

An Introduction to Transfer Learning and Domain Adaptation

Amaury Habrard

Laboratoire Hubert Curien, UMR CNRS 5516, Université de Saint-Etienne

EPAT 2014

Some resources

- List of transfer learning papers
<http://www1.i2r.a-star.edu.sg/~jspan/conferenceTL.htm>
- List of available softwares
<http://www.cse.ust.hk/TL/index.html>
- Surveys
 - Patel, Gopalan, Chellappa. Visual Domain Adaptation: An Overview of Recent Advances. Tech report, 2014.
 - Qi Li. Literature Survey: Domain Adaptation Algorithms for Natural Language Processing, Tech report, 2012
 - Margolis. A Literature Review of Domain Adaptation with Unlabeled Data. Tech report 2011.
 - Pan and Yang. A survey on Transfer Learning', TKDE 2010.
 - J. Quionero-Candela and M. Sugiyama and A. Schwaighofer and N.D. Lawrence. Dataset Shift in Machine Learning. MIT Press.

Credits and acknowledgments

Documents used for this talk:

- D. Xu, K. Saenko, I. Tsang. Tutorial on Domain Transfer Learning for Vision Applications, CVPR'12.
- S. Pan, Q. Yang and W. Fan. Tutorial: Transfer Learning with Applications, IJCAI'13.
- S. Ben-David. Towards Theoretical Understanding of Domain Adaptation Learning, workshop LNIID at ECML'09.
- F. Sha and B. Kingsbury. Domain Adaptation in Machine Learning and Speech Recognition, Tutorial - Interspeech 2012.
- K. Grauman. Adaptation for Objects and Attributes, workshop VisDA at ICCV'13
- J. Blitzer and H. DauméIII. Domain Adaptation, Tutorial - ICML 2010.
- A. Habrard, JP Peyrache, M. Sebban. Iterative Self-labeling Domain Adaptation for Linear Structured Image Classification, IJAIT 2013.
- A. Habrard, JP Peyrache, M. Sebban. Boosting for unsupervised domain adaptation, ECML 2013

Acknowledgments: B. Fernando, P. Germain, E. Morvant, JP Peyrache, M. Sebban.

Transfer Learning

Definition [Pan, TL-IJCAI'13 tutorial]

Ability of a system to recognize and apply knowledge and skills learned in previous domains/tasks to novel domains/tasks

An example

- We have **labeled** images from a **Web image corpus**
- Is there a Person in **unlabeled** images from a **Video corpus** ?



Person



no Person



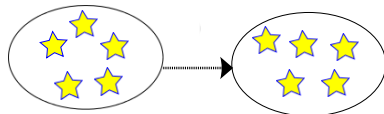
Is there a Person?

Outline

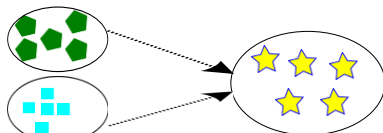
- 1 Introduction/Motivation
- 2 Reweighting/Instance based methods
- 3 Theoretical frameworks
- 4 Feature/projection based methods
- 5 Adjusting/Iterative methods
- 6 A quick word on model selection

Introduction

Settings



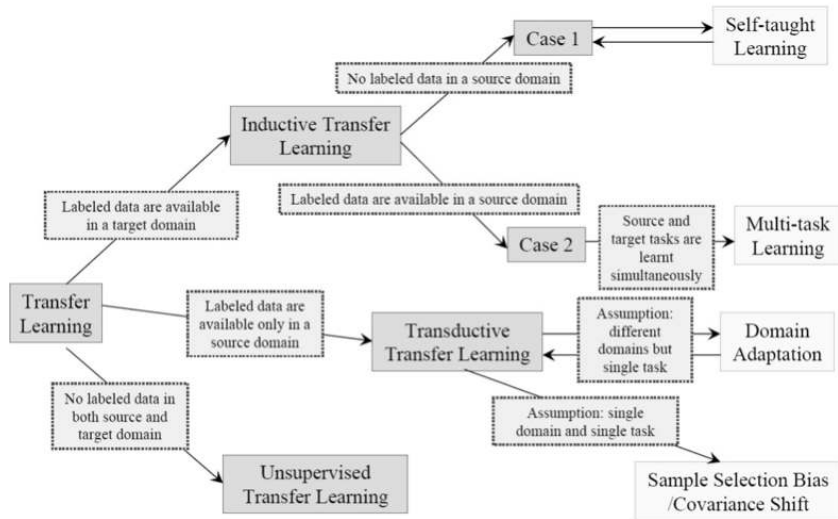
Training and test data are
from the same domain



Training and test data are
from different domains

- Domains are modeled as probability distributions over an instance space
- Tasks associated to a domain (classification, regression, clustering, ...)
- **Objective:** From **source** to **target**
⇒ Improve a target predictive function in the **target** domain using knowledge from the **source** domain

A Taxonomy of Transfer Learning



“A survey on Transfer Learning” [Pan and Yang, TKDE 2010]

In this presentation

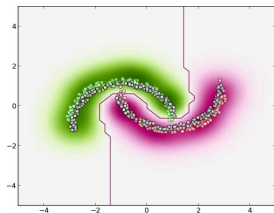
- We will make a focus on **domain adaptation**
- We will focus on classification tasks

⇒ How can we learn, using labeled data from a **source distribution**, a low-error classifier for **another related target distribution**?

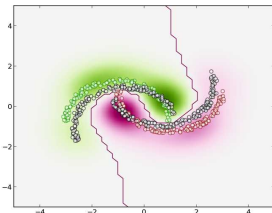
⇒ “Hot topic” - tutorials at ICML 2010, CVPR 2012, Interspeech 2012, workshops at ICCV 2013, NIPS 2013, ECML 2014

⇒ Motivating examples

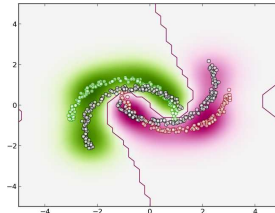
A toy problem: Inter-twinning moons



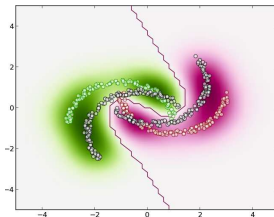
(a) 10°



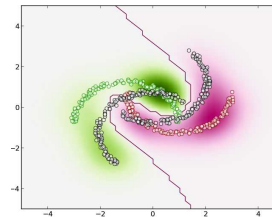
(b) 20°



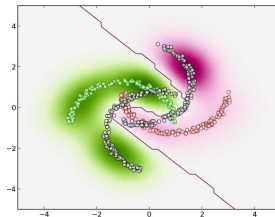
(c) 30°



(d) 40°



(e) 50°



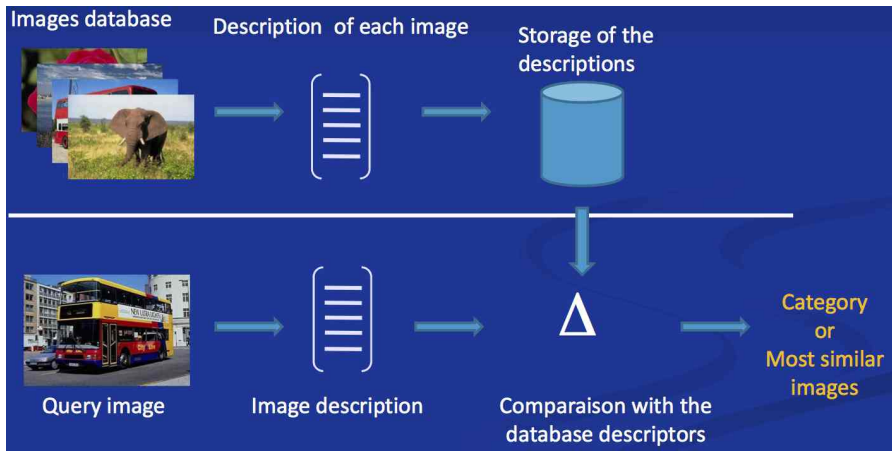
(f) 70°

Intuition and motivation from a CV perspective

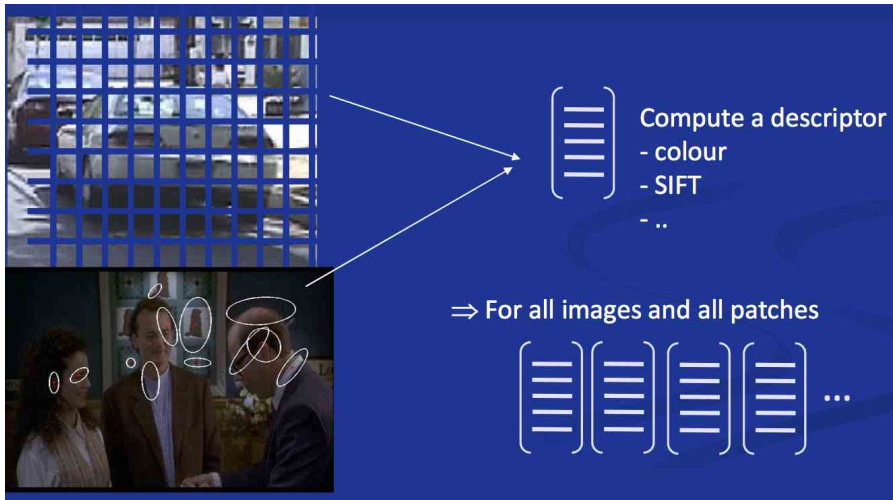


- “Can we train classifiers with Flickr photos, as they have already been collected and annotated, and hope the classifiers still work well on mobile camera images?” [Gong et al., CVPR 2012]
- “object classifiers optimized on benchmark dataset often exhibit significant degradation in recognition accuracy when evaluated on another one” [Gong et al., ICML 2013, Torralba et al., CVPR 2011, Perronnin et al., CVPR 2010]
- “Hot topic” -Visual domain adaptation [Tutorial CVPR’12, ICCV’13]

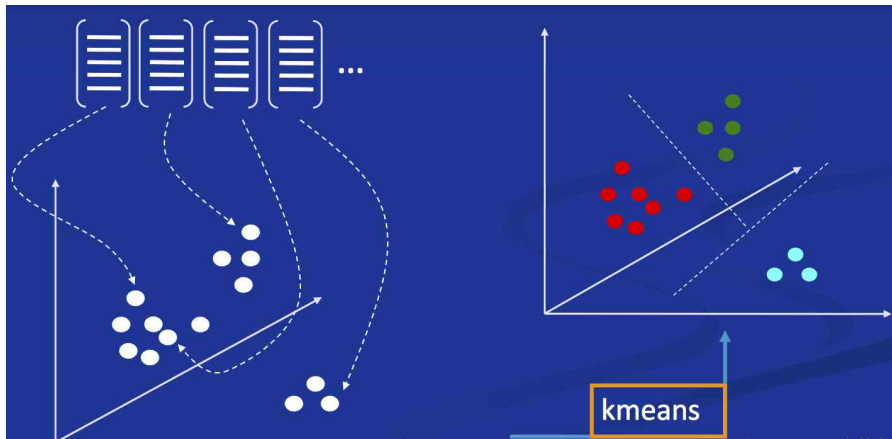
Brief recap on computer vision issues [Slides from J. Sivic]



