The Theory-Practice Interplay in Machine Learning – Emerging Theoretical Challenges

Shai Ben-David

University of Waterloo

ECML/PKDD 2009

uccess and limitations
f Machine Learning Theory

- Some remarkable successes, like *Boosting* and *Support Vector Machines* are founded on theoretical insights.

- However, our mainstream theory is based on simplistic models of learning and uses simplifying assumptions that many potential applications fail to meet.

- Major challenge –

  Extend ML theory to cover a wider variety of realistic learning settings.

If it works, why bother with theory?

- To provide success guarantees to common heuristics.
- To help understand under what circumstances common learning techniques may fail.
- To help choose the right leaning paradigm for a given task.
- To guide the development of novel learning algorithms.
- .....

The modeling of prior knowledge

- ***No learning is possible without applying prior knowledge.***

(This is the "no free lunch" phenomena).

- A central, yet not always explicit, questions in learning is "How should one model prior knowledge?"

Common PK modeling frameworks

- Hypothesis classes –

  Allow clean mathematical analysis, yet far from being user friendly.

- Kernels, or better yet, similarity functions- much more user friendly.

  Still, they require too detailed commitment to prior knowledge.

  In practice, choice of kernels is often ad hoc.

Prior Knowledge as "Meta-Bias"

Rather than coming up with a learning bias, in the form of a hypothesis class or a kernel function, a learner may incorporate a prior belief about the relevance of some external data:

Semi-supervised learning (SSL) is based on the premise that the spatial distribution of the unlabeled data is related to the actual labels of the data points.

Multi-task learning is based on the premise that the labels of points under one task are relevant to their labels under another task.

The theory-practice gap

Learning paradigms based on incorporating domain knowledge in the form of "meta-bias" , such as Semi-supervised learning, domain adaptation learning and multi-task learning, are being commonly used in practice, but theory does not yet provide them with satisfactory support and understanding.

Negative results for SSL

In work with my students, David pal and Tyler Lu, we showed that, contrary to common belief,

The SSL paradigm (or the use of unlabeled data) cannot ***provide guarantees of performance enhancement, unless one makes*** further strong assumptions ***about the data-generation process***.

Another type of meta-bias—
Multi-Task Learning

Our experience with human learning is that it is easier to learn a task if related tasks have already been mastered:

Learning to drive different vehicles

Learning new languages

Empirical work indeed shows that data from extra "related" tasks does improve accuracy. [Intrator and Adelman, '96], [Thrun, '96] [Caruana, '97], [Heskes, '00] [Bakker and Heskes '03]

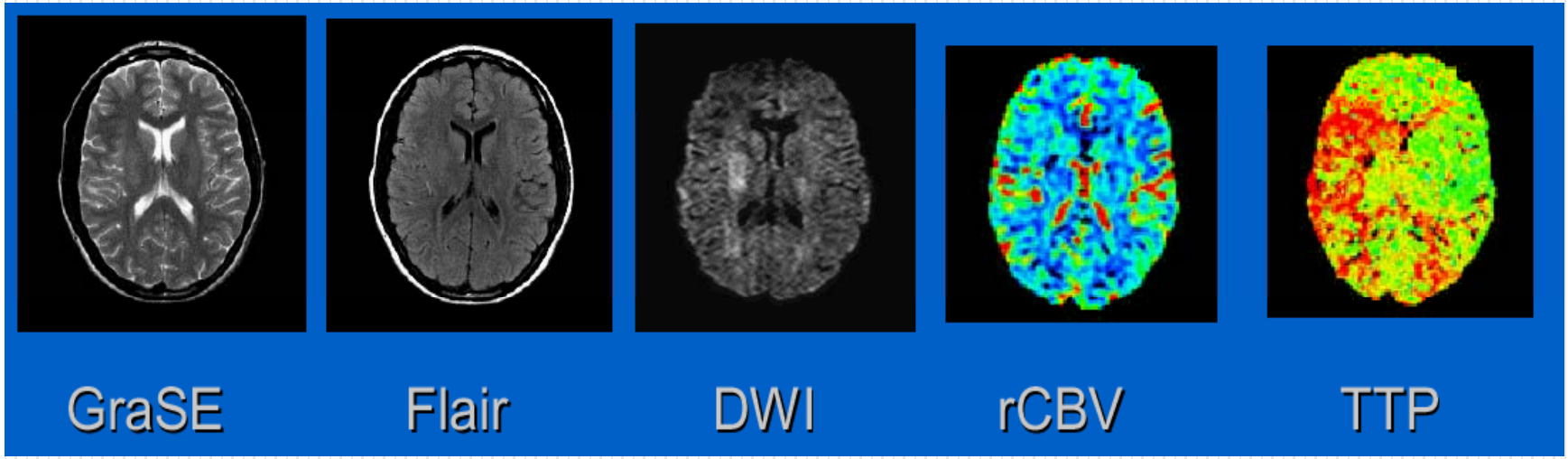
Can this be backed up by solid theory?

