Uczenie się częściowo nadzorowane

Semi-supervised learning

Część 2 – self learning, co-training, multi-view oraz modyfikacje S3VM

Wykład spec. projekt eksploracji danych

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska
Poznań

Poznań, 2020

Plan wykładu

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
   - Rozszerzanie SVM
   - Podejścia grafowe z propagacją etykiet
4. Inne zagadnienia i wyzwania

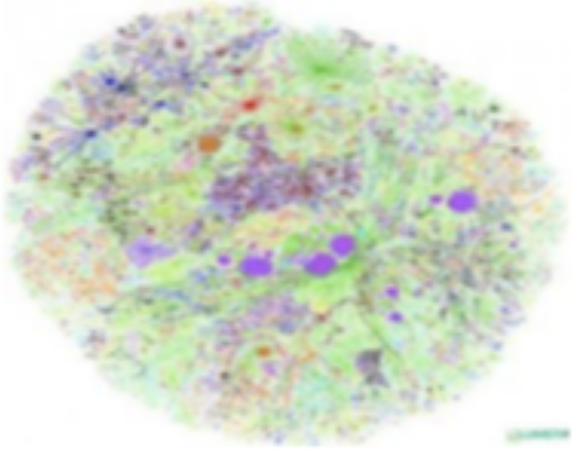
Dzisiaj cz. 2

Współczesne zastosowania - inne wyzwania

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.

