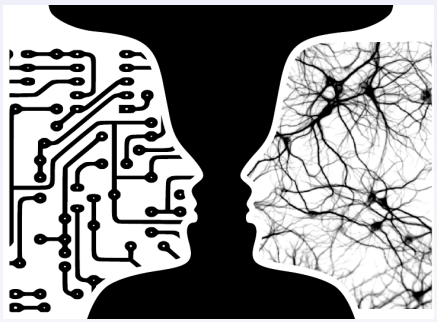
Un réseau de neurones pour l'adaptation de domaine

Pascal Germain

Travail conjoint avec
Hana Ajakan, Hugo Larochelle, François Laviolette et Mario Marchand

Groupe de recherche en apprentissage automatique de l'Université Laval (GRAAL)



23 janvier 2014

Plan

1. Mise en contexte
   - Apprentissage automatique et classification
   - Adaptation de domaine
2. Talk : *NIPS 2014 Workshop on Transfer and Multi-task learning*
   - Domain Adaptation Setting
   - Theoretical Foundations
   - Neural Network for Domain Adaptation
   - Empirical Results
3. Conclusion
   - Travaux futurs

Une définition de l'apprentissage automatique

"Field of study that gives computers the ability to learn without being explicitly programmed"

– Arthur Samuel, 1959

Exemple

critiques de films

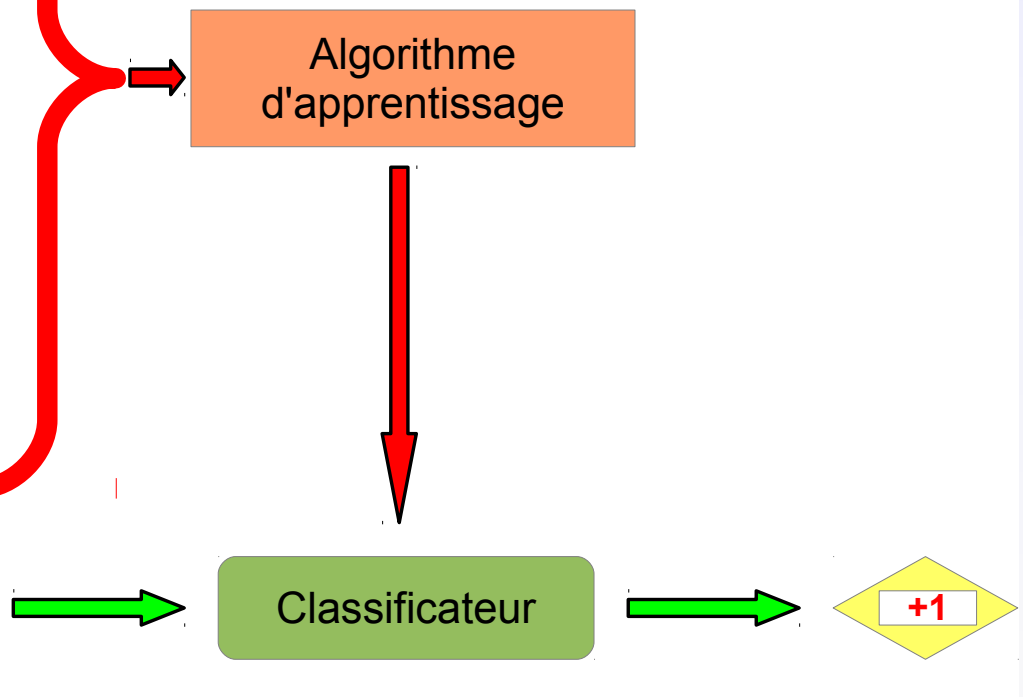
-1 **An insult to Douglas Adams' memory**
I agree entirely with "darkgenius'" comments. This movie is a travesty of the book and the TV series; a cutesy version totally lacking in the wit and satire of the original. Read more
Published 5 months ago by John W Beare

+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

??? **Mindbending**
I will not recommend this movie for people who haven't read at least two or three of Douglas Adams' books on hitchhiking. Read more
Published on Mar 28 2006 by alper bac



Définitions de base

Échantillon de données

$$S \stackrel{\text{def}}{=} \{ (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m) \} \in (\mathcal{X} \times \mathcal{Y})^m$$

Pour cette présentation,

Espace d'entré : $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ *(Attributs à valeurs réelles)*

