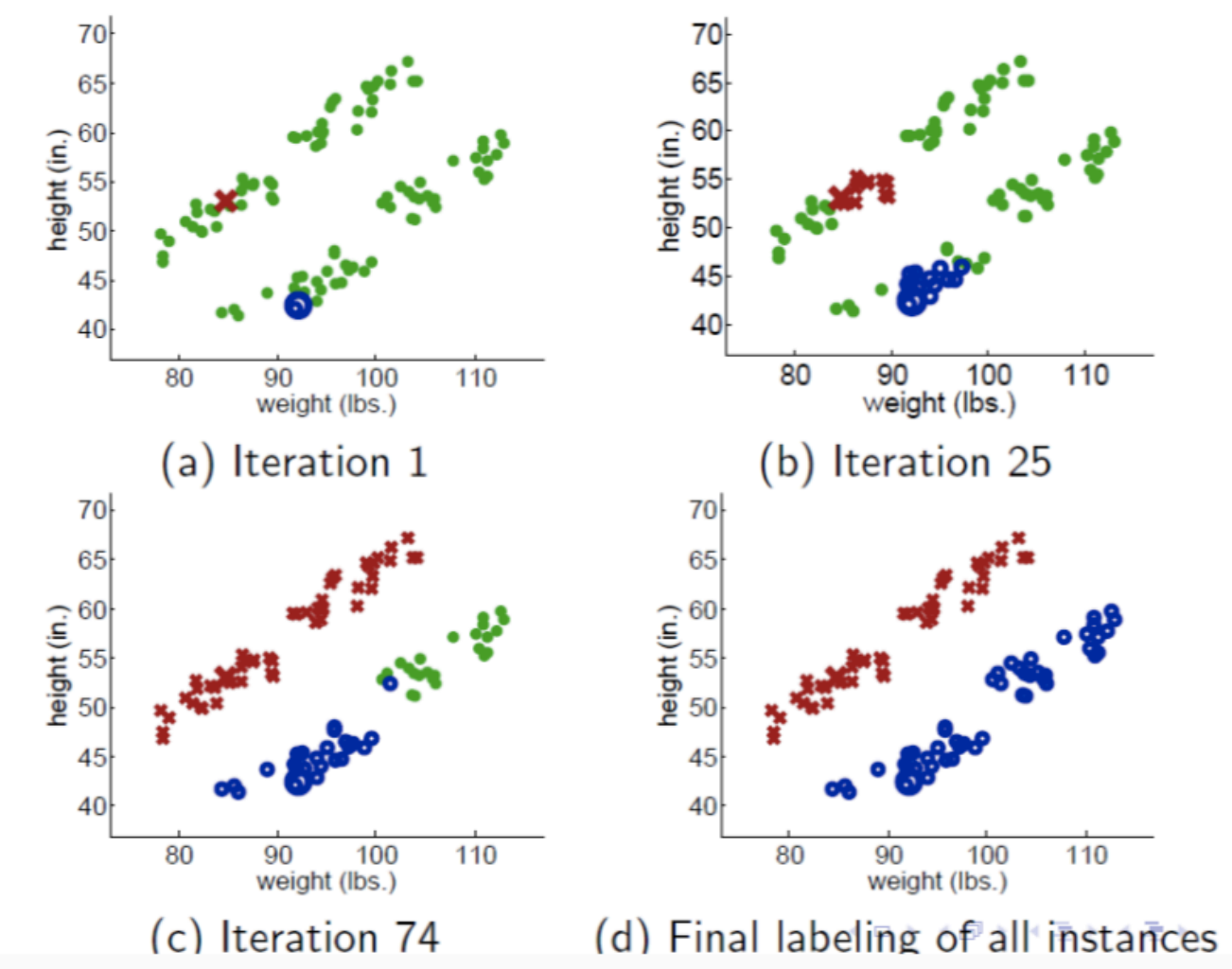
Algorithm 1 Self-training

1: **repeat**
2: $m \leftarrow train\_model(L)$
3: **for** $x \in U$ **do**
4: **if** $\max m(x) > \tau$ **then**
5: $L \leftarrow L \cup \{(x, p(x))\}$
6: **until** no more predictions are confident

Podejście dość ogólne (typu wrapper) stosowalne wookół klasycznych algorytmów / Yarowsky, 1995; McClosky et al., 2006

Pomimo ciekawych zastosowań, złożone dane i niewłaściwe predykcje mogą źle ukierunkować kolejne iteracje

Przykład self-learning



Propagowanie etykiet z wykorzystaniem 1-NN

Materiały dodatkowe

Artykuły:

J.Engelen, H.Hoos: A survey on semi-supervised learning. Machine learning 2020.

Xiaojin Zhu: Semi-Supervised Learning Literature Survey

Wykłady:

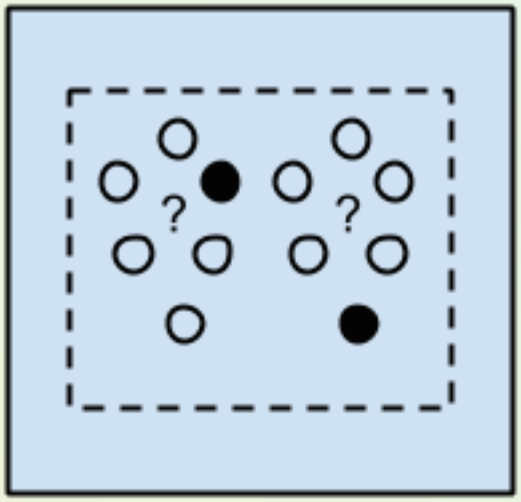
P.Rai: CS5350/6350 Semi-supervised Learning

A.Zien: Semi-supervised learning: Tutorial at ANN summer school

Barnabas Poczos: Semi-supervised learning

S.Ruder: An overview of proxy-label approaches for semi-supervised learning

Dziękuję za uwagę



Semi-supervised
Learning Algorithms

Pytania lub komentarze?

Więcej informacji w publikacjach!
oraz następnym wykładzie

Kontakt:
Jerzy.Stefanowski@cs.put.poznan.pl