Protein sequences



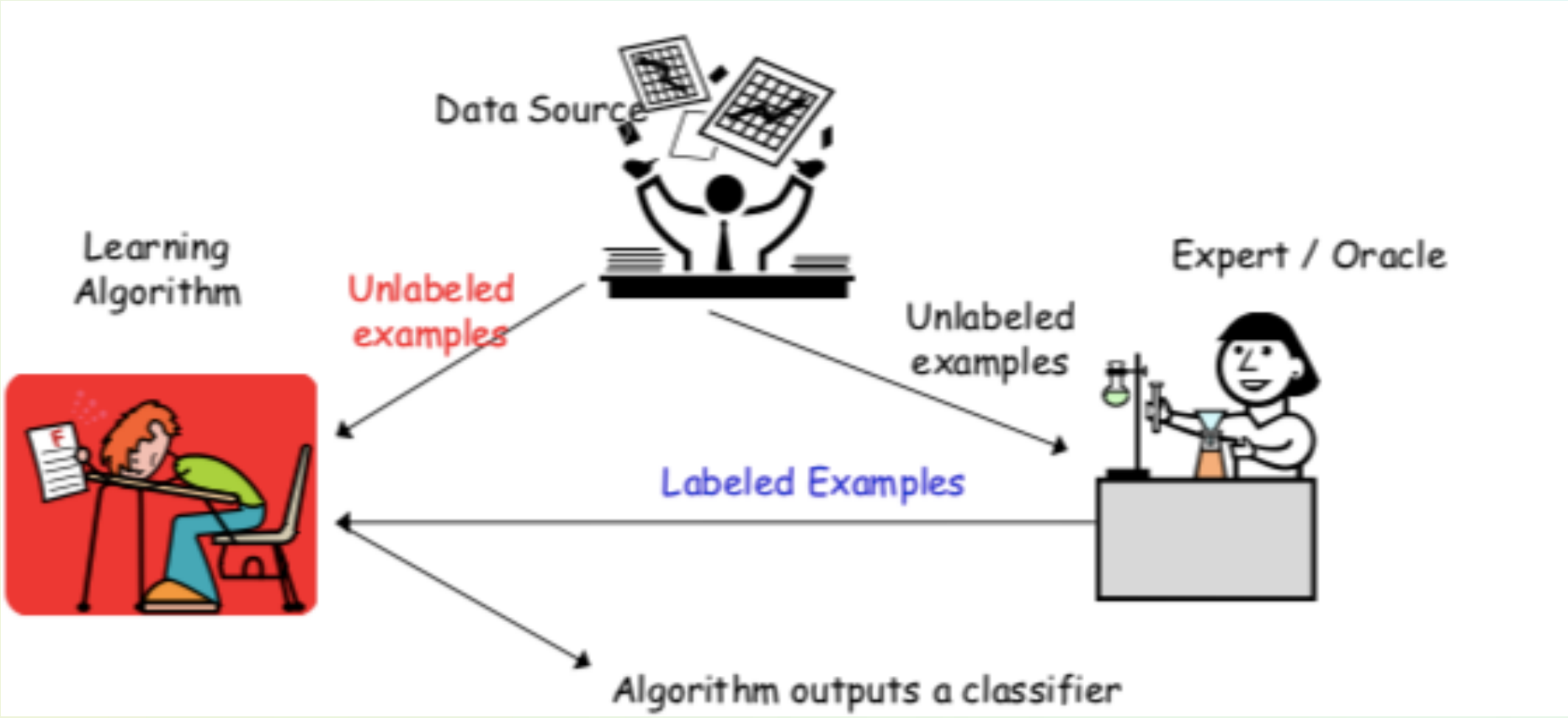
Billions of webpages

Images

Etykietowanie – uciążliwe, trudne, drogie, wymaga interakcji z ekspertem
Nieetykietowane – ławto pozyskiwalne,

za wykładem Balcan Semi-supervised learning

Ogólne motywacje – skorzystajmy z nieetykietowanych przykładów do ulepszenia klasyfikatora



za wykładem N. Balcan

Różne podejścia do uczenia się z przykładów

- Uczenie w pełni nadzorowane (supervised) – etykietowane przykłady

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

- Uczenie nienadzorowane (unsupervised) – nieetykietowane

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

- 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

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) / Yarowsky, 1995; McClosky et al., 2006

Pomimo ciekawych zastosowań, także otwarte pyania

Samo-uczenie, cd.

Otwarte pytania:

Q1: tzw. Wrapper approach – zastanów się dlaczego

Q2: Parametryzacja – co i jak dobieramy?

Q3: Kiedy ma szanse na działanie (tzw. bootstrap classifier)?

Podejścia Self-learning - pytania

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 różnych klasyfikatorów – lecz w miarę skutecznych (mały błąd) z tzw. dobrym oszacowaniem marginesu pewności

Wybór miary pewności predykcji (wiele możliwości), także wybór warunku zatrzymania

Złożone dane i niewłaściwe predykcje mogą źle ukierunkować kolejne iteracje

“The main downside of self-training is that the model is unable to correct its own mistakes. If the model's predictions on unlabelled data are confident but wrong, the erroneous data is nevertheless incorporated into training and the model's errors are amplified.” S.Ruder

Inny paradygmat -> co-training

Co-training wprowadzony przez (Blum & Mitchell, 1998) (Mitchell, 1999) wykorzystuje:

- tzw. **dwa różne, komplementarne spojrzenia** (views) na dane uczące, które są przydatne do budowy klasyfikatorów
- zbiór atrybutów X podzielony na **dwa rozłączne** podzbiory $\{X_1, X_2\}$, każdy z nich dostarcza wystarczająco dużo informacji aby wytrenować poprawny klasyfikator przy odpowiednio dużej ilości danych

Hipoteza zgodności -> klasyfikatory wyuczone z części

c_1, c_2 s.t. $c_1(X_1)=c_2(X_2)=C^*(X)$ mogę być potencjalnie zgodne dla wystarczająco dużej liczby przykładów

Dwa niezależnie zbiory atrybutów

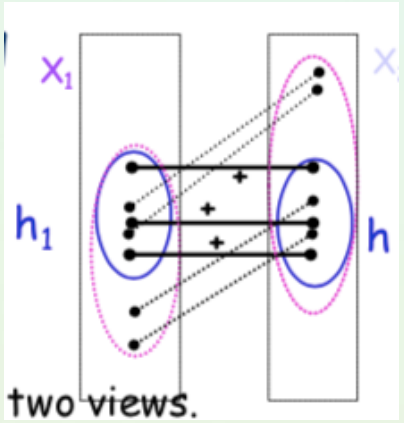
Naturalne w części zastosowań:

- Obrazy: - różne metody przetwarzania, np. view 1 - pixel features; view 2 - fourier coefficients
- Teksty/emaile: nagłówek (header) vs. wnętrze (tekst)

Strony WWW = oryginalne zadania Blum & Mitchell (zawartość tekstowa strony vs. linki i ich opisy anchor text of hyperlinks pointing to the webpage)

Idea iteracyjnego uczenia się wzajemnego

- Każdy klasyfikator jest uczony niezależnie na etykietowanych przykładach ze swoim zbiorem atrybutów (L_1 oddzielnie L_2)
- Następnie każdy klasyfikator jest użyty do predykcji klasyfikacji puli nieetykietowanych przykładów U
- Jeden klasyfikator przekazuje kilka najpewniejszych “swoich predykcji” dla nieetykietowanych przykładów drugiemu klasyfikatorowi
- Zbiory uczące L1 i L2 są poszerzone i można ponownie wyuczyć klasyfikatory (z oddzielnymi zbiorami atrybutów)
- Idea: “Each view teaching (training) the other view”



Przykład A.Blum

Oryginalne zadanie – klasyfikacja stron jako “faculty vs. no-faculty ones”

Poszukiwanie reguł klasyfikacyjnych z każdego zbioru cech

Idea: Use small labeled sample to learn initial rules.