Espace de sortie : $y \in \mathcal{Y} = \{-1, 1\}$ *(Classification binaire)*

Algorithme d'apprentissage

$$A(S) \longrightarrow \eta$$

Classificateur

Cas général : $\eta : \mathcal{X} \to \mathcal{Y}$

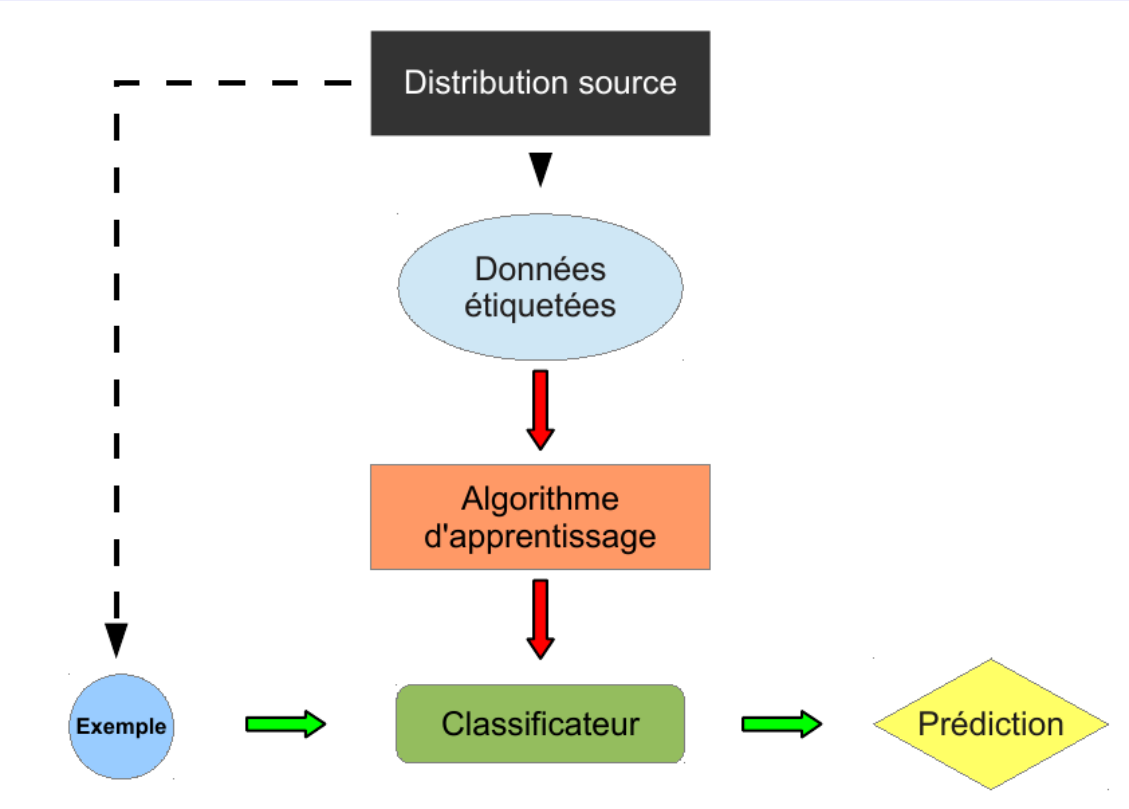
Pour cette présentation : $\eta : \mathbb{R}^d \to \{-1, 1\}$

Apprentissage inductif

Hypothèse

Les exemples sont générées *i.i.d.* par une distribution $\mathcal{D}$ sur $\mathcal{X} \times \mathcal{Y}$.

$$S \sim \mathcal{D}^m$$



Apprentissage inductif

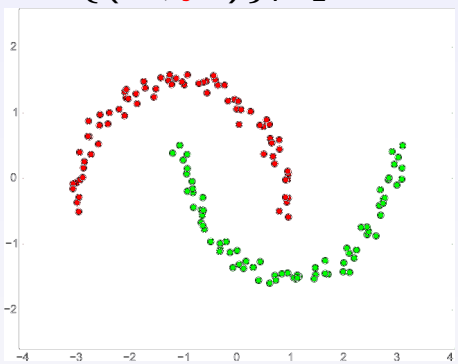
Hypothèse **(classification binaire)**

Les exemples sont générées *i.i.d.* par une distribution $\mathcal{D}$ sur $\mathbb{R}^d \times \{-1, 1\}$.

$$S \sim \mathcal{D}^m$$

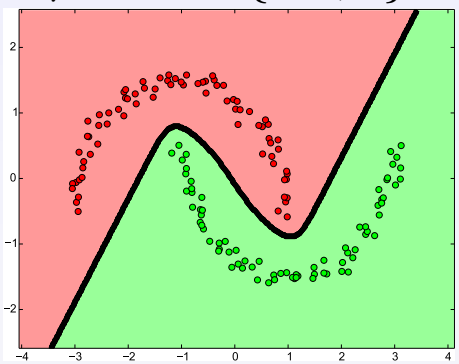
Un **algorithme d'apprentissage** reçoit un **échantillon d'entraînement**

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m,$$



et retourne un **classificateur**

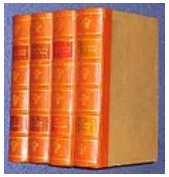
$$\eta : \mathbb{R}^d \to \{-1, 1\}.$$



On veut un classificateur avec un faible **risque**

$$R_{\mathcal{D}}(\eta) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x},y)\sim\mathcal{D}} \Big(\eta(\mathbf{x}) \neq y\Big).$$

Exemple



critiques de livres

critiques de films

??? **The end of the series.**
This book was written to provoke those who wanted Adams to continue the trilogy but I loved it. Aurthor setteled down on a bob fearing planet where he has aquired the prestigous...
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

Algorithme d'apprentissage

Classificateur

-1

-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

Définitions

Deux distributions de données

Distribution source $\mathcal{D}_S$; Distribution cible $\mathcal{D}_T$.