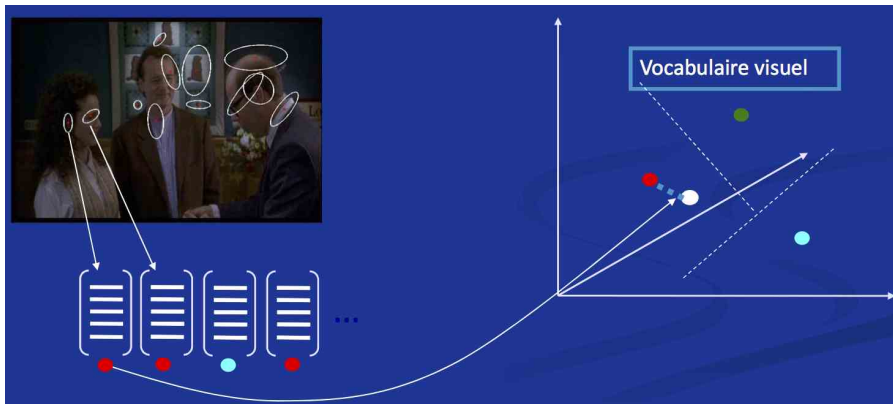
Brief recap on computer vision issues (2) [Slides from J. Sivic]



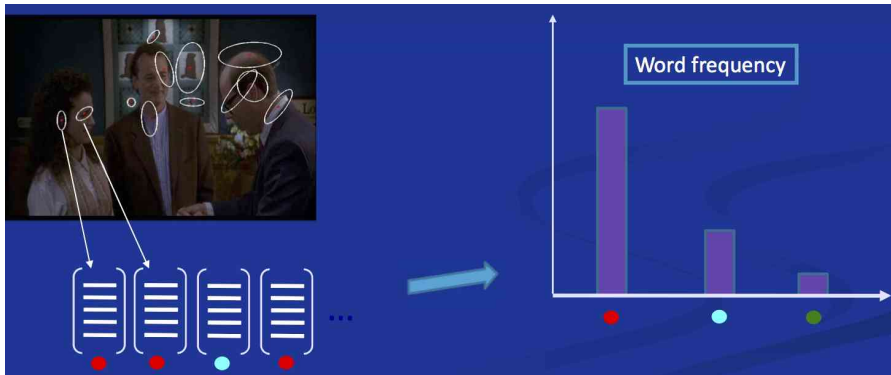
Brief recap on computer vision issues (3) [Slides from J. Sivic]



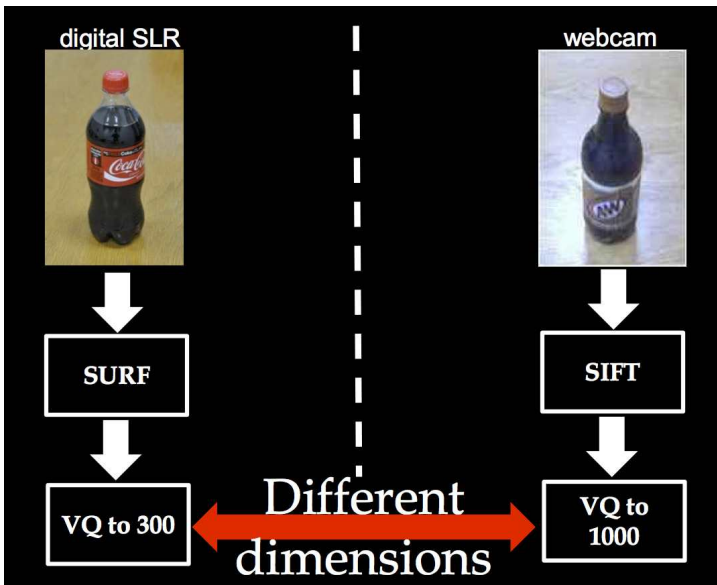
Brief recap on computer vision issues (4) [Slides from J. Sivic]



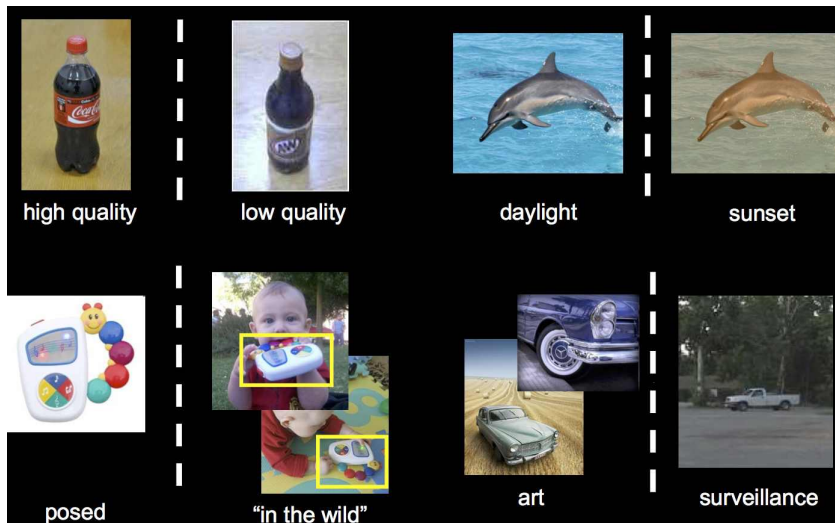
Brief recap on computer vision issues (5) [Slides from J. Sivic]



Problems with data representations



Hard to predict what will change in the new domain



[Xu, Saenko, Tsang, Domain Transfer Tutorial - CVPR'12]

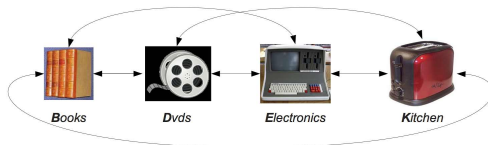
Natural Language Processing

Text are represented by “words” (Bag of Words)

- Part of Speech Tagging: Adapt a tagger learned from medical papers to a journal (Wall Street Journal) - Newsgroup

Biomedical	WSJ
the signal <i>required</i> to	investment <i>required</i>
stimulatory signal <i>from</i>	buyouts <i>from</i> buyers
essential signal <i>for</i>	to jail <i>for</i> violating

- Spam detection: Adapt a classifier from one mailbox to another



- Sentiment analysis:

Domain Adaptation for sentiment analysis



critiques de livres

??? The end of the series.

This book was written to provoke those who wanted Adams to continue the trilogy but I loved it. Author settled down on a bob fearing planet where he has aquired the prestigious...

[Read more](#)

Published on Mar 18 2002 by dan

??? Mostly Harmless is Underrated

I think most of the reviews for this book downplay it seriously. While the ending is kind of disappointing, the book overall is wonderful.

[Read more](#)

Published on Jan 22 2002 by A Big Adams Fan

??? Please pretend this book was never written.

I have long been a fan of the Hitchhikers series as they are comic genius. The book Mostly Harmless, however, should never have come about. It is frustration at its peak. [Read more](#)

Published on Jan 14 2002 by Paul Norrod

??? Kinda like horror movies...

...in that the last one usually isn't all that appealing. I liked it fine, with some of Adams's wit, but it was a bit disappointing. [Read more](#)

Published on Nov 4 2001 by Kristopher Vincent

??? A Terrible End to A Great Series

The ending for this books was so bad that I vowed never to read another Douglas Adams book. Adams was obviously sick and tired of the series and used this book to kill it off with...

[Read more](#)

Published on Oct 17 2001 by David A. Lessnau

Exemple



critiques de film

-1 An insult to Douglas Adams' memory

I agree entirely with "darkgenius" comments. This movie is a travesty of the book and the TV series; a cutesy version totally lacking in the wit and satire of the original. [Read more](#)

Published 5 months ago by John W Beare

+1 Don't Panic!

If you haven't listened to the BBC radio-play, this isn't bad! Purists, no doubt, will dispute my verdict but the fact of the matter is THGTTG (see title) does have Douglas Adams'...

[Read more](#)

Published on Mar 13 2011 by Sid Matheson

+1 On Blu-ray, even better

I've seen this movie on TV and wanted to add it to my collection. I couldn't find it locally so when I saw it on amazon and on Blu-ray, I picked it up. [Read more](#)

Published on April 18 2009 by J. W. Little

-1 An insult to Douglas Adams' memory

The filmmaker's reverence for Adams' legacy? What kind of rubbish statement is that? As a loyal fan of Douglas Adams for more than a quarter of a century, I was appalled and...

[Read more](#)

Published on Aug 22 2006 by Daniel Jolley

Algorithme d'apprentissage

Classificateur

-1

Domain Adaptation for sentiment analysis - ex [Pan-IJCAI'13 tutorial]

	Electronics	Video games
✓	(1) <u>Compact</u> ; easy to operate; very good picture quality; looks <u>sharp</u> !	(2) A very <u>good</u> game! It is action packed and full of excitement. I am very much <u>hooked</u> on this game.
✓	(3) I purchased this unit from Circuit City and I was very <u>excited</u> about the quality of the picture. It is really <u>nice</u> and <u>sharp</u> .	(4) Very <u>realistic</u> shooting action and good plots. We played this and were <u>hooked</u> .
✗	(5) It is also quite <u>blurry</u> in very dark settings. I will <u>never_buy</u> HP again.	(6) It is so boring. I am extremely <u>unhappy</u> and will probably <u>never_buy</u> UbiSoft again.

- Source specific: *compact, sharp, blurry*.
- Target specific: *hooked, realistic, boring*.
- Domain independent: *good, excited, nice, never_buy, unhappy*.

Other applications

- Speech recognition [Tutorial at Interspeech'12]
- Medecine
- Biology
- Time series
- Wifi localization

Notations

Notations

- $X \subseteq \mathbb{R}^d$ input space, $Y = \{-1, +1\}$ output space
- P_S **source domain**: distribution over $X \times Y$
 D_S marginal distribution over X
- P_T **target domain**: different distribution over $X \times Y$
 D_T marginal distribution over X
- $\mathcal{H} \subseteq Y^X$: hypothesis class

Expected error of a hypothesis $h : X \rightarrow Y$

- $R_{P_S}(h) = \mathbf{E}_{(\mathbf{x}^s, y^s) \sim P_S} \mathbf{I}[h(\mathbf{x}^s) \neq y^s]$ **source** domain error
- $R_{P_T}(h) = \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \mathbf{I}[h(\mathbf{x}^t) \neq y^t]$ **target** domain error

Domain Adaptation: find $h \in \mathcal{H}$ with R_{P_T} **small** from data $\sim D_T$ and P_S

Classical result in supervised learning

Empirical error

- $R_S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s} \sim (P_S)^{m_s}$ a labeled sample drawn i.i.d. from P_S
- Associated **empirical error** of an hypothesis h :

$$R_S(h) = \frac{1}{m_s} \sum_{i=1}^{m_s} \mathbf{I}[h(\mathbf{x}_i^s) \neq y_i^s]$$

Classical PAC result: From the same distribution

$$R_{P_S}(h) \leq R_S(h) + O\left(\frac{\text{complexity}(h)}{\sqrt{m_s}}\right)$$

\Rightarrow Occam razor principle

What about R_{P_T} if we have no or very few labeled data? \rightarrow try to make use of source information

Domain Adaptation

Setting

- **Labeled Source** Sample

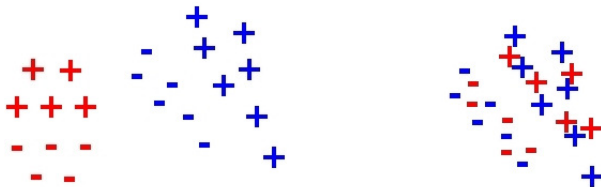
$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ **Source** sample drawn i.i.d. from P_S

- **Unlabeled Target** Sample

$T = \{\mathbf{x}_j\}_{j=1}^{m_t}$ **Target** Sample drawn i.i.d. from D_T

optionnal: a few labeled target examples

If h is learned from **source** domain, how does it perform on **target** domain?