- E.g., “my advisor” pointing to a page is a good indicator it is a faculty home page.
- E.g., “I am teaching” on a page is a good indicator it is a faculty home page.

Idea: Use unlabeled data to propagate learned information.

my advisor

Iterative co-training

Then look for unlabeled examples where one rule is confident and the other is not. Have it label the example for the other.

Training 2 classifiers, one on each type of info. Using each to help train the other.

Za wykładem A.Blum

Zapis algorytmu co-training (wersja ogólna)

- Given labeled data L and unlabeled data U
- Create two labeled datasets L_1 and L_2 from L using views 1 and 2
- Learn classifier $f^{(1)}$ using L_1 and classifier $f^{(2)}$ using L_2
- Apply $f^{(1)}$ and $f^{(2)}$ on unlabeled data pool U to predict labels
  - Predictions are made only using their own set (view) of features
- Add K most confident predictions $((\mathbf{x}, f^{(1)}(\mathbf{x}))$ of f_1 to L_2
- Add K most confident predictions $((\mathbf{x}, f^{(2)}(\mathbf{x}))$ of f_2 to L_1
- Note: Absolute margin could be used to measure confidence
- Remove these examples from the unlabeled pool
- Re-train $f^{(1)}$ using L_1, $f^{(1)}$ using L_2
- Like self-training but two classifiers teaching each other
- Finally, use a voting or averaging to make predictions on the test data

Co-training

Uwagi badawcze:

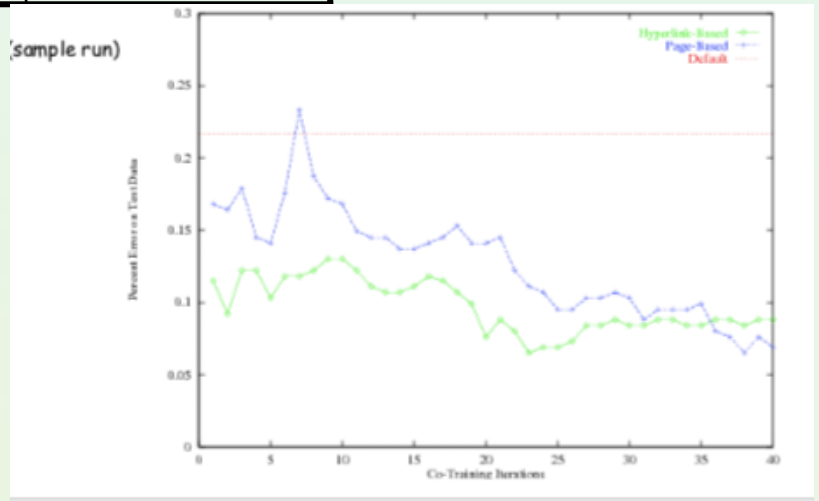
- Views – niezależne
- Klasyfikatory - weakly-useful predictor, tzn. It has higher probability of saying positive on a true positive than it does on a true negative, by at least some thresold
- Posiadamy wystarczającą liczbę przykładów startowych

Badania M.Balcan : Direct Optimization of Agreement bewteen h1 and h2

Przykłady eksperymentów [Blum, Mitchell]

- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: learning from labeled data 11.1%;
- average error: co-training 5.0%

	Page-base classifier	Link-based classifier	Combined classifier
Supervised training	12.9	12.4	11.1
Co-training	6.2	11.6	5.0



Za wykład A.Blum

Inne zastosowania - obrazy

Results: images [Levin-Viola-Freund '03]:

- Visual detectors with different kinds of processing

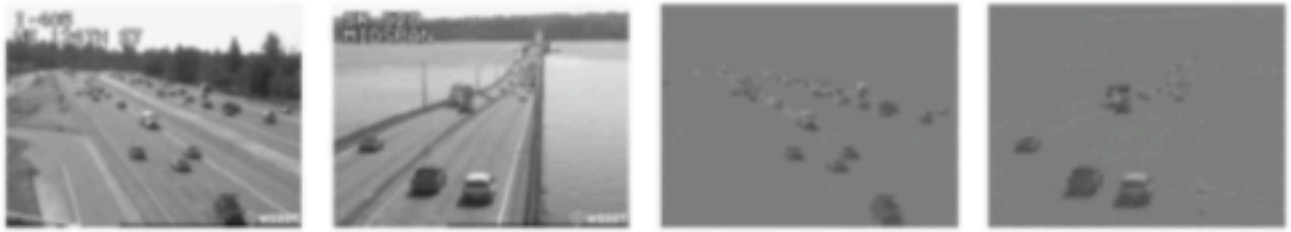
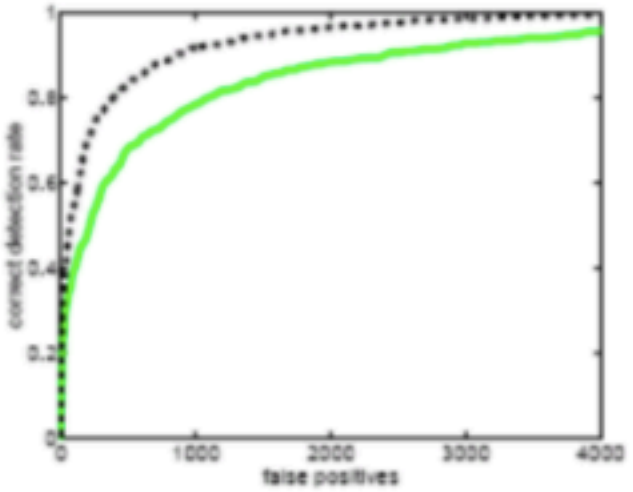


Figure 1: Example images used to test and train the car detection system. On the left are the original images. On the right are background subtracted images.

- Images with 50 labeled cars. 22,000 unlabeled images.
- Factor 2-3+ improvement.

From [LVF03]



Multi-view learning- rozszerzenie

- Więcej niż dwa spojrzenia na dane (zbiory atrybutów)
- Naucz niezależne klasyfikatory na oddzielnych zbiorach (views – zbiory atrybutów)
- Modus operandi: predykcje klasyfikatorów wystarczająco zgodne na przykładach nieetykietowanych
- Dla przykładów testowych – agregacja głosowania jak w zespole klasyfikatorów
- Co-training specjalny pod-przypadek

Spojrzeć do: J.Zhao et al: Multi-view Learning Overview: Recent Progress and New Challenges.

Podejścia inspirowane

Spójrz do bloga Sebastian Ruder'a An overview of proxy-label approaches for semi-supervised learning

Powiązane z multi-view training

- Democratic co-training
- Tri-training
- Tri-training with disagreement
- Asymmetric tri-training
- Multi-task tri-training
- ....

Democratic co-training

Wielokrotne uczenie zróżnicowanych klasyfikatorów (z tzw. Innymi inductive bias) z L i ocena ich predykcji na U

M – zbiór klasyfikatorów z tą samą predykcją j-tej klasy dla przykładu x

w – ocena pewności predykcji klasyfikatora

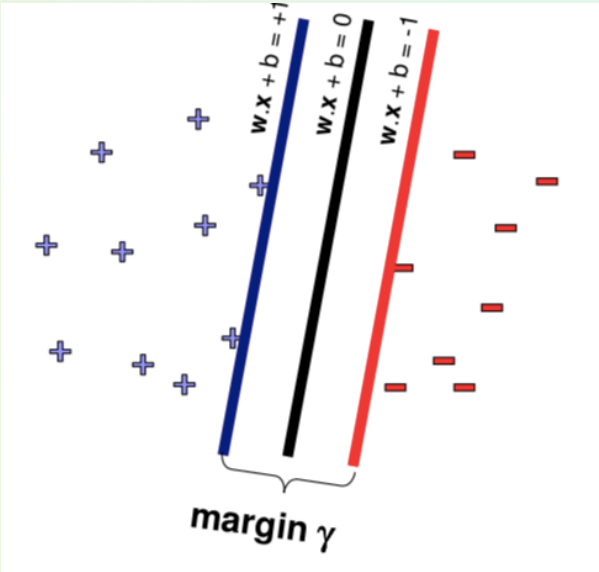
Przykłady nieetykietowane dla które większość klasyfikatorów osiąga wysoką zgodność mogą być dodane do zbioru uczącego