Deux chantillons d'entraînement

Échantillon source : $S \stackrel{\text{def}}{=} \{ (\mathbf{x}_1^s, y_1^s), (\mathbf{x}_2^s, y_2^s), \ldots, (\mathbf{x}_m^s, y_m^s) \} \sim (\mathcal{D}_S)^m$

Échantillon cible : $T \stackrel{\text{def}}{=} \{ \mathbf{x}_1^t, \mathbf{x}_2^t, , \ldots, \mathbf{x}_m^t \} \sim (\mathcal{D}_T)^m$ **(non étiqueté !)**

Algorithme d'apprentissage

$$A(S, T) \longrightarrow \eta$$

Classificateur

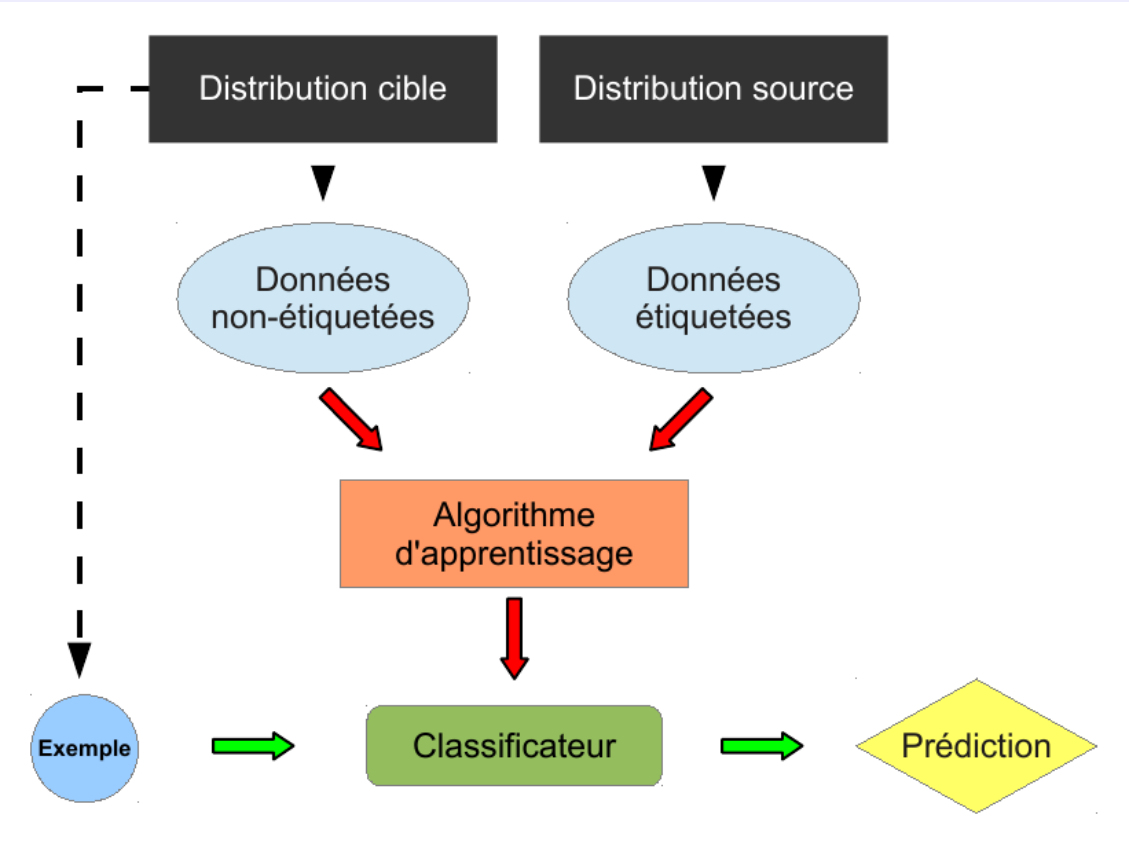
$\eta : \mathcal{X} \to \mathcal{Y}$ doit classifier des exemples cibles ($\mathbf{x}^t \sim \mathcal{D}_T$).

Adaptation de domaine

Hypothèse

Les exemples sont générées *i.i.d.* par les distributions $\mathcal{D}_S$ et $\mathcal{D}_T$ sur $\mathcal{X} \times \mathcal{Y}$.

$$S \sim (\mathcal{D}_S)^m \quad \text{et} \quad T \sim (\mathcal{D}_T)^m$$



Adaptation de domaine (deux approches)

Question

Dans quelle(s) situation(s) l'adaptation est-elle possible de $\mathcal{D}_S$ vers $\mathcal{D}_T$?