Key component for theoretical analysis– modeling task relatedness

- Bayesian: Probability model for task generation [Baxter '00], [Heskes '00], [The, Seeger, Jordan '05], [Zhang, Gharamani, Yang '05]
- Between-task noise correlations [Greene '02]
- Hidden common data structure
  - Implicit structure (common kernels) [Evgeniou, Micchelli, Pontil '05]
  - Explicit structure (PCA) [Ando, Zhang '04]

Transformation relatedness [BD-Schuler'07]

Consider the task of learning to detect images of tumors through different brain-imaging techniques.



One can assume that there exist functions of data-transformation between these tasks.

Evaluation of the Transformation-Relatedness Approach

- Allow utilizing advantages from a *small number of tasks*.
- Allows the derivation of *generalization error bounds* that are provably superior to the corresponding single task learning bounds.
- Applies to scenarios not handled by previous approaches.
- **Big weakness** – applies to *only limited sub-domain* of multitask learning.

A major common assumption-
The data-generating distribution is stationary

Most of the statistical learning guarantees are based on assuming that

The learning environment is unchanged throughout the learning process.

Formally, it is common to assume that

both the training and the test examples are generated i.i.d. by the **same fixed** *probability distribution.*

This is unrealistic for many ML applications

Learning when Training and Test distributions differ

- *Driving lessons* – train on one car, test on another.
- *Spam filters* – train on email arriving at one address, test on a different mailbox.
- *Natural Language Processing* tasks- train on some content domains, test on others.

Main issue-
MODELING TASK RELATEDNES

Preliminary convention *– a learning task is a joint distribution over points and labels,* P *over* $X \times \{0,1\}$.

- Relatedness of unlabeled distributions (How should we measure that?)
- Relatedness of the labels in both tasks (again, how to measure it?)

- Most obvious desired relatedness – labels are the same.

  this is the "covariate shift" assumption.

Evaluating the generalization bound

We can bound the error of a predictor h on the target task, $\varepsilon_T(\mathrm{h})$ in terms of its training error on the source task, $\varepsilon_S(\mathrm{h})$,

$$\varepsilon_T(\mathrm{h}) \le \varepsilon_S(\mathrm{h}) + \mathrm{d}_{\Delta \mathrm{H}}(U_S, U_T) + \inf_{h \in \mathrm{H}} (\mathrm{Er}_T(h) + \mathrm{Er}_S(h))$$

(Plus a term that goes to zero with the sample sizes)

Can this bound be improved?