Algorithm 3 Democratic Co-learning

repeat
  for $i \in \{1..n\}$ **do**
    $m_i \leftarrow train\_model(L)$
  for $x \in U$ **do**
    for $j \in \{1..C\}$ **do**
      $M \leftarrow \{i \mid p_i(x) = j\}$
      if $|M| > n/2$ **and** $\sum_{i \in M} w_i > \sum_{i \notin M} w_i$ **then**
        $L \leftarrow L \cup \{(x, j)\}$
until none of m_i changes
apply weighted majority vote over m_i

SVM w obecności częściowo etykietowanych

Przypomnijmy zasady klasycznego SVM



$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \ldots, l$$

W trudniejszych nakładających się rozkładach przykładów element regularyzacyjny ze zmiennymi osłabiającymi

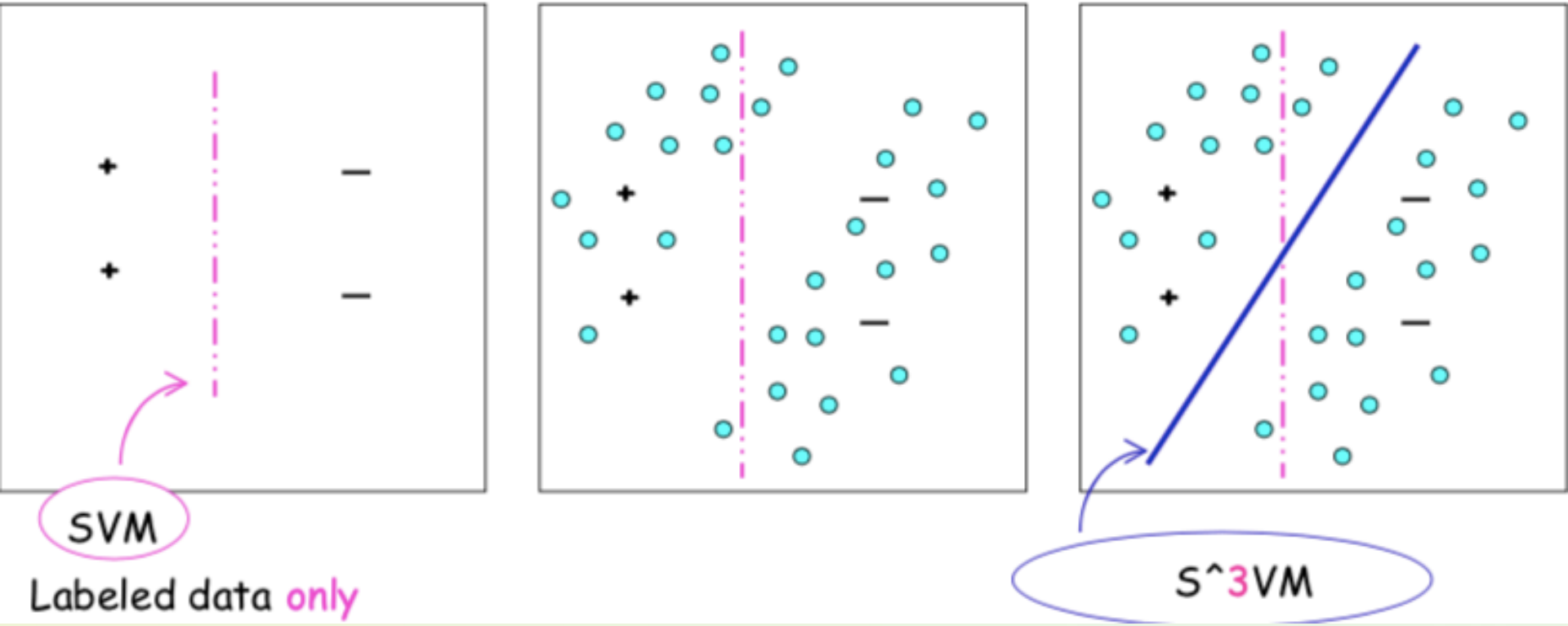
$$L(w) = \frac{\|\vec{w}\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k\right)$$

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 - \xi_i \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 + \xi_i \end{cases}$$

Semi-supervised SVM

Granica decyzyjna powinna przejść przez **obszar o małej gęstości przykładów** (etykietowanych i nieetykietowanych) pomiędzy dwoma klasami (-1;1)

Rozszerzenie idei sformułowania zadania SVM, aby uwzględnić obecność nieetykietowanych przykładów



SVM -> S3VM

Zadanie optymalizacji SVM

$$\min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$$
$$s.t. \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, l$$

where $\xi_i (i = 1, 2, \ldots, n)$ are the slack variables.