Réponse (partielle)

Lorsque les domaines $\mathcal{D}_S$ et $\mathcal{D}_T$ sont «semblables».

Un outil

Notion de distance $d_\eta(\mathcal{D}_S, \mathcal{D}_T)$ entre les distributions.

Deux approches pour la conception d'algorithmes

1. Trouver un classificateur $\eta \in \mathcal{H}$ tel que $d_\eta(\mathcal{D}_S, \mathcal{D}_T)$ est faible.
2. Modifier la représentation des exemples :
   $\Rightarrow$ Trouver une fonction $\mathbf{h}$ telle que $d_\eta(\mathbf{h}(\mathcal{D}_S), \mathbf{h}(\mathcal{D}_T))$ est faible.

Our Domain Adaptation Setting

Binary classification tasks

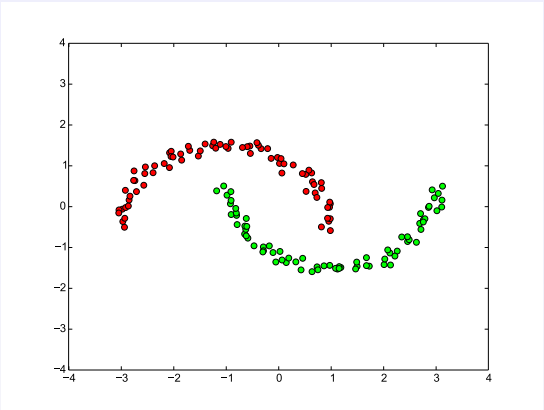
- Input space : $\mathbb{R}^d$
- Labels : $\{-1, 1\}$

Two different data distributions

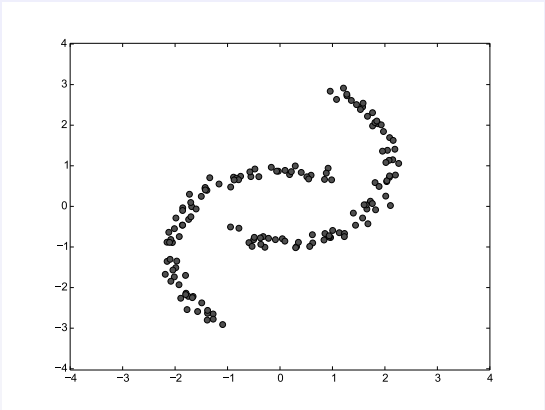
- Source domain : $\mathcal{D}_S$
- Target domain : $\mathcal{D}_T$

A **domain adaptation** learning algorithm is provided with

a **labeled source sample**
$S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,



an **unlabeled target sample**
$T = \{\mathbf{x}_i^t\}_{i=1}^m \sim (\mathcal{D}_T)^m$.



The goal is to build a classifier $\eta : \mathbb{R}^d \to \{-1, 1\}$ with a low **target risk**

$$R_{\mathcal{D}_T}(\eta) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x}^t, y^t) \sim \mathcal{D}_T} [\eta(\mathbf{x}^t) \neq y^t] .$$

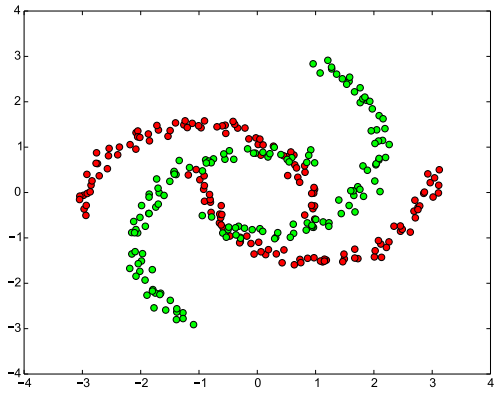
Divergence between source and target domains

Definition (Ben David et al., 2006)

Given two domain distributions $\mathcal{D}_S$ and $\mathcal{D}_T$, and a **hypothesis class** $\mathcal{H}$, the $\mathcal{H}$-**divergence** between $\mathcal{D}_S$ and $\mathcal{D}_T$ is

$$d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \stackrel{\text{def}}{=} 2 \sup_{\eta \in \mathcal{H}} \left| \Pr_{\mathbf{x}^s \sim \mathcal{D}_S} \left[\eta(\mathbf{x}^s) = 1\right] + \Pr_{\mathbf{x}^t \sim \mathcal{D}_T} \left[\eta(\mathbf{x}^t) = -1\right] - 1 \right|.$$

The $\mathcal{H}$-**divergence** measures the ability of an hypothesis class $\mathcal{H}$ to **discriminate** between source $\mathcal{D}_S$ and target $\mathcal{D}_T$ distributions.