Domain Adaptation

Setting

- **Labeled Source** Sample

$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ **Source** sample drawn i.i.d. from P_S

- **Unlabeled Target** Sample

$T = \{\mathbf{x}_j\}_{j=1}^{m_t}$ **Target** Sample drawn i.i.d. from D_T

optionnal: a few labeled target examples

If h is learned from **source** domain, how does it perform on **target** domain?

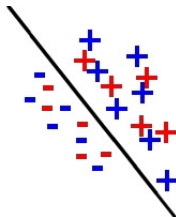
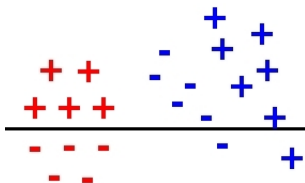
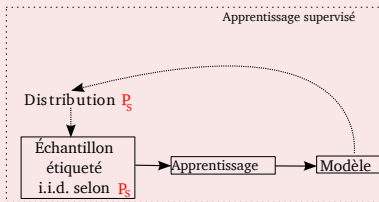
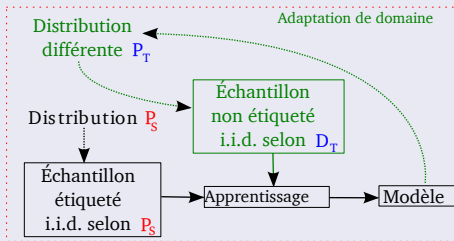


Illustration settings

Classical supervised learning



Domain adaptation



A bit of vocabulary

Unsupervised Transfer Learning

No labels

Unsupervised DA

Presence of source labels, no target labels

Semi-supervised DA

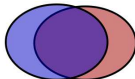
Presence of source labels, few target labels and a lot of unlabeled data

≠ Semi-supervised learning

No distribution shift, few labeled data and a lot of unlabeled data from the same domain

Some key points

- Estimating of the distribution shift

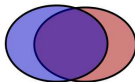


- Deriving generalization guarantees

$$R_{P_T}(h) \leq R_{P_S}(h) + ?$$

Some key points

- Estimating of the distribution shift



- Deriving generalization guarantees

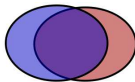
$$R_{P_T}(h) \leq ? R_{P_S}(h) ? + ?$$

- Characterizing when the adaptation is possible



Some key points

- Estimating of the distribution shift



- Deriving generalization guarantees

$$R_{P_T}(h) \leq ? R_{P_S}(h) ? + ?$$

- Characterizing when the adaptation is possible

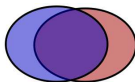


- Defining algorithms

Underlying idea: Try to move closer the two distributions

Some key points

- Estimating of the distribution shift



- Deriving generalization guarantees

$$R_{P_T}(h) \leq ? R_{P_S}(h) + ?$$

- Characterizing when the adaptation is possible



- Defining algorithms

Underlying idea: Try to move closer the two distributions

- Applying model selection principle

How to tune hyperparameters with no labeled information from target

3 main classes of algorithms

Reweighting/Instance-based methods

Correct a sample bias by reweighting source labeled data: source instances close to target instances are more important.



Feature-based methods/Find new representation spaces

Find a common space where source and target are close (projection, new features, etc)



Adjustment/Iterative methods

Modify the model by incorporating pseudo-labeled information



Reweighting/Instance based Methods

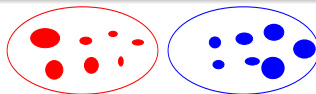
Context

Motivation

- Domains share the **same** support (*i.e.* bag of words)
- Distribution shift is caused by **sampling bias/shift between marginals**

Idea

Reweight or **select** instances to reduce the discrepancy between **source** and **target** domains.



A first analysis

$$R_{P_T}(h) = \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \mathbf{I}[h(\mathbf{x}^t) \neq y^t]$$

A first analysis

$$\begin{aligned} R_{P_T}(h) &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \frac{P_S(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \end{aligned}$$

A first analysis

$$\begin{aligned} R_{P_T}(h) &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \frac{P_S(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \sum_{(\mathbf{x}^t, y^t)} P_T(\mathbf{x}^t, y^t) \frac{P_S(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \end{aligned}$$

A first analysis

$$\begin{aligned} R_{P_T}(h) &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_T} \frac{P_S(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \sum_{(\mathbf{x}^t, y^t)} P_T(\mathbf{x}^t, y^t) \frac{P_S(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_S} \frac{P_T(\mathbf{x}^t, y^t)}{P_S(\mathbf{x}^t, y^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \end{aligned}$$

Covariate shift [Shimodaira,'00]

⇒ Assume similar tasks, $P_S(y|\mathbf{x}) = P_T(y|\mathbf{x})$, then:

$$\begin{aligned} &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_S} \frac{D_T(\mathbf{x}^t) P_T(y^t|\mathbf{x}^t)}{D_S(\mathbf{x}^t) P_S(y^t|\mathbf{x}^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_S} \frac{D_T(\mathbf{x}^t)}{D_S(\mathbf{x}^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \\ &= \mathbf{E}_{(\mathbf{x}^t) \sim D_S} \frac{D_T(\mathbf{x}^t)}{D_S(\mathbf{x}^t)} \mathbf{E}_{y^t \sim P_S(y^t|\mathbf{x}^t)} \mathbf{I}[h(\mathbf{x}^t) \neq y^t] \end{aligned}$$

⇒ **weighted error** on the **source domain**: $\omega(\mathbf{x}^t) = \frac{D_T(\mathbf{x}^t)}{D_S(\mathbf{x}^t)}$

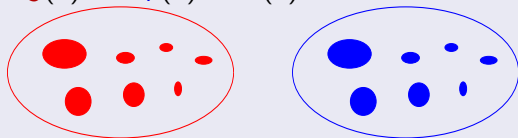
Idea reweight labeled **source** data according to an estimate of $\omega(\mathbf{x}^t)$:

$$\mathbf{E}_{(\mathbf{x}^t, y^t) \sim P_S} \omega(\mathbf{x}^t) \mathbf{I}[h(\mathbf{x}^t) \neq y^t]$$

Illustration

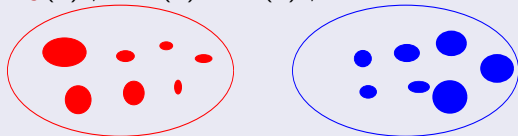
No Bias

$$D_S(\mathbf{x}) = D_T(\mathbf{x}) \Rightarrow \omega(\mathbf{x}) = 1$$



With Bias

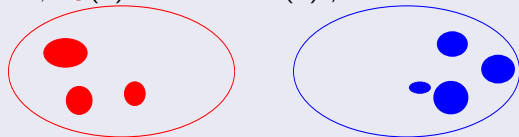
$$D_S(\mathbf{x}) \neq D_T(\mathbf{x}) \Rightarrow \omega(\mathbf{x}) \neq 1$$



Difficult case

No shared support

$\exists \mathbf{x}, D_S(\mathbf{x}) = 0$ and $D_T(\mathbf{x}) \neq 0$



Shared support

$D_S(\mathbf{x}) = 0$ if and only if $D_T(\mathbf{x}) = 0$

Intuition: the quality of the adaptation depends on the magnitude on the weights

How to deal with the sample selection bias?

Setting

A source sample $\mathcal{S} = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s}$ and a target sample $\mathcal{T} = \{\mathbf{x}_j^t\}_{j=1}^{m_t}$

Estimate new weights without using labels

$$\hat{w}(\mathbf{x}_i^s) = \frac{\hat{P}_{\mathcal{T}}(\mathbf{x}_i^s)}{\hat{P}_{\mathcal{S}}(\mathbf{x}_i^s)}$$

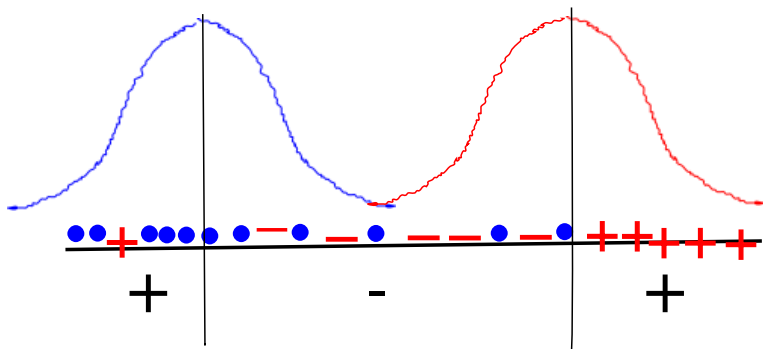
Learn a classifier on the classifier w.r.t. \hat{w}

$$\sum_{(\mathbf{x}_i^s, y_i^s) \in \mathcal{S}} \hat{w}(\mathbf{x}_i^s) I[h(\mathbf{x}_i^s) \neq y_i^s]$$

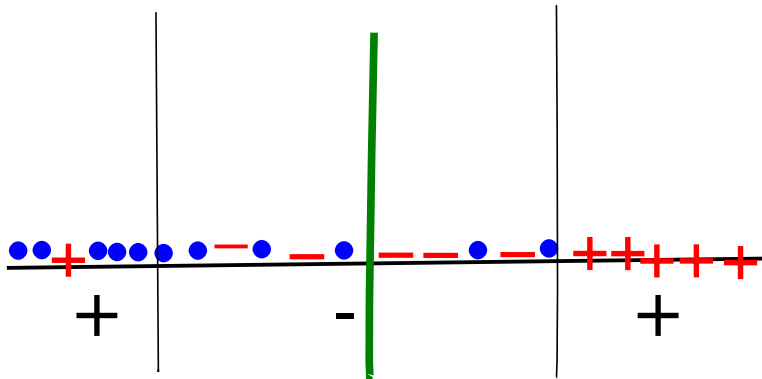
(Other losses: margin-based hinge-loss, least-square)



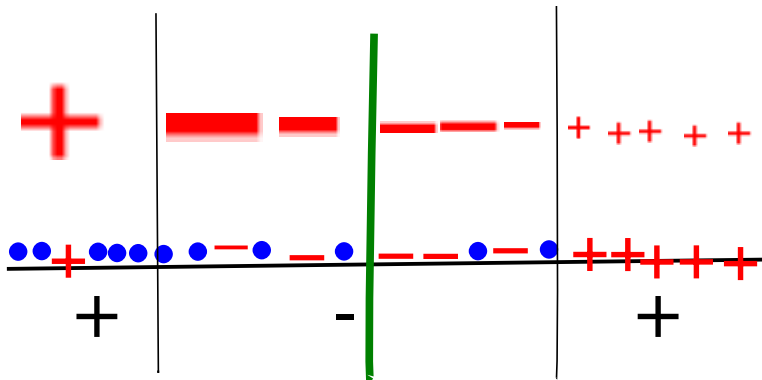
Illustration



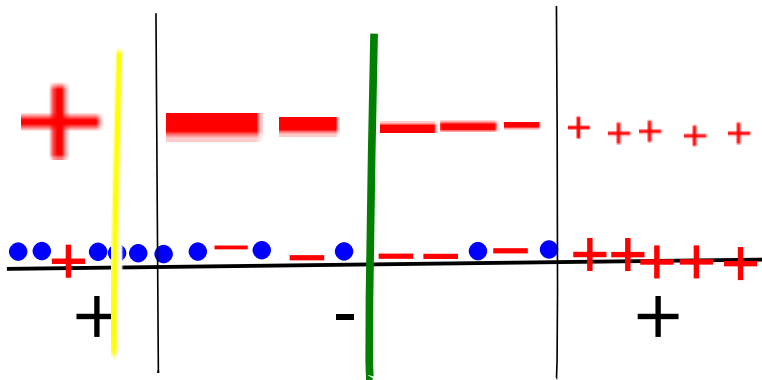
Illustration



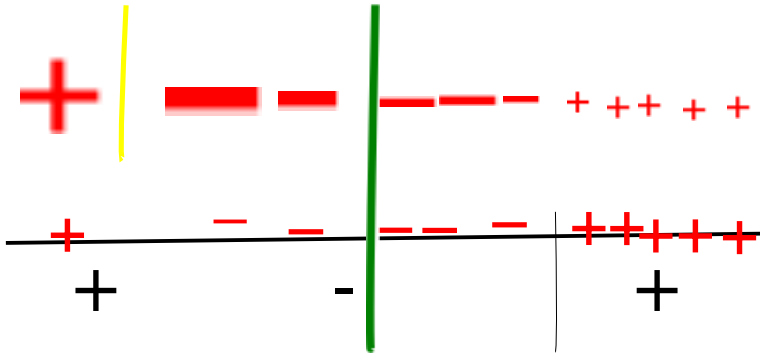
Illustration



Illustration



Illustration



Some existing approaches (1/2)