Funkcja celu jest przetworzona z wykorzystaniem funkcji straty (Hinge loss)

$$\min \Phi(w) = \min_{w,b}\left\{\frac{1}{2}||w||^2 + C\sum_{i}^{l}\max(1 - y_i[\mathbf{w} \cdot x_i + b], 0)\right\} \quad (10)$$

Here, $\frac{1}{2}||\mathbf{w}||^2$ can be seen as a regularization term, and $\max(1 - y_i[\mathbf{w} \cdot x_i + b], 0)$ is the loss function of labeled data.

za: Ding. S. et al. An overview on semi-supervised support vector machine

S3VM

Przykłady nieetykietowane – analogiczna funkcja straty

$$\max(1-|f(x)|,0)$$

- ma wartość dodatnią, gdy -1 < f(x) < 1 oraz 0 poza tym.

Odpowiednik kary za naruszenie separawalności marginesu

Dla wszystkich przykładów nieetykietowanych w funkcji celu pojawi się element

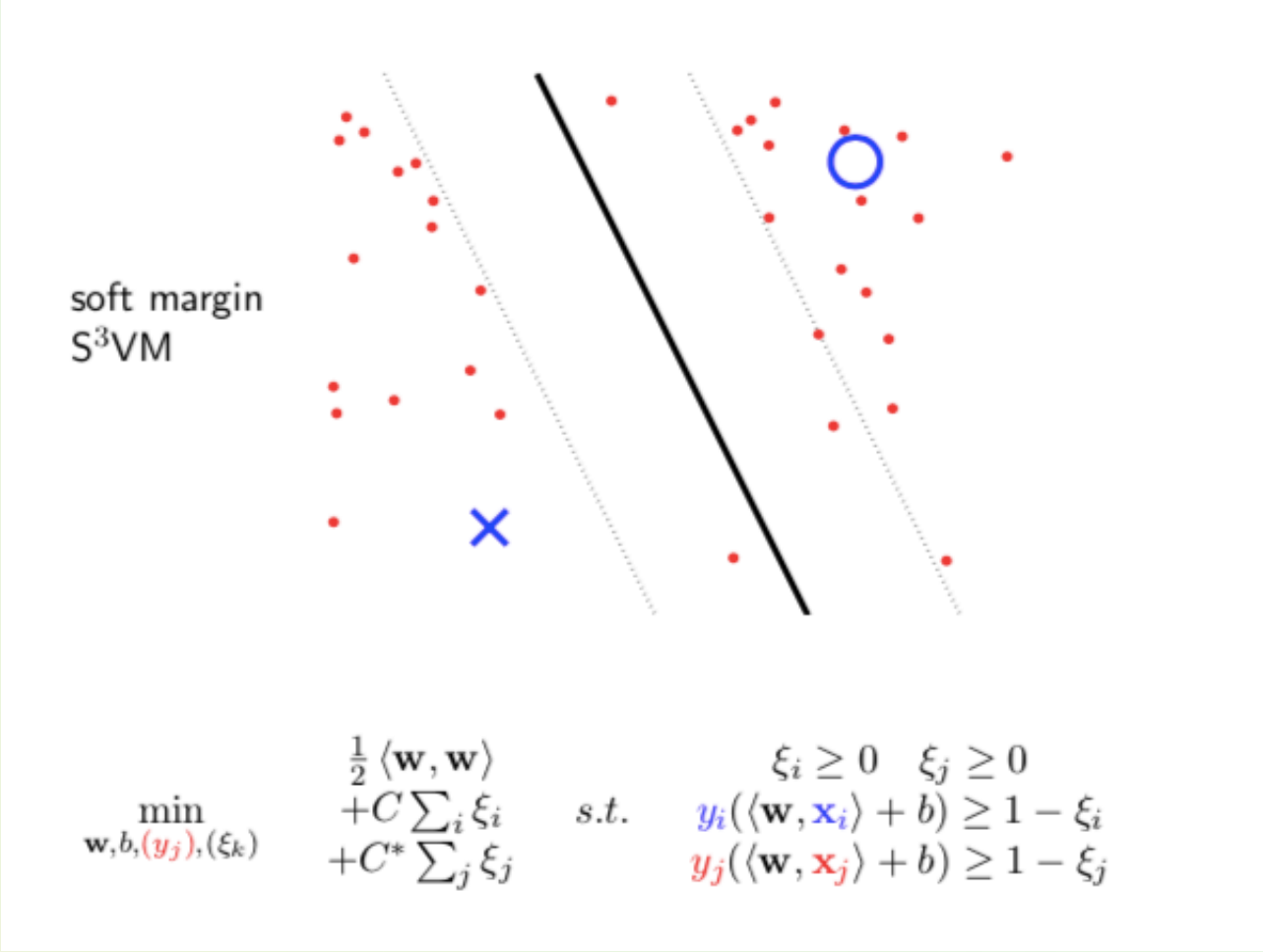
$$\frac{1}{u}\sum_{i=l+1}^{l+u}\max(1-|f(\mathbf{x})|,0).$$