Bound on the target risk

Theorem (Ben David et al., 2006)

Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. With probability $1-\delta$ over the choice of samples $S \sim (\mathcal{D}_S)^m$ and $T \sim (\mathcal{D}_T)^m$, for every $\eta \in \mathcal{H}$:

$$R_{\mathcal{D}_T}(\eta) \leq R_S(\eta) + \frac{4}{m}\sqrt{d \log \frac{2em}{d} + \log \frac{4}{\delta}} + \hat{d}_{\mathcal{H}}(S,T) + \frac{4}{m^2}\sqrt{d \log \frac{2m}{d} + \log \frac{4}{\delta}} + \beta$$

with $\beta \geq \inf_{\eta^* \in \mathcal{H}} \left[R_{\mathcal{D}_S}(\eta^*) + R_{\mathcal{D}_T}(\eta^*) \right]$.

Empirical risk on the source sample :

$$R_S(\eta) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I[\eta(\mathbf{x}_i^s) \neq y_i^s] .$$

Empirical $\mathcal{H}$-divergence :

$$\hat{d}_{\mathcal{H}}(S,T) \stackrel{\text{def}}{=} 2 \max_{\eta \in \mathcal{H}} \left[\frac{1}{m}\sum_{i=1}^{m} I[\eta(\mathbf{x}_i^s) = 1] + \frac{1}{m}\sum_{i=1}^{m} I[\eta(\mathbf{x}_i^t) = -1] - 1 \right] .$$

Bound on the target risk

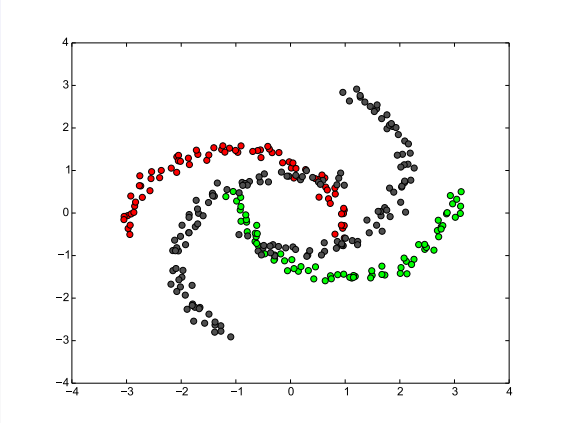
Theorem (Ben David et al., 2006)

Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. With probability $1-\delta$ over the choice of samples $S \sim (\mathcal{D}_S)^m$ and $T \sim (\mathcal{D}_T)^m$, for every $\eta \in \mathcal{H}$:

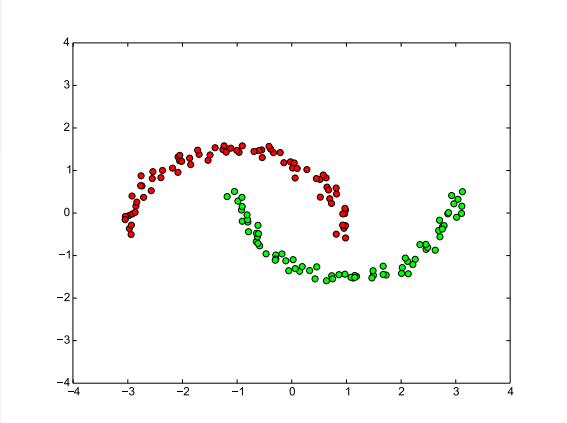
$$R_{\mathcal{D}_T}(\eta) \leq R_S(\eta) + \frac{4}{m}\sqrt{d \log \frac{2e\,m}{d} + \log \frac{4}{\delta}} + \hat{d}_{\mathcal{H}}(S,T) + \frac{4}{m^2}\sqrt{d \log \frac{2\,m}{d} + \log \frac{4}{\delta}} + \beta$$

with $\beta \geq \inf_{\eta^* \in \mathcal{H}} \left[R_{\mathcal{D}_S}(\eta^*) + R_{\mathcal{D}_T}(\eta^*)\right]$.

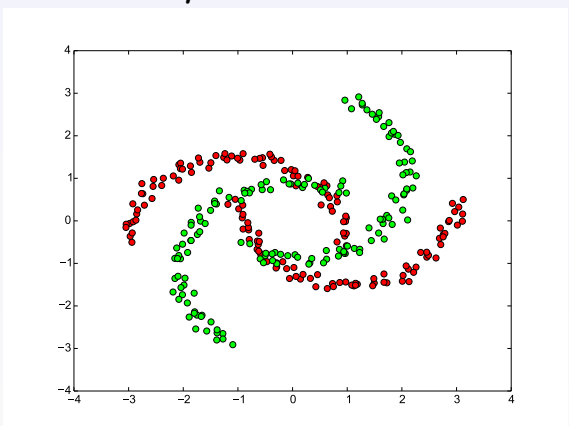
Target risk $R_{\mathcal{D}_T}(\eta)$ is low
if, given S and T,



$R_S(\eta)$ **is small**,
i.e., $\eta \in \mathcal{H}$ is good on



and $\hat{d}_{\mathcal{H}}(S,T)$ is small,
i.e., all $\eta' \in \mathcal{H}$ are bad on