The algorithmic conclusion

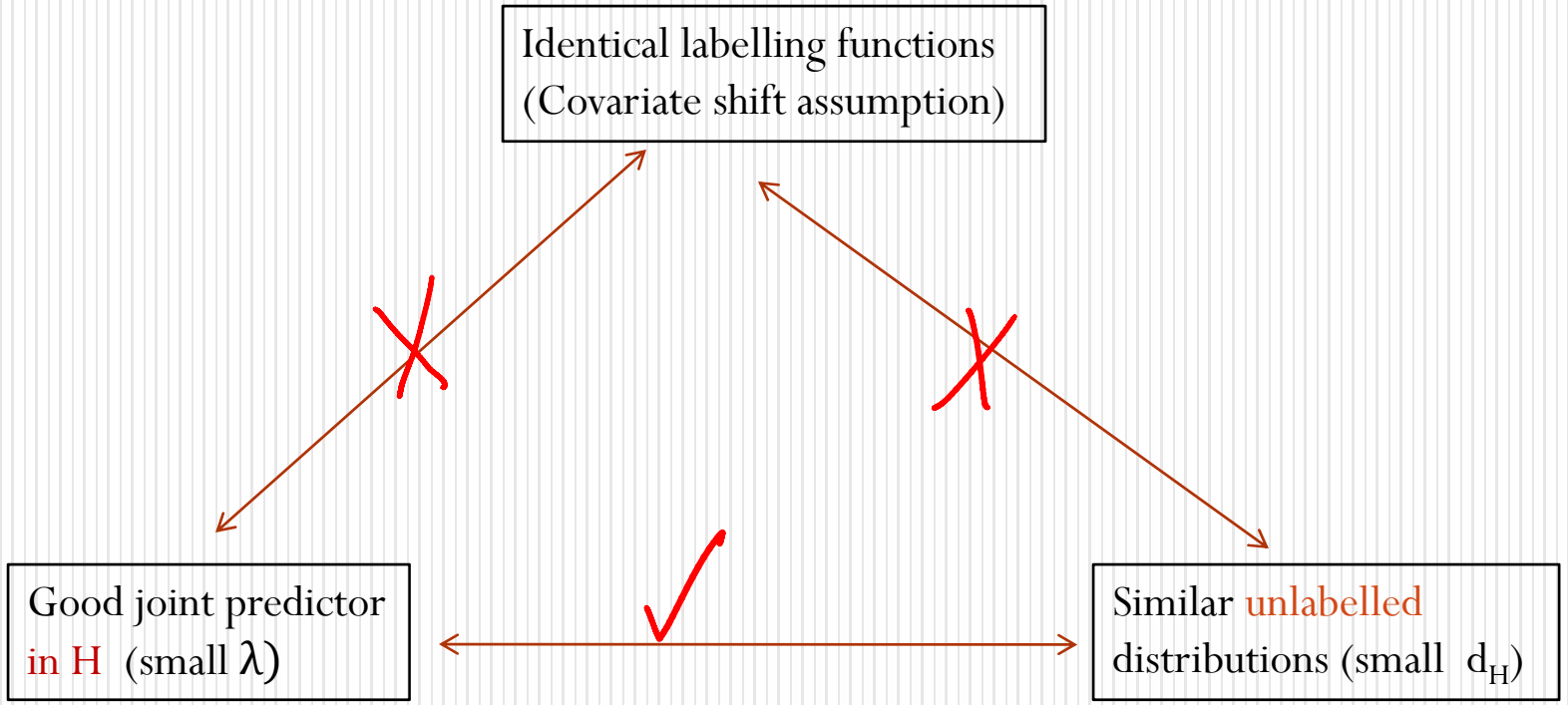
Find a feature space representation, R, such that:

1. The (unlabeled) distributions induced by the *Source* and *Target* distributions under a representation *R* are similar.
2. There *exist* a predictor in *H* that works reasonably well for the training data (in the feature space).

To predict:

Represent *your test point in the feature space, and use a good training classifier to predict its label.*

Can the covariate shift assumption help?



Major remaining open questions

- Improve our basic generalization-error bound.
- Find relatedness parameters under which different paradigms work (e.g., ERM with respect to task-reweighted training sample).
- Come up with different adaptive (rather than conservative) learning algorithms.
- Come up with more user-friendly useful notions of relatedness.

Summary: Prior-Knowledge Expression -- a major (under-researched?) Challenge

In all the above settings the first challenge is to find suitable formalisms for expressing high-level prior knowledge about the relationships between the external source of data and the target classification task.

Such formalisms should, on one hand, allow natural expression of domain-expert beliefs and, at the same time, allow derivation of provably significant merits of that knowledge.

Major challenge- clustering

Clustering is one of the most widely used tool for exploratory data analysis.

Social Sciences

Biology

Astronomy

Computer Science

.

.

All apply clustering to gain a first understanding of the structure of large data sets.

Yet, there exist distressingly little theoretical understanding of clustering

Questions that research of **fundamentals of clustering** should address

- Can clustering be given an *formal* and *general* definition?
- What is a "good" clustering?
- Can we distinguish "clusterable" from "structureless" data?
- Can we distinguish meaningful clustering from random structure?
- Given a clustering task, how should a user choose a suitable clustering algorithm?

Worthy challenge:
classification of clustering paradigms

Given a clustering task, how should a user choose a suitable clustering algorithm?

We wish to formulate properties of clustering functions that would allow translating prior knowledge about a clustering task into guidance concerning the choice of suitable clustering functions.

Next, I will show two examples of such results.

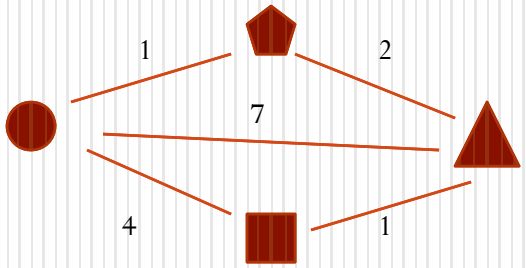
Bosagh Zadeh, Ben-