Co prowadzi do re-definicji zadania programowania matematycznego

Trudne do rozwiązania – nieciągłe zadanie programowania kwadratowego

Przeformułowanie S3VM w pracy [Bennett and Demiriz]

S3VM



Za wykładem Alex Zien Semi-supervised learning. S3VM

S3VM

Supervised Support Vector Machine (SVM)

$$\min_{\mathbf{w},b,(\xi_k)} \quad \begin{array}{l} \frac{1}{2}\langle \mathbf{w},\mathbf{w}\rangle \\ +C\sum_i \xi_i \end{array} \quad s.t. \quad \begin{array}{c} \xi_i \geq 0 \\ y_i(\langle \mathbf{w},\mathbf{x}_i\rangle + b) \geq 1-\xi_i \end{array}$$

- maximize margin on (labeled) points
- convex optimization problem (QP, quadratic programming)

Semi-Supervised Support Vector Machine (S^3VM)

$$\min_{\mathbf{w},b,(y_j),(\xi_k)} \quad \begin{array}{l} \frac{1}{2}\langle \mathbf{w},\mathbf{w}\rangle \\ +C\sum_i \xi_i \\ +C^*\sum_j \xi_j \end{array} \quad s.t. \quad \begin{array}{c} \xi_i \geq 0 \quad \xi_j \geq 0 \\ y_i(\langle \mathbf{w},\mathbf{x}_i\rangle + b) \geq 1-\xi_i \\ y_j(\langle \mathbf{w},\mathbf{x}_j\rangle + b) \geq 1-\xi_j \end{array}$$

- maximize margin on labeled and unlabeled points
- also QP?

Za wykładem Alex Zien Semi-supervised learning. S3VM

Transductive Reasoning

Komentarz:

> ”Transductive” means here ”reasoning from particular to particular” as opposed to ”Inductive” - ”reasoning from particular to general”. So the motivation is that the performance of the model outside the current set of labeled and unlabeled points is not of interest

lub [Wikipedia]

> Transduction or transductive inference is reasoning from observed, specific (training) cases to specific (test) cases.

jako przeciwieństwo klasycznej zasady indukcji :

> induction is reasoning from observed training cases to general rules, which are then applied to the test cases

Transductive SVM

Formalizacja Joachims [99] wykorzystuje założenia skupień podobnych obiektów i podobny problem optymalizacyjny

Input: $S_l=\{(x_1, y_1), \ldots,(x_{m_l}, y_{m_l})\}$

$S_u=\{x_1, \ldots, x_{m_u}\}$

$$\operatorname{argmin}_w \left|\left|w\right|\right|^2 + C\sum_i \xi_i + C\sum_u \widehat{\xi_u}$$

- $y_i\, w \cdot x_i \geq 1-\xi_i$, for all $i \in \{1, \ldots, m_l\}$
- $\widehat{y_u} w \cdot x_u \geq 1 - \widehat{\xi_u}$, for all $u \in \{1, \ldots, m_u\}$
- $\widehat{y_u} \in \{-1, 1\}$ for all $u \in \{1, \ldots, m_u\}$

Zmienne osłabiające – funkcje straty Hinge loss – jak poprzednio

Transductive SVM [Joachims]

Wykorzystuje optymalizacje na podzbiorze przykładów L i U
i próbuje zaetykietować część przykładów U (tak jakby były testowe)
Następnie przełącza się etykiety niektórych przykładów, jeśli powiększają margines

1. Train SVM on labeled data
2. Classify the unlabeled data using the SVM, assign q examples with highest $f(x)$ examples to the positive class, rest to negative class (q is user-set parameter)
3. Initialize $C* = 1E-5$ (small value, almost disregading unlabeled data points)
4. Iterate while increasing slack coefficient $C*$ of unlabeled data points:
   - 4.1 Find two unlabeled examples that have different predicted labels, and a total Hinge loss > 2
   - 4.2 Swap the predicted labels of the examples $\Longrightarrow$ decrease Hinge loss
   - 4.3 Increase slack parameter $C*$ and and retrain SVM

Parę uwag nt. implementacji

Sklearn Python – tylko wersje label propagation

Indywidulane projekty – patrz listy mat. dodatkowe

SVM – dostępny w nowych bibliotekach R -> RSSL Package

Package 'RSSL'

February 4, 2020

Version 0.9.1

Title Implementations of Semi-Supervised Learning Approaches for Classification

Depends R(>= 2.10.0)

Imports methods, Rcpp, MASS, kernlab, quadprog, Matrix, dplyr, tidyr, ggplot2, reshape2, scales, cluster

LinkingTo Rcpp, RcppArmadillo

Suggests testthat, rmarkdown, SparseM, numDeriv, LiblineaR

Description A collection of implementations of semi-supervised classifiers and methods to evaluate their performance. The package includes implementations of, among others, Implicitly Constrained Learning, Moment Constrained Learning, the Transductive SVM, Manifold regularization, Maximum Contrastive Pessimistic Likelihood estimation, S4VM and WellSVM.

License GPL (>= 2)

URL http://www.github.com/jkrijthe/RSSL

BugReports http://www.github.com/jkrijthe/RSSL

Inne zagadanienia

Podejścia grafowe i propagacja etykiet – kolejny wykład

Inne metody, nieomawiane w tej edycji przedmiotu

- Zaawansowane metody generatywne (np. wykorzystujące Hidden Markov Models)
- Więcej o Multi-view learning
- Sieci neuronowe
  - w szczególności: GAN - Generative Adversarial Networks - w uczeniu częściowo nadzorowanym (z mała liczbą etykietowych obrazów)
  - oraz inne DANN, i tzw. Restricted Boltzmann Machines

Materiały dodatkowe

Artykuły:

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

Jing Zhao, Xijiong Xie, Xin Xu, Shiliang Sun: Multi-view Learning Overview: Recent Progress and New Challenges

Avrim Blum, Tom Mitchell: Combining Labeled and Unlabeled Data with Co- Training. COLT 1998

Książka: Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge, Mass.: MIT Press

Wykłady:

N.Balcan: Semi-supervised Learning [inspiracja dla obecnego wykładu)

Xiaojin Zhu: Semi-Supervised Learning Tutorial [także inspiracja]

Barnabas Poczos: Semi-supervised learning

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

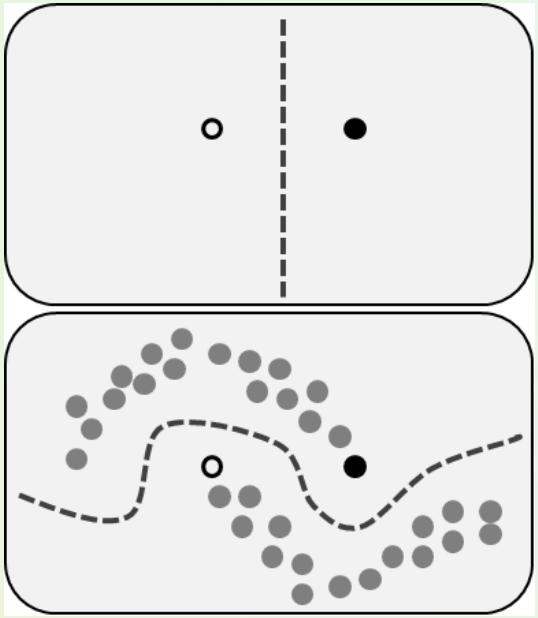
Strona przeglądowa

https://sci2s.ugr.es/ssl - repozytorium Univ. Granada

Inne związane zagadnienia

- Semi-supervised clustering
- Active learning i semi-supervised learning
- Weakly – supervised learning
- Labeling only one class
- …

Dziękuję za uwagę



Pytania lub komentarze?

Więcej informacji w publikacjach
oraz materiały dodatkowe!

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