Density estimators

Build density estimators for source and target domains and estimate the ratio between them - Ex [Sugiyama et al., NIPS'07]:

- $\hat{\omega}(\mathbf{x}) = \sum_{l=1}^b \alpha_l \psi_l(\mathbf{x})$
- Learning: $\operatorname{argmin}_{\alpha} KL(\hat{\omega} D_S, D_T)$

Learn the weights discriminatively [Bickel et al., ICML'07]

- Assume $\frac{D_T(\mathbf{x}_i)}{D_S(\mathbf{x}_i)} \propto \frac{1}{p(q=1|\mathbf{x},\theta)}$
- Label **source** with label 1, **target** with label 0 and train a classifier ($\hat{\theta}$) to classify examples **1** or **0** (e.g. with logistic regression)
- Compute the new weights $\hat{\omega}(\mathbf{x}_i^s) = \frac{1}{p(q=1|\mathbf{x}_i^s; \hat{\theta})}$

Some existing approaches (2/2)

Kernel Mean Matching [Huang et al., NIPS'06]

- Maximum Mean Discrepancy

$$\text{MMD}(\mathcal{S}, \mathcal{T}) = \left\| \frac{1}{m_S} \sum_{i=1}^{m_S} \phi(\mathbf{x}_i^S) - \frac{1}{m_T} \sum_{j=1}^{m_T} \phi(\mathbf{x}_j^T) \right\|_H$$

- $\min_{\beta} \left\| \frac{1}{m_S} \sum_{i=1}^{m_S} \beta(\mathbf{x}_i^S) \phi(\mathbf{x}_i^S) - \frac{1}{m_S} \sum_{j=1}^{m_T} \phi(\mathbf{x}_j^T) \right\|_H$

$$\text{s.t. } \beta(\mathbf{x}_i^S) \in [0, B] \text{ and } \left| \frac{1}{m_S} \sum_{i=1}^{m_S} \beta(\mathbf{x}_i^S) - 1 \right| < \epsilon$$

- $\min_{\beta} \frac{1}{2} \beta^T \mathbf{K}_T \beta - \kappa_{\mathcal{S}, \mathcal{T}}^T \text{ s.t. } \beta_i \in [0, B] \text{ and } \left| \sum_{i=1}^{m_S} \beta(\mathbf{x}_i^S) - m_S \right| < m_S \epsilon$

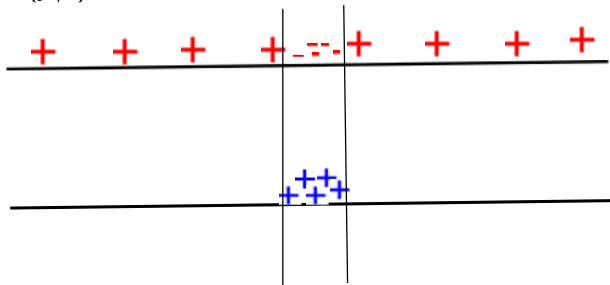
Guarantees [Gretton et al., 2008] - Under covariate shift assumptions

$$|R_{P_T}(h) - \text{weighted}(R_S(h))| < \sqrt{\frac{O(1/\delta) + O(\max_{\mathbf{x}} \omega(\mathbf{x})^2)}{m_S}} + C\epsilon \text{ and}$$

$$\left\| \frac{1}{m_S} \sum_{i=1}^{m_S} \omega(\mathbf{x}_i^S) \phi(\mathbf{x}_i^S) - \frac{1}{m_T} \sum_{j=1}^{m_T} \phi(\mathbf{x}_j^T) \right\| \leq O((1/\delta) \sqrt{\omega_{\max}^2 / m_S} + 1/m_T)$$

Bad news

- DA is hard, even under covariate shift [Ben-David et al., ALT'12]
⇒ To learn a classifier the number of examples depend on $|\mathcal{H}|$ (finite) or exponentially on the dimension of X
- Covariate shift assumption may fail: Tasks are not similar in general
 $P_S(y|\mathbf{x}) \neq P_T(y|\mathbf{x})$



- We did not consider the hypothesis space.
- Can define a general theory about DA?

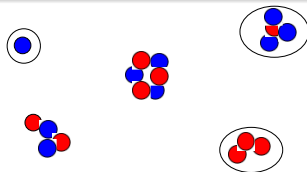
Theoretical frameworks for Domain Adaptation

A keypoint: estimating the distribution shift

First idea: Total variation measure

$$d_{L_1}(D_S, D_T) = \sup_{B \subseteq X} |D_S(B) - D_T(B)|$$

Subset of points maximizing the divergence



But:

- Not computable in general, and thus not estimable from finite samples

A keypoint: estimating the distribution shift

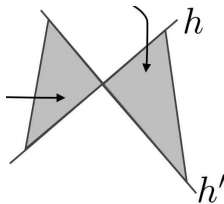
First idea: Total variation measure

$$d_{L_1}(D_S, D_T) = \sup_{B \subseteq X} |D_S(B) - D_T(B)|$$

Subset of points maximizing the divergence

But:

- Not computable in general, and thus not estimable from finite samples
- Not related to the hypothesis class
- Do not characterize the difficulty of the problem for \mathcal{H}



The $\mathcal{H}\Delta\mathcal{H}$ -divergence [Ben-David et al., NIPS'06; MLJ'10]

Definition

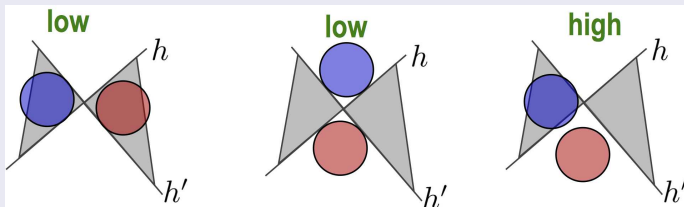
$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) &= \sup_{(h, h') \in \mathcal{H}^2} \left| R_{D_T}(h, h') - R_{D_S}(h, h') \right| \\ &= \sup_{(h, h') \in \mathcal{H}^2} \left| \mathbf{E}_{\mathbf{x}^t \sim D_T} \mathbf{I}[h(\mathbf{x}^t) \neq h'(\mathbf{x}^t)] - \mathbf{E}_{\mathbf{x}^s \sim D_S} \mathbf{I}[h(\mathbf{x}^s) \neq h'(\mathbf{x}^s)] \right| \end{aligned}$$

The $\mathcal{H}\Delta\mathcal{H}$ -divergence [Ben-David et al., NIPS'06; MLJ'10]

Definition

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) &= \sup_{(h, h') \in \mathcal{H}^2} \left| R_{D_T}(h, h') - R_{D_S}(h, h') \right| \\ &= \sup_{(h, h') \in \mathcal{H}^2} \left| \mathbf{E}_{\mathbf{x}^t \sim D_T} \mathbf{I}[h(\mathbf{x}^t) \neq h'(\mathbf{x}^t)] - \mathbf{E}_{\mathbf{x}^s \sim D_S} \mathbf{I}[h(\mathbf{x}^s) \neq h'(\mathbf{x}^s)] \right| \end{aligned}$$

Illustration with **only 2 hypothesis in \mathcal{H}** h and h'



Note: With a larger \mathcal{H} , the distance will be **high** since we can easily find two hypothesis able to **distinguish** the two domains

Computable from samples

Consider two samples \mathcal{S} , \mathcal{T} of size m from $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) \leq d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + O(\text{complexity}(\mathcal{H})\sqrt{\frac{\log(m)}{m}})$$

complexity(\mathcal{H}): VC-dimension [Ben-david et al.,'06;'10], Rademacher [Mansour et al.,'09]

Empirical estimation

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \left(1 - \min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{\mathbf{x}: h(\mathbf{x})=-1} I[\mathbf{x} \in \mathcal{S}] + \frac{1}{m} \sum_{\mathbf{x}: h(\mathbf{x})=1} I[\mathbf{x} \in \mathcal{T}] \right] \right)$$

\Rightarrow Already seen: label **source** examples as -1, **target** ones as +1 and try to learn a classifier in \mathcal{H} minimizing the associated empirical error



Going to a generalization bound

Preliminaries

- $R_{P_T}(h, h') = \mathbf{E}_{(\mathbf{x}, y) \sim P_S} I[h(\mathbf{x}) \neq h'(\mathbf{x})] = \mathbf{E}_{\mathbf{x} \sim D_T} I[h(\mathbf{x}) \neq h'(\mathbf{x})]$
 $R_{P_T}(R_{P_S})$ fulfills the triangle inequality
- $|R_{P_T}(h, h') - R_{P_S}(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$
since $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) = 2 \sup_{(h, h') \in \mathcal{H}^2} |R_{D_T}(h, h') - R_{D_S}(h, h')|$
- $h_S^* = \operatorname{argmin}_{h \in \mathcal{H}} R_{P_S}(h)$: best on **source**
- $h_T^* = \operatorname{argmin}_{h \in \mathcal{H}} R_{P_T}(h)$: best on **target**

Ideal joint hypothesis

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}} R_{P_S}(h) + R_{P_T}(h) ; \lambda = R_{P_S}(h^*) + R_{P_T}(h^*)$$

A first bound

$$R_{P_T}(h) \leq$$

A first bound

$$R_{P_T}(h) \leq R_{P_T}(h^*) + R_{P_T}(h, h^*)$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \end{aligned}$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + |R_{P_T}(h, h^*) - R_{P_S}(h, h^*)| \end{aligned}$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + |R_{P_T}(h, h^*) - R_{P_S}(h, h^*)| \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \end{aligned}$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + |R_{P_T}(h, h^*) - R_{P_S}(h, h^*)| \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h) + R_{P_S}(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \end{aligned}$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + |R_{P_T}(h, h^*) - R_{P_S}(h, h^*)| \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h) + R_{P_S}(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \lambda \end{aligned}$$

A first bound

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h^*) + R_{P_T}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + R_{P_T}(h, h^*) - R_{P_S}(h, h^*) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + |R_{P_T}(h, h^*) - R_{P_S}(h, h^*)| \\ &\leq R_{P_T}(h^*) + R_{P_S}(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq R_{P_T}(h^*) + R_{P_S}(h) + R_{P_S}(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \\ &\leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \lambda \\ &\leq R_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(S, T) + O(\text{complexity}(\mathcal{H}) \sqrt{\frac{\log(m)}{m}}) + \lambda \end{aligned}$$

Main theoretical bound

Theorem [Ben-David et al., MLJ'10, NIPS'06]

Let \mathcal{H} a symmetric hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in (0, 1]$, with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H}, \quad R_{P_T}(h) \leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \lambda$$

Formalizes a natural approach: Move closer the two distributions while ensuring a low error on the source domain.

Justifies many algorithms:

- reweighting methods,
- feature-based methods,
- adjusting/iterative methods.

Another analysis [Mansour et al.,COLT'09]

$$R_{P_T}(h) \leq$$

Another analysis [Mansour et al., COLT'09]

$$R_{P_T}(h) \leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*)$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \end{aligned}$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*) + \nu \end{aligned}$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + |R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*)| + \nu \end{aligned}$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + |R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*)| + \nu \\ &\leq R_{P_S}(h, h_S^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu \end{aligned}$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + |R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*)| + \nu \\ &\leq R_{P_S}(h, h_S^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu \\ &(\leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + R_{P_S}(h_S^*) + \nu) \text{ if } h_S^* \text{ is not the true lab} \end{aligned}$$

Another analysis [Mansour et al., COLT'09]

$$\begin{aligned} R_{P_T}(h) &\leq R_{P_T}(h, h_S^*) + R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*) \\ &= R_{P_T}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*) + \nu \\ &\leq R_{P_S}(h, h_S^*) + |R_{P_T}(h, h_S^*) - R_{P_S}(h, h_S^*)| + \nu \\ &\leq R_{P_S}(h, h_S^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu \\ &(\leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + R_{P_S}(h_S^*) + \nu) \text{ if } h_S^* \text{ is not the true lab} \end{aligned}$$

- This analysis can lead to smaller when adaptation is possible
- Leads to the same type of bound, just the constant changes

Characterization of the possibility of domain adaptation



Constants characterize when adaptation is possible

- $\lambda = R_{P_S}(h^*) + R_{P_T}(h^*)$, $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R_{P_S}(h) + R_{P_T}(h)$
There must exist an ideal joint hypothesis with small error
- $\nu = R_{P_T}(h_S^*, h_T^*) + R_{P_T}(h_T^*)$
there must exist a very good hypothesis on the target and the best hypothesis on source must be close to the best on target w.r.t to D_T

Other settings

Discrepancy [Mansour et al.,COLT'09]

- instead of the 0-1 loss, more general loss functions ℓ (i.e. L_p norms)
 $\text{disc}^\ell(D_S, D_T) = \sup_{h, h' \in \mathcal{H}} |R_{D_S}^\ell(h, h') - R_{D_T}^\ell(h, h')|$
- This discrepancy can be minimized and used as a reweighting method ([Mansour et al.,COLT'09] - polynomial for L_2 norm for example)

Using some target labeled data

- Weighting the empirical source and target risks [Ben David et al.,2010]
- Using a divergence taking into account target labels [Zhang et al.,NIPS'12] (a divergence must take into account marginals over X and Y , the λ constant counts for Y)

Other settings

Averaged quantities [Germain et al., ICML'13]

- Consider a probability distribution ρ (posterior) over \mathcal{H} to learn and the following risk: $\mathbf{E}_{h \sim \rho} R_{P_T}(h)$

- Definition of an averaged distance

$$\text{dis}(D_S, D_T) = \left| \mathbf{E}_{h, h' \sim \rho^2} [R_{D_T}(h, h') - R_{D_S}(h, h')] \right|$$

- Similar generalization bound

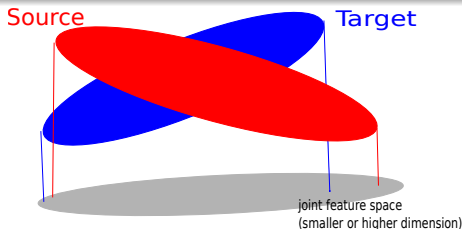
$$\mathbf{E}_{h \sim \rho} R_{P_T}(h) \leq \mathbf{E}_{h \sim \rho} R_{P_S}(h) + \text{dis}(D_S, D_T) + \lambda_{\rho^*}$$

- Estimation from samples controlled by PAC-Bayesian theory
- Bound tighter without a supremum

Feature/Projection based Approaches

Idea

- Change the feature representation X to better represent shared characteristics between the two domains
 - some features are domain-specific,
 - others are generalizable
 - or there exist mappings from the original space
- \Rightarrow Make source and target domain explicitly similar
- \Rightarrow Learn a new feature space by embedding or projection



Metric Learning [Kulis et al., '11; Saenko et al., '10]

- Mahalanobis: $d_{\mathbf{W}}^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \mathbf{W} (\mathbf{x} - \mathbf{x}')$
- PSD matrix $\mathbf{W} = L^T L$,
 L projection space of dimension $\mathbb{R}^{rank(\mathbf{W}) \times d}$
 $(L\mathbf{x} - L\mathbf{x}')^T (L\mathbf{x} - L\mathbf{x}')$
- Pair-wise constraints: source ex. (\mathbf{x}_i^s, y_i^s) and target (\mathbf{x}_j^t, y_j^t)
 - $d_{\mathbf{W}}^2(\mathbf{x}_i^s, \mathbf{x}_j^t) \leq u$ if $y_i^s = y_j^t$ (source and target must be similar)
 - $d_{\mathbf{W}}^2(\mathbf{x}_i^s, \mathbf{x}_j^t) \geq l$ if $y_i^s \neq y_j^t$ (source and target must be dissimilar)
 - Require some target labels

Metric Learning [Kulis et al., CVPR'11; Saenko et al., ECCV'10]



(a) Domain shift problem



(b) Pairwise constraints



(c) Invariant space

[Saenko et al., ECCV'10]

Formulation (based on ITML [Davis et al., ICML'07])

$$\min_{\mathbf{W} \succeq 0} \quad \text{Tr}(\mathbf{W}) - \log \det \mathbf{W}$$

$$\text{s.t.} \quad d_{\mathbf{W}}^2(\mathbf{x}_i^s, \mathbf{x}_j^t) \leq u, \forall (\mathbf{x}_i^s, \mathbf{x}_j^t) \in \text{SimilarSet}$$
$$d_{\mathbf{W}}^2(\mathbf{x}_i^s, \mathbf{x}_j^t) \geq l, \forall (\mathbf{x}_i^s, \mathbf{x}_j^t) \in \text{DissimilarSet}$$

⇒ Can be kernelized

(Simple) Feature augmentation [Daume III et al., '07;'10]

- $\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}, 0 \rangle$ for source instances
- $\phi(\mathbf{x}) = \langle \mathbf{x}, 0, \mathbf{x} \rangle$ for target instances
- \Rightarrow Share some relevant features and not irrelevant ones (e.g. in text sentiment analysis: find shared words)
 \Rightarrow a way to allow the existency of the ideal joint hypothesis h^*

Learn in the new space ϕ

- Require target labels
- Bound: $R_{D_T} \leq \frac{1}{2}(R_T + R_S) + O(\text{complexity}) + (\frac{1}{m_s} + \frac{1}{m_t})O(\frac{1}{\delta}) + O(d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T))$
- Kernelized and semi-supervised versions [add: $(+1, \langle 0, \mathbf{x}, -\mathbf{x} \rangle)$ and $(-1, \langle 0, \mathbf{x}, -\mathbf{x} \rangle)$ to learning sample]

Find latent spaces - Structural Correspondence Learning [Blitzer et al., '07]

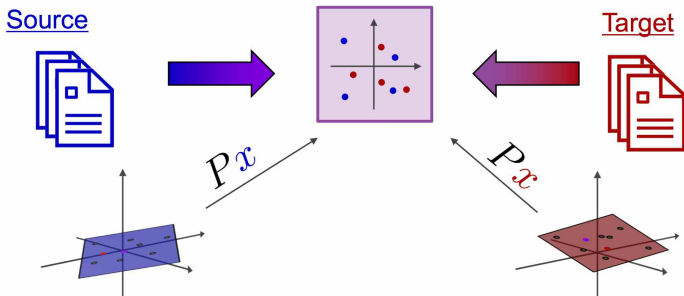
Identify shared features

Domains	Negative	Positive
Books	plot <num>_pages predictable reading_this page_<num>	reader grisham engaging must_read fascinating
Kitchen	the_plastic poorly_designed leaking awkward_to defective	excellent_product espresso are_perfect years_now a_breeze
Pivot features	weak don't_waste awful	and_easy loved_it a_wonderful a_must highly_recommended

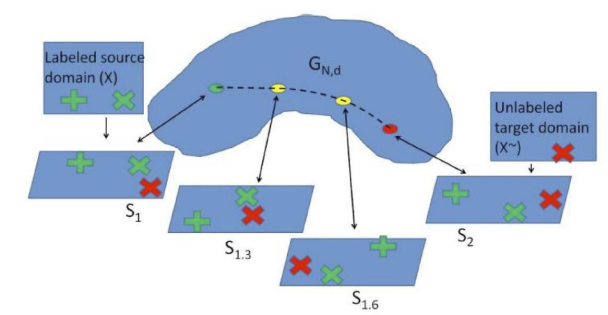
- Sentiment analysis - Bag of words (bigrams)
- Choose K **pivot** features (frequent words in both domains, highly correlated with labels)
- Learn K classifiers to predict pivot features from remaining features
- For each feature add K new features
- Represents source and target data with these features

Find latent spaces - Structural Correspondence Learning [Blitzer et al., '07]

- Apply PCA source+target new features to get a low rank latent representation
- Learn a classifier in the new projection space defined by PCA

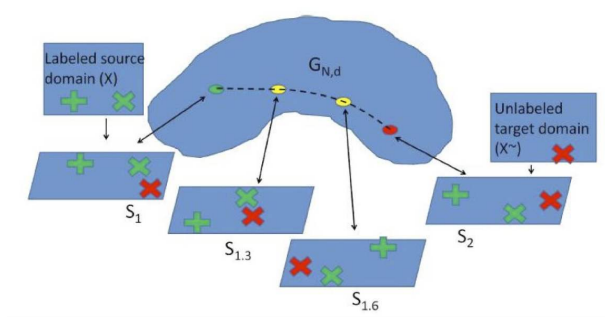


Manifold-based methods



- Assume $X \subseteq \mathbb{R}^N$
- Apply PCA on source data \Rightarrow matrix \mathbf{S}_1 of rank d
- Apply PCA on target data \Rightarrow matrix \mathbf{S}_2 of rank d
- Geodesic path on the Grassman manifold $\mathbb{G}_{N,d}$ (d -dimensional vector subspaces $\subset \mathbb{R}^N$) between \mathbf{S}_1 and \mathbf{S}_2

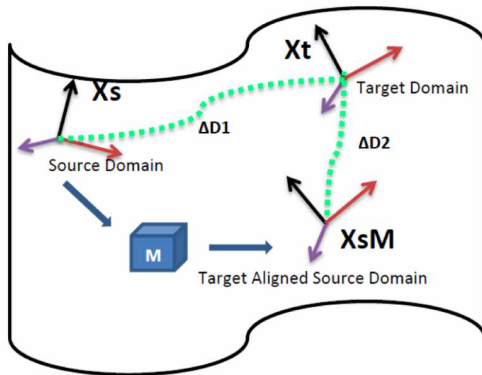
Manifold-based methods



[Gopalan et al., '10]

- Use of an exponential flow $\psi(t') = \mathbf{Q} \exp(t' \mathbf{B}) \mathbf{J}$ with \mathbf{Q} $N \times N$ matrix with determinant 1 s.t. $\mathbf{Q}^T \mathbf{S}_1 = \mathbf{J}$ and $\mathbf{J}^T = [\mathbf{I}_d \mathbf{0}_{N-d,d}]$ intermediate subspaces are obtained by computing \mathbf{B} (skew block-diagonal matrix) and varying t' between 0 and 1
- Take a collection S' of l subspaces between \mathbf{S}_1 and \mathbf{S}_2 on the manifold
- Project the data on \mathbf{S}' and learn in that new space

A simpler approach - Subspace alignment [Fernando et al., ICCV'13]



- Move closer PCA-based representations
- Totally unsupervised

Subspace alignment algorithm

Algorithm 1: Subspace alignment DA algorithm

Data: Source data \mathbf{S} , Target data \mathbf{T} , Source labels Y_S , Subspace dimension d

Result: Predicted target labels Y_T

$\mathbf{S}_1 \leftarrow \text{PCA}(\mathbf{S}, d)$ (source subspace defined by the first d eigenvectors) ;

$\mathbf{S}_2 \leftarrow \text{PCA}(\mathbf{T}, d)$ (target subspace defined by the first d eigenvectors);

$\mathbf{X}_a \leftarrow \mathbf{S}_1 \mathbf{S}_1' \mathbf{S}_2$ (operator for aligning the source subspace to the target one);

$\mathbf{S}_a = \mathbf{S} \mathbf{X}_a$ (new source data in the aligned space);

$\mathbf{T}_T = \mathbf{T} \mathbf{S}_2$ (new target data in the aligned space);

$Y_T \leftarrow \text{Classifier}(\mathbf{S}_a, \mathbf{T}_T, Y_S)$;

- $\mathbf{M}^* = \mathbf{S}_1' \mathbf{S}_2$ corresponds to the “subspace alignment matrix”:
 $\mathbf{M}^* = \text{argmin}_{\mathbf{M}} \|\mathbf{S}_1 \mathbf{M} - \mathbf{S}_2\|$
- $\mathbf{X}_a = \mathbf{S}_1 \mathbf{S}_1' \mathbf{S}_2 = \mathbf{S}_1 \mathbf{M}^*$ projects the source data to the target subspace
- A natural similarity: $\text{Sim}(\mathbf{x}_s, \mathbf{x}_t) = \mathbf{x}_s \mathbf{S}_1 \mathbf{M}^* \mathbf{S}_1' \mathbf{x}_t' = \mathbf{x}_s \mathbf{A} \mathbf{x}_t'$

Some results



- Adaptation from Office/Caltech-10 datasets (four domains to adapt) is used as source and one as target
- Comparisons
 - Baseline 1: projection on the source subspace
 - Baseline 2: projection on the target subspace
 - 2 related methods : GFK [Gong et al., CVPR'12] and GFS [Gopalan et al., ICCV'11]

Some results

- Office/Caltech-10 datasets with 4 domains A, B, C, D

Method	C→A	D→A	W→A	A→C	D→C	W→C
NA	21.5	26.9	20.8	22.8	24.8	16.4
Baseline 1	38.0	29.8	35.5	30.9	29.6	31.3
Baseline 2	40.5	33.0	38.0	33.3	31.2	31.9
GFS [8]	36.9	32	27.5	35.3	29.4	21.7
GFK [7]	36.9	32.5	31.1	35.6	29.8	27.2
OUR	39.0	38.0	37.4	35.3	32.4	32.3

Method	A→D	C→D	W→D	A→W	C→W	D→W
NA	22.4	21.7	40.5	23.3	20.0	53.0
Baseline 1	34.6	37.4	71.8	35.1	33.5	74.0
Baseline 2	34.7	36.4	72.9	36.8	34.4	78.4
GFS [8]	30.7	32.6	54.3	31.0	30.6	66.0
GFK [7]	35.2	35.2	70.6	34.4	33.7	74.9
OUR	37.6	39.6	80.3	38.6	36.8	83.6

Table 2. Recognition accuracy with unsupervised DA using a NN classifier (Office dataset + Caltech10).

Method	C→A	D→A	W→A	A→C	D→C	W→C
Baseline 1	44.3	36.8	32.9	36.8	29.6	24.9
Baseline 2	44.5	38.6	34.2	37.3	31.6	28.4
GFK	44.8	37.9	37.1	38.3	31.4	29.1
OUR	46.1	42.0	39.3	39.9	35.0	31.8

Method	A→D	C→D	W→D	A→W	C→W	D→W
Baseline 1	36.1	38.9	73.6	42.5	34.6	75.4
Baseline 2	32.5	35.3	73.6	37.3	34.2	80.5
GFK	37.9	36.1	74.6	39.8	34.9	79.1
OUR	38.8	39.4	77.9	39.6	38.9	82.3

Table 3. Recognition accuracy with unsupervised DA using a SVM classifier (Office dataset + Caltech10).

- Divergences

Method	NA	Baseline 1	Baseline 2	GFK	SA
TDAS	1.25	3.34	2.74	2.84	4.26
H4H	98.1	99.0	99.0	74.3	53.2

Feature-based method

- Feature-based approaches are very popular
Many other (SVM/kernel-based, MKL, deep learning [Glorot et al., ICML'11], ...) methods not covered here,
- Subspace-based methods \Rightarrow "hot topic"
- Embed similarity map: define feature as similarity to landmarks points
- labeled source instances distributed similarly to the target domain



[Grauman, VisDA-WS ICCV'13] \rightarrow subsampling: work with instances facilitating adaptation

or use distances to headphones as a representation

$\langle k(\cdot, \mathbf{x}_1), k(\cdot, \mathbf{x}_2), k(\cdot, \mathbf{x}_3), k(\cdot, \mathbf{x}_4), k(\cdot, \mathbf{x}_5), \dots \rangle, \dots$

Adjusting/Iterative methods

Principle

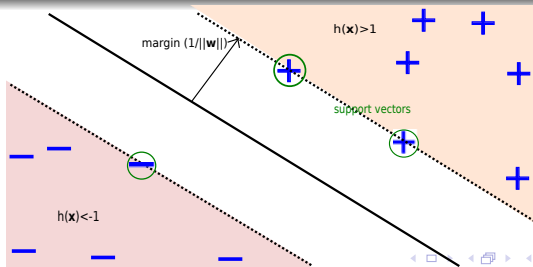
- Integrate some information about the target samples iteratively
⇒ use of pseudo-labels
- “Move” closer distributions
⇒ Remove/add some instances ⇒ take into account a divergence measure
- Repeat the process until convergence or no remaining instances

DASVM [Bruzzone et al., '10]

A brief recap on SVM

- Learning sample $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Learn a classifier $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$

$$\begin{aligned} \text{Formulation: } \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to } & \ell_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ & \xi_i \geq 0 \end{aligned}$$

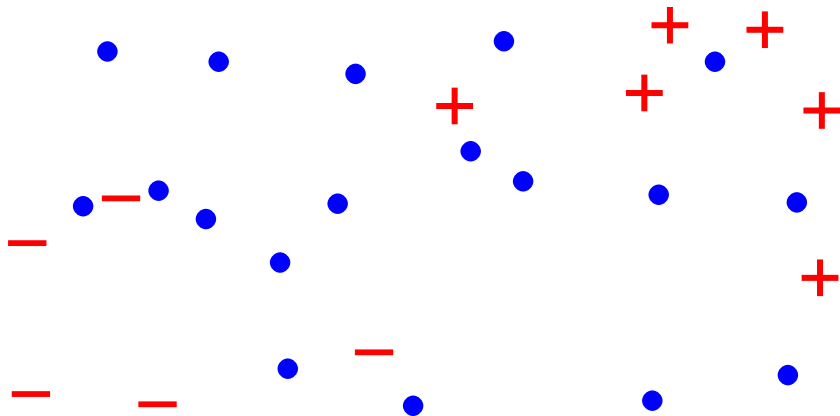


DASVM principle

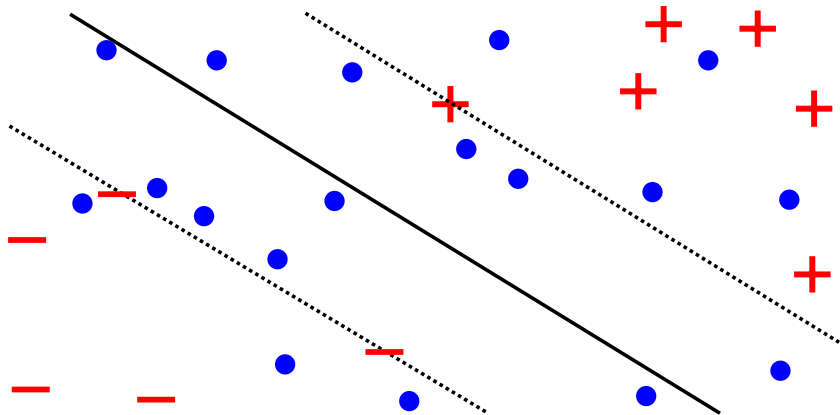
- 1 $LS = S$
- 2 Learn a classifier h^0 from the learning sample LS
- 3 Repeat until stopping criterion
 - Select the first k target examples \mathbf{x}^t s.t. $0 < h(\mathbf{x}^t) < 1$ with highest margin and affect the pseudo-label -1
 - Select the first k target examples \mathbf{x}^t s.t. $-1 < h(\mathbf{x}^t) < 0$ with highest margin and affect them the pseudo-label $+1$
 - Add these $2k$ examples (pseudo-labeled) to LS
 - Remove from LS the first k positive and k negative source instances with highest margin
- 4 Output the last classifier

Algorithm stops when the number of selected instances at each step downs to a threshold.

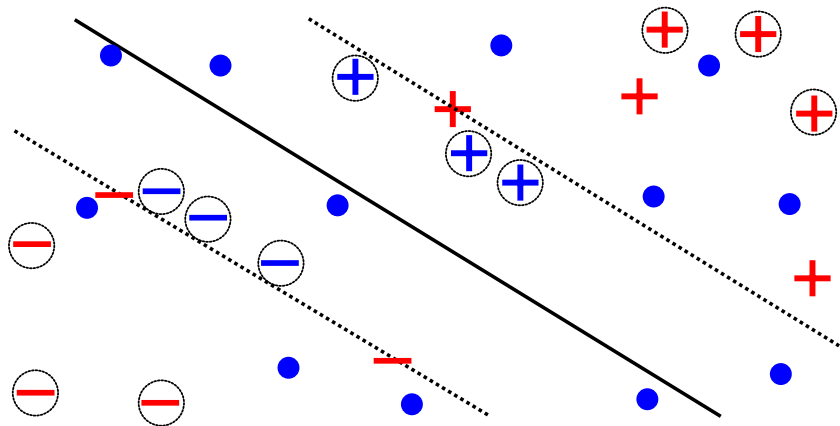
DASVM - graphical illustration



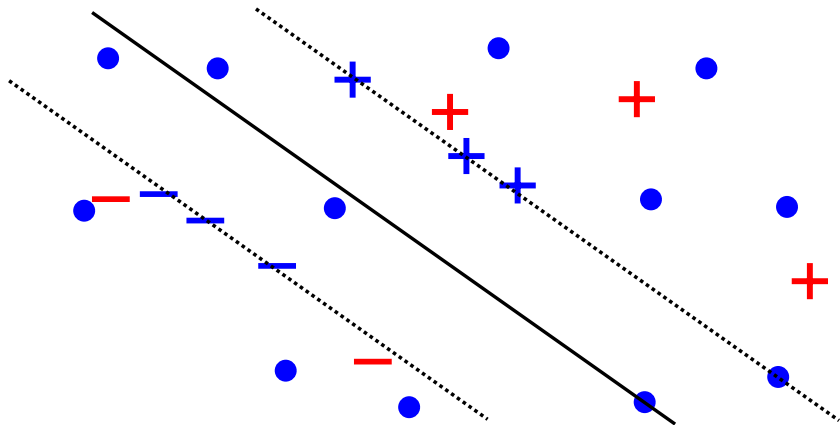
DASVM - graphical illustration



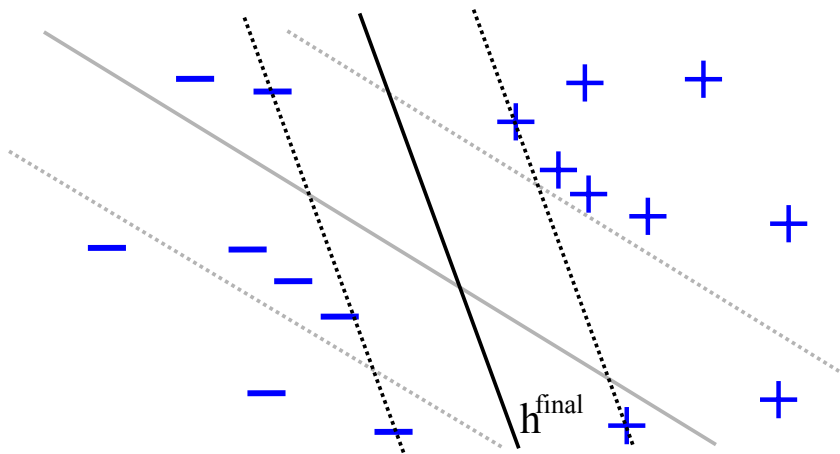
DASVM - graphical illustration



DASVM - graphical illustration



DASVM - graphical illustration



Convergence - theoretical guarantees

- What we need?: At each step pseudo-labels on target are sufficiently reasonable.
- \Rightarrow Can be tackled with a notion of weak classifier

Weak classifier on a single domain

An hypothesis h^n , learned at iteration n , is a weak learner over a labeled sample S if it performs a bit better than a random guessing: $\exists \gamma_n \in]0; \frac{1}{2}$

$$R_S(h^n) = \hat{P}_{r_{(\mathbf{x}_i^s, y_i) \sim S^n}}[h^n(\mathbf{x}_i) \neq y_i] = \frac{1}{2} - \gamma_n$$

A notion of weak learner for controlling pseudo-labels

Self labeling weak learner

A classifier $h^{(i)}$ learned at iteration over a current learning sample LS^i is self labeling weak learner w.r.t. a set SL^j of $2k$ pseudo-labeled examples introduced at step j if its true error over SL^j is strictly better than random guessing:

$$R_{SL^j}(h^{(i)}) = Pr_{\mathbf{x}_i^t \in SL^j}[h(\mathbf{x}_i^t) \neq y_i^t] < \frac{1}{2}$$

A first necessary condition

Theorem

Let $h^{(i)}$ a weak learner output at iteration i from $LS^{(i)}$, let $\tilde{R}_{LS^i}(h^{(i)}) = \frac{1}{2} - \gamma_{LS}^i$ the associated empirical error.

Let $R_T^{(i)} = \frac{1}{2} - \gamma_T^{(i)}$ the true empirical error over T .

$h^{(i)}$ is a self-labeling weak learner if $\gamma_{LS}^i > 0$

$\Rightarrow h^{(i)}$ will be able to correctly classify (w.r.t. their unknown true label) more than k pseudo-labeled target examples among $2k$ if at least half of them have been correctly pseudo-labeled.

A second result

Theorem

Let S a labeled source sample of m_S instances and T a target sample of $m_t \geq m_S$ unlabeled instances. Let \mathcal{A} an iterative labeling algorithm using $2k$ examples at each step. \mathcal{A} is able to perform an adaptation if

- $\gamma_S^{(i)} \geq \gamma_T^{(i)}, \forall i = 1 \dots \frac{m_S}{2k}$
- $\gamma_S^{\max} > \sqrt{\frac{\gamma_T^{(0)}}{2}}$

$\Rightarrow h^{(i)}$ has to perform sufficiently well on the data it has been learned from
 $\Rightarrow \mathcal{A}$ the final classifier has to work better on T than a classifier learned only from source data.

A simple illustration

Task: handwritten digit recognition. P_1 scaling problem between 1 and 0 and P_2 rotation problem between 5 and 7.

Iteration	P_1			P_2		
	$\gamma_S^{(i)}$	$\gamma_T^{(i-1)}$	$1 - \hat{\epsilon}_T^{(i)}$	$\gamma_S^{(i)}$	$\gamma_T^{(i-1)}$	$1 - \hat{\epsilon}_T^{(i)}$
1	0.5	0	0.585	0.50	-0.1	0.32
2	0.475	0.085	0.75	0.50	-0.18	0.285
3	0.48	0.25	0.73	0.50	-0.215	0.285
4	0.49	0.23	0.795	0.50	-0.215	0.24
5	0.49	0.295	0.875	0.50	-0.26	0.18
6	0.49	0.375	0.94	0.50	-0.32	0.205
7	0.49	0.44	0.94	0.50	-0.295	0.19
8	0.49	0.44	0.94	0.50	-0.31	0.12
9	0.49	0.44	0.94	0.50	-0.38	0.145
10	0.495	0.44	0.985	0.50	-0.355	0.115
11	0.5	0.485	0.99	0.495	-0.385	0.115

For P_1 , we can check $\gamma_S^{(i)} \geq \gamma_T^{(i)}, \forall i = 1 \dots \frac{m_S}{2k}$, and $\gamma_S^{\max} > \sqrt{\frac{\gamma_T^{(0)}}{2}}$.

Interpretation summary

- $h^{(i)}$ must work well on T
- $h^{(i)}$ must work well on S
- \mathcal{A} works better than a non adaptation process

⇒ Necessary and reasonable conditions
⇒ Condition on T hard to check in practise
⇒ Boosting

Ensemble Methods and Boosting

Definition

Ensemble methods infer a set of classifiers h_1, \dots, h_N whose individual decisions are combined in some way to classify new examples.

Necessary conditions for an ensemble method to be efficient

- the individual classifiers are better than random guessing.
- they are diverse, *i.e.* they make different errors on new data points.

ADABOOST

- Learns step by step **weak** binary classifiers.
- **Optimizes a convex loss** by increasing the weights of misclassified examples.
- Builds a **convex combination** of the weak classifiers.

ADABOOST

Data: A learning sample S , a number of iterations N , a weak learner L

Result: A global hypothesis H_N

for $i = 1$ **to** m **do** $D_1(\mathbf{x}_i) = 1/m$;

for $t = 1$ **to** T **do**

$h_n = \text{LEARN}(S, \mathbf{D}_n)$;

$\hat{\epsilon}_n = \sum_{\mathbf{x}_i \text{ s.t. } y_i \neq h_n(\mathbf{x}_i)} D_n(\mathbf{x}_i)$;

$\alpha_n = \frac{1}{2} \ln \frac{1 - \hat{\epsilon}_n}{\hat{\epsilon}_n}$;

for $i = 1$ **to** m **do**

$D_{t+1}(\mathbf{x}_i) = D_n(\mathbf{x}_i) \exp(-\alpha_t y_i h_t(\mathbf{x}_i)) / Z_n$;

 /* Z_n is a normalization coefficient */

end

end

$f(\mathbf{x}) = \sum_{t=1}^T \alpha_t h_t(\mathbf{x})$;

$H_N(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$;

Theoretical result on the empirical error

Theorem

Upper bound on the empirical error of the final classifier H_N

$$\hat{\epsilon}_{H_N} \leq \prod_{t=1}^T Z_t \leq \exp(-2 \sum_{t=1}^T \gamma_t^2)$$

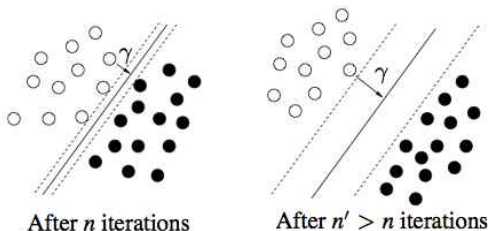
where $\hat{\epsilon}_n = \frac{1}{2} - \gamma_n$ (weak hypothesis).

$\hat{\epsilon}_{H_N}$ is optimized with $\alpha_n = \frac{1}{2} \ln \frac{1-\hat{\epsilon}_n}{\hat{\epsilon}_n}$.

Meaning

This theorem means that the empirical error **exponentially decreases towards 0** with the number T of iterations.

Explanation in terms of margins of the training examples



Theorem

$\forall \gamma > 0$, with probability $1 - \delta$, any classifier ensemble H_T satisfies:

$$\epsilon_{H_T} \leq \mathbb{E}_{\mathbf{x} \in S}[\text{margin}(\mathbf{x}) \leq \gamma] + \mathcal{O} \left(\sqrt{\frac{d_h \log^2(m/d_h)}{m \gamma^2}} + \log(1/\delta) \right),$$

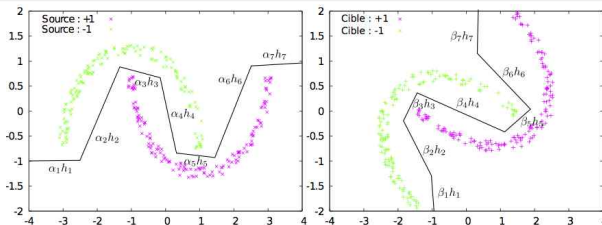
where $\mathbb{E}_{\mathbf{x} \in S}[\text{margin}(\mathbf{x}) \leq \gamma]$ exponentially decreases towards 0 with T .

\Rightarrow apply this idea on source and target (pseudo-labels)

Idea for domain adaptation

Double weighting strategy

- Keep the same weak hypothesis for both domains $h_1, \dots, h_n, \dots, h_N$
- Learn two functions
 - Source domain: $F_S^N = \sum_{n=1}^N \alpha_n h_n(\mathbf{x})$
 - Target domain: $F_T^N = \sum_{n=1}^N \beta_n h_n(\mathbf{x})$
- β_n depends on the (pseudo-)margin of the examples and a divergence measure



A notion of weak DA learner

Weak DA learner

A classifier h_n learned at iteration n from a S and T and a divergence $g_n \in [0, 1]$ between S and T and $f_{DA}(h_n(\mathbf{x}_i)) = |h_n(\mathbf{x}_i)| - \lambda g_n$, is a weak DA learner for T if:

- 1 h_n is a weak learner for S .
- 2 $\hat{L}_n = \mathbb{E}_{\mathbf{x}_i \sim T}[|f_{DA}(h_n(\mathbf{x}_i))| \leq \gamma] < \frac{\gamma}{\gamma + \max(\gamma, \lambda g_n)}$

- f_{DA} : obtaining high margin with small divergence
- if $\max(\gamma, \lambda g_n) = \gamma$: divergence is small, we are close to a semi-supervised setting
- if $\max(\gamma, \lambda g_n) = \lambda g_n$: divergence is high and the reweighting scheme requires a specific attention to the divergence.

Algorithm SLDAB

SLDAB

Data: Learning sample \mathcal{S} , Nb of iterations N , unlabeled sample \mathcal{T} , $\gamma \in [0, 1]$, $\lambda \in [0, 1]$

Result: Source and target hypothesis $H_{\mathcal{S}}$, $H_{\mathcal{T}}$

foreach $\forall (\mathbf{x}_i^s, y_i^s) \in \mathcal{S}$, $\mathbf{x}_j^t \in \mathcal{T}$ **do** $D_1^{\mathcal{S}}(\mathbf{x}_i^s) = 1/m_s$; $D_1^{\mathcal{T}}(\mathbf{x}_j^t) = 1/m_t$;

for $n = 1$ **to** N **do**

Learn h_n to produce a **weak DA** learner; **compute** g_n ;

$$\alpha_n = \frac{1}{2} \ln \frac{1 - \hat{\epsilon}_{\mathcal{S}^n}(h^n)}{\hat{\epsilon}_{\mathcal{S}^n}(h^n)}; \beta_n = \frac{1}{\gamma + \max(\gamma, \lambda g_n)} \ln \frac{\gamma W_n^+}{\max(\gamma, \lambda g_n) W_n^-};$$

for $(\mathbf{x}_i^s, y_i^s) \in \mathcal{S}$ **do** $D_{n+1}(\mathbf{x}_i) = D_n(\mathbf{x}_i^s) \exp(-\alpha_n y_i^s \text{sign}(h_n(\mathbf{x}_i^s))) / Z_{\mathcal{S}}^n$;

for $\mathbf{x}_j^t \in \mathcal{T}$ **do**

$$D_{\mathcal{T}}^{n+1}(\mathbf{x}_j) = D_{\mathcal{T}}^n(\mathbf{x}_j^t) \exp(-\beta_n y_j^n f_{DA}(h_n(\mathbf{x}_j^t))) / Z_{\mathcal{T}}^n;$$

where $y_j^n = -\text{sign}(f_{DA}(h_n(\mathbf{x})))$ if $|f_{DA}(h_n(\mathbf{x}))| > \gamma$ and $y_j^n = -\text{sign}(f_{DA}(h_n(\mathbf{x})))$ otherwise;

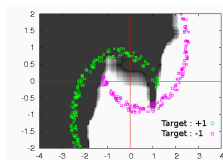
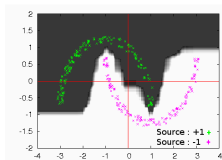
end

end

$$F_{\mathcal{S}}^N(\mathbf{x}^s) = \sum_{n=1}^N \alpha_n \text{sign}(h_n(\mathbf{x}^s));$$

$$H_{\mathcal{T}}^N(\mathbf{x}^t) = \sum_{n=1}^N \beta_n \text{sign}(h_n(\mathbf{x}^t));$$

Conclusion



Theoretical results

- Convergence of the empirical losses (exponentially fast to 0) \Rightarrow (pseudo-)margin increases
- No generalization bound over the true error on P_T

Difficult aspects

- Defining divergence g_n : avoid degenerate cases
- Finding/Learning weak DA learned: need to take into account both source error and divergence information

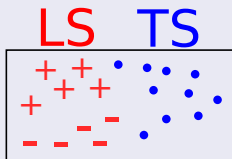
Model selection?

Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r

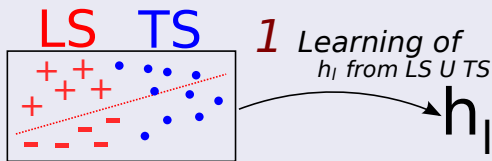
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



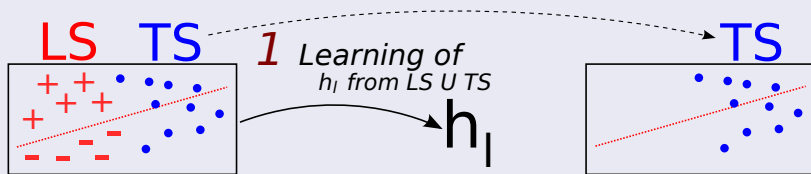
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



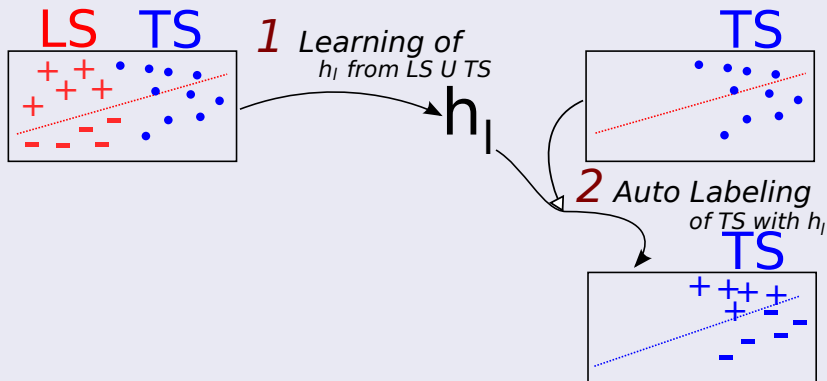
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



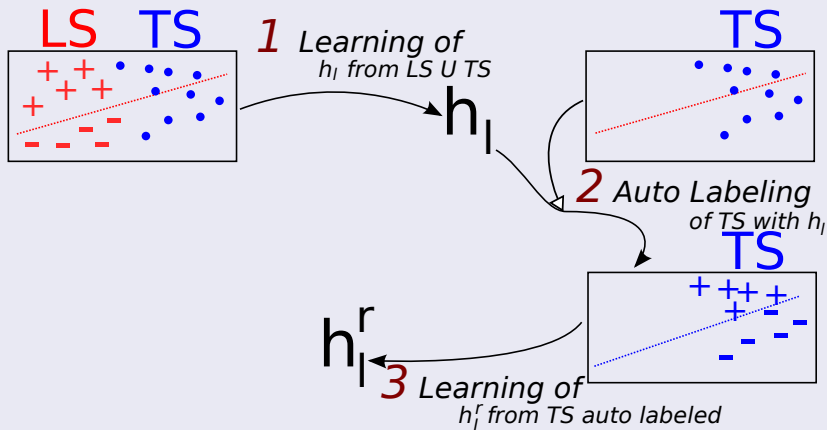
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



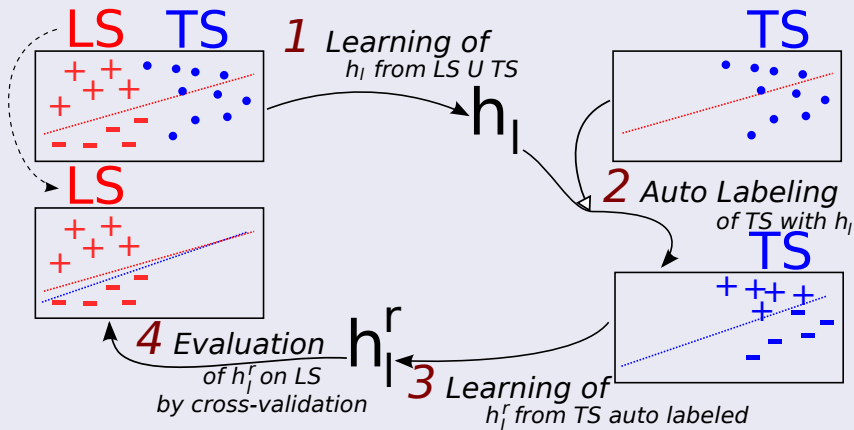
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



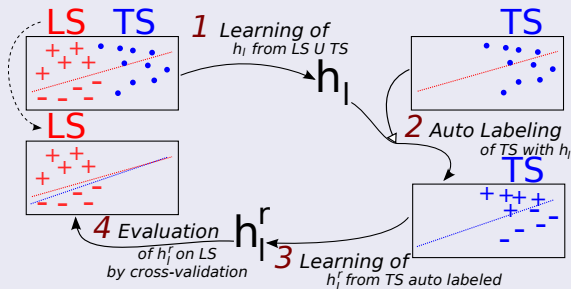
Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



Reverse validation [Zhong et al., ECML'10; Bruzzone et al., PAMI'10]

Reverse classifier h^r



- Two domains are related $\Rightarrow h_l^r$ performs well on the source domain
- Used with target labels to have an estimation of R_{P_T}
- Used to heuristically estimate theoretical constants of adaptability (λ) [Morvant et al., ICDM'11; KAIS'12]

Conclusion

Conclusion

- Very active domains - Lots of methods (Sometimes difficult to follow)
Approaches not covered here: probabilistic-based, bayesian, deep learning methods, etc.
- Same idea: Moving closer the distributions while ensuring good accuracy on labeled data
- Can we imagine general efficient frameworks
⇒ probably No: DA is difficult [Ben-David et al., ALT'12]
⇒ Choose a method in function of the task/data
- Importance of data preparation
- Importance of divergence measures

Perspectives

- Understanding negative transfer
- Model selection
- Heterogeneous data
- Large scale
- Links with multi-tasks and multi-source learning, lifelong learning, concept drift, etc.
- \Rightarrow A large room for more research

References (partial) (1/5)

- J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," , NIPS 2007.
- S. J. Pan and Q. Yang, "A survey on transfer learning", TKDE 2010.
- H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function", Journal of Statistical Planning and Inference, 2000.
- H. Daumé III, "Frustratingly easy domain adaptation", ACL, 2007.
- H. Daumé III, A. Kumar, et A. Saha. Co-regularization based semi-supervised domain adaptation. NIPS 2010.
- R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach", ICCV'11.
- B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation", CVPR 2012.
- K. Saenko, B. Kulis, M. Fritz, and T. Darrell., "Adapting visual category models to new domains", ECCV'2010.
- J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning", ICML 2007.

References (partial) (2/5)

- B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms", CVPR'11
- B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment", ICCV 2013
- L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy", PAMI 2010.
- B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation", ICML 2013
- X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach", ICML 2011.
- P. Germain, A. Habrard, F. Laviolette, et E. Morvant. "PAC-Bayesian domain adaptation bound with specialization to linear classifiers", ICML 2013.
- C. Cortes et M. Mohri. Domain adaptation in regression. ALT 2011.
- S. Ben-David, J. Blitzer, K. Crammer, et F. Pereira. Analysis of representations for domain adaptation. NIPS 2006.

References (partial) (3/5)

- S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, et J.W. Vaughan. A theory of learning from different domains. Machine Learning 2010.
- S. Ben-David et R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. ALT 2012.
- S. Ben-David et R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. AISTAT 2010.
- J. Ye C. Zhang, L. Zhang. Generalization bounds for domain adaptation. NIPS 2012.
- M. Chen, K. Q. Weinberger, et J. Blitzer. Co-training for domain adaptation. NIPS 2011.
- A. Habrard, JP Peyrache, M. Sebban. Iterative Self-labeling Domain Adaptation for Linear Structured Image Classification, IJAIT 2013.
- A. Habrard, JP Peyrache, M. Sebban. Boosting for unsupervised domain adaptation, ECML 2013
- Y. Mansour, M. Mohri, et A. Rostamizadeh. Domain adaptation : Learning bounds and algorithms. COLT 2009.
- Y. Mansour, M. Mohri, et A. Rostamizadeh. Multiple source adaptation and the rényi divergence. UAI 2009.

References (partial) (4/5)

- M. Mohri et A.M. Medina. New analysis and algorithm for learning with drifting distributions. ALT 2012.
- E. Morvant, A. Habrard, et S. Ayache. Sparse domain adaptation in projection spaces based on good similarity functions. ICDM 2011.
- E. Morvant, A. Habrard, et S. Ayache. Parsimonious Unsupervised and Semi-Supervised Domain Adaptation with Good Similarity Functions. KAIS 2012.
- E. Zhong, W. Fan, Q. Yang, O. Verscheure, et J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. ECML 2010.
- C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. Theoretical Computer Science, 2013.
- M. Sugiyama, S. Nakajima, H. Kashima, P. von Bunau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. NIPS 2008.
- C. Cortes, M. Mohri, A. Medina. Adaptation Algorithm and Theory Based on Generalized Discrepancy, 2014.
- S. Bickel, M. Bruckner, and T. Scheffer, "Discriminative learning for differing training and test distributions", 2007

References (partial) (4/5)

- B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation", CVPR 2012
- A. Torralba and A. A. Efros, "Unbiased look at dataset bias", CVPR 2011.
- A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Scholkopf. Covariate Shift by Kernel Mean Matching. 2008.