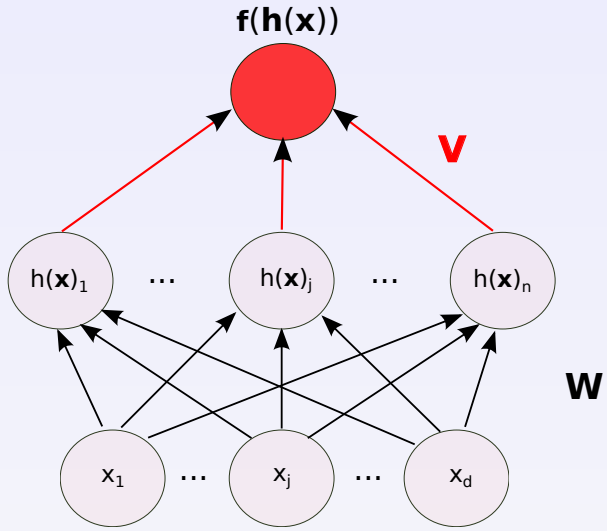
Standard Neural Network

Let consider a neural network architecture with one hidden layer

$$\mathbf{h}(\mathbf{x}) = \mathrm{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x}), \quad \text{and} \quad \mathbf{f}(\mathbf{h}(\mathbf{x})) = \mathrm{softmax}(\mathbf{c} + \mathbf{V}\mathbf{h}(\mathbf{x})).$$

$$\min_{\mathbf{W},\mathbf{V},\mathbf{b},\mathbf{c}} \underbrace{\left[\frac{1}{m} \sum_{i=1}^{m} -\log\left(f_{y_i^s}(\mathbf{x}_i^s) \right) \right]}_{\text{source loss}}.$$

where $f_y(\mathbf{x})$ denotes the conditional probability that the neural network assigns $\mathbf{x}$ to class y.



Given a source sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,

1. Pick a $\mathbf{x}^s \in S$
2. Update $\mathbf{V}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
3. Update $\mathbf{W}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$

The hidden layer learns a **representation** $\mathbf{h}(\cdot)$ from which linear hypothesis $\mathbf{f}(\cdot)$ can **classify source examples**.

Domain-Adversarial Neural Network (DANN)

Empirical $\mathcal{H}$-divergence

$$\hat{d}_{\mathcal{H}}(S, T) \stackrel{\text{def}}{=} 2 \max_{\eta \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^{m} I[\eta(\mathbf{x}_i^s) = 1] + \frac{1}{m} \sum_{i=1}^{m} I[\eta(\mathbf{x}_i^t) = -1] - 1 \right].$$

We estimate the $\mathcal{H}$-divergence by a logistic regressor that model the probability that a given input (either $\mathbf{x}^s$ or $\mathbf{x}^t$) is from the source domain :

$$o(\mathbf{h}(\mathbf{x})) \stackrel{\text{def}}{=} \text{sigm}(d + \mathbf{w}^\top \mathbf{h}(\mathbf{x})).$$

Given a representation output by the hidden layer $\mathbf{h}(\cdot)$:

$$\hat{d}_{\mathcal{H}}\Big(\mathbf{h}(S), \mathbf{h}(T)\Big) \approx 2 \max_{\mathbf{w}, d} \left[\frac{1}{m} \sum_{i=1}^{m} \log\big(o(\mathbf{h}(\mathbf{x}_i^s))\big) + \frac{1}{m} \sum_{i=1}^{m} \log\big(1 - o(\mathbf{h}(\mathbf{x}_i^t))\big) - 1 \right].$$

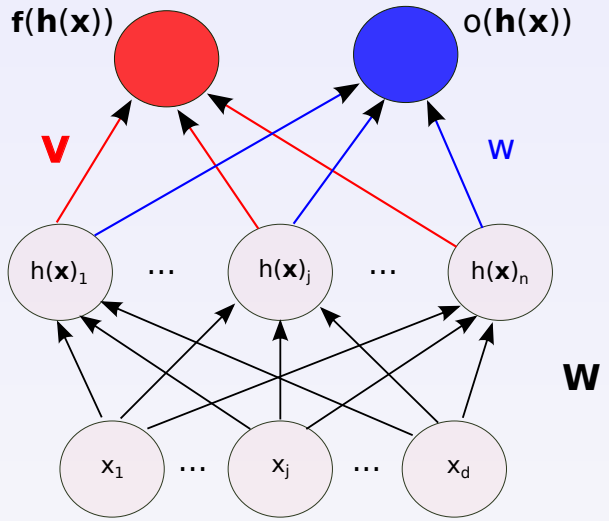
Domain-Adversarial Neural Network (DANN)

$$\min_{\mathbf{W},\mathbf{V},\mathbf{b},\mathbf{c}} \left[\underbrace{\frac{1}{m}\sum_{i=1}^{m} -\log\left(f_{y_i^s}(\mathbf{x}_i^s)\right)}_{\text{source loss}} + \lambda \underbrace{\max_{\mathbf{w},d} \left(\frac{1}{m}\sum_{i=1}^{m} \log\left(o(\mathbf{h}(\mathbf{x}_i^s))\right) + \frac{1}{m}\sum_{i=1}^{m} \log\left(1-o(\mathbf{h}(\mathbf{x}_i^t))\right) \right)}_{\text{adaptation regularizer}} \right],$$

where $\lambda > 0$ weights the domain adaptation regularization term.

Given a source sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,
and a target sample $T = \{(\mathbf{x}_i^t)\}_{i=1}^m \sim (\mathcal{D}_T)^m$,

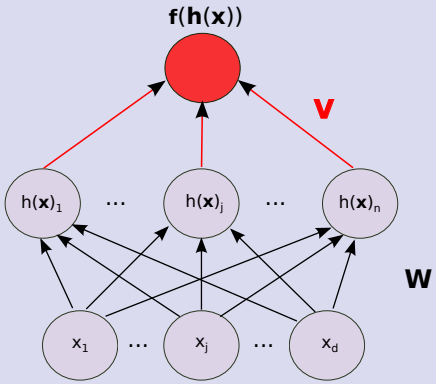
1. Pick a $\mathbf{x}^s \in S$ and $\mathbf{x}^t \in T$
2. Update $\mathbf{V}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
3. Update $\mathbf{W}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
4. Update $\mathbf{w}$ towards $o(\mathbf{h}(\mathbf{x}^s)) = 1$ and $o(\mathbf{h}(\mathbf{x}^t)) = -1$
5. Update $\mathbf{W}$ towards $o(\mathbf{h}(\mathbf{x}^s)) = -1$ and $o(\mathbf{h}(\mathbf{x}^t)) = 1$



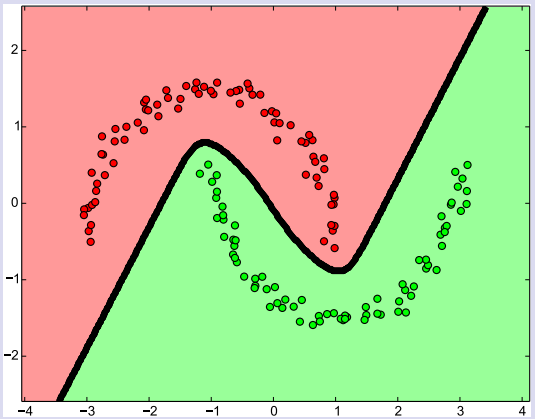
DANN finds a representation $\mathbf{h}(\cdot)$ that are good on S;
but unable to discriminate between S and T.

Toy Dataset

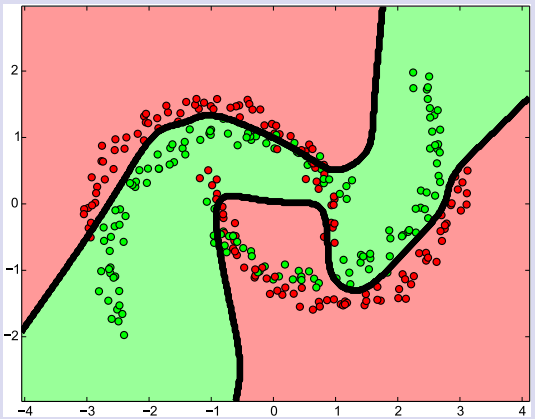
Standard Neural Network (NN)



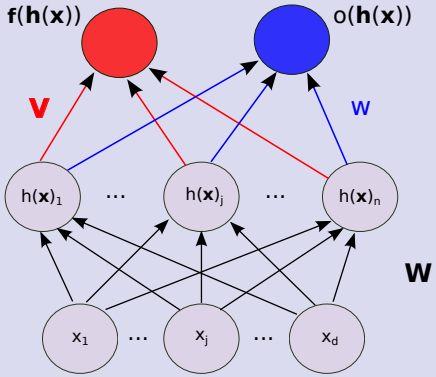
Trained to classify source



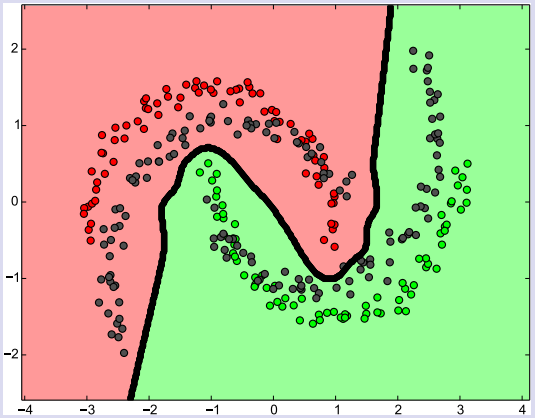
Trained to classify domains



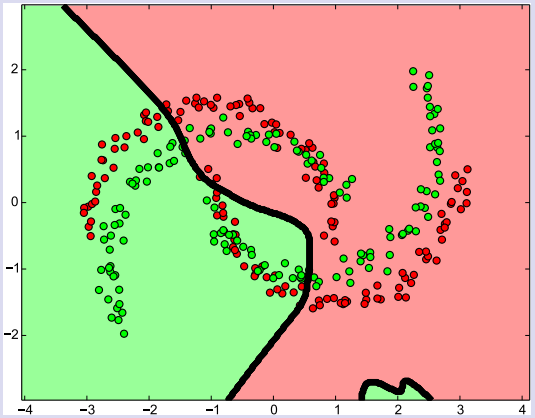
Domain-Adversarial Neural Networks (DANN)



Classification output : $\mathbf{f}(\mathbf{h}(\mathbf{x}))$

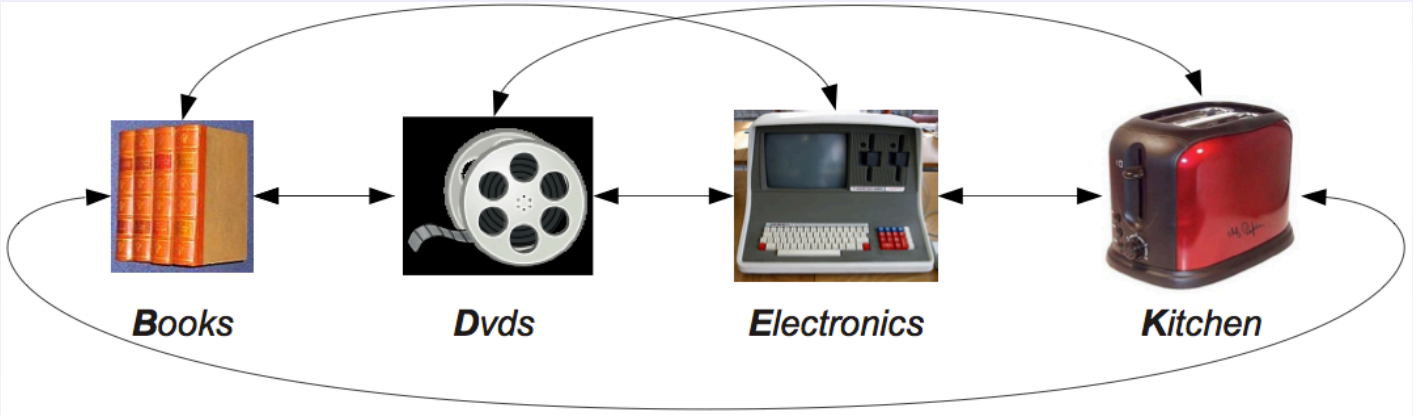


Domain output : $o(\mathbf{h}(\mathbf{x}))$



Amazon Reviews

Input : product review (bag of words) — **Output :** positive or negative rating.



Amazon Reviews

Input : product review (bag of words) — **Output :** positive or negative rating.

Dataset	**DANN**	**NN**
books → dvd	0.201	**0.199**
books → electronics	**0.246**	0.251
books → kitchen	**0.230**	0.235
dvd → books	**0.247**	0.261
dvd → electronics	**0.247**	0.256
dvd → kitchen	0.227	0.227
electronics → books	**0.280**	0.281
electronics → dvd	**0.273**	0.277
electronics → kitchen	**0.148**	0.149
kitchen → books	**0.283**	0.288
kitchen → dvd	0.261	0.261
kitchen → electronics	0.161	0.161

Note : We use a *small labeled subset* of 100 target examples to select the hyperparameters.

Marginalized Stacked Denoising Autoencoders (mSDA)

Question

Does DANN can be combined with other representation learning techniques for domain adaptation ?

The autoencoders mSDA (Chen et al. 2012) provides a new common representation for source and target (unsupervised)

With **mSDA+SVM**, Chen et al. (2012) obtained *state-of-the-art* results on Amazon Reviews :
– Train a linear SVM on mSDA source representations.

We try **mSDA+DANN** :
– Train DANN on source representations and target representations.

Amazon Reviews

Input : product review (bag of words) — **Output :** positive or negative rating.

Dataset	**mSDA+DANN**	**mSDA+SVM**
books → dvd	0.176	**0.175**
books → electronics	**0.197**	0.244
books → kitchen	**0.169**	0.172
dvd → books	0.176	0.176
dvd → electronics	**0.181**	0.220
dvd → kitchen	**0.151**	0.178
electronics → books	0.237	**0.229**
electronics → dvd	**0.216**	0.261
electronics → kitchen	**0.118**	0.137
kitchen → books	**0.222**	0.234
kitchen → dvd	**0.208**	0.209
kitchen → electronics	0.141	**0.138**

Note : We use a *small labeled subset* of 100 target examples to select the hyperparameters. The *noise parameter* of mSDA representations is fixed to 50%.

Travaux futurs

Quelques avenues à explorer :

- Réseau de neurones profonds (*deep learning*) .
- Problèmes multi-classes et multi-étiquettes.
- Adaptation avec plusieurs ensembles sources.

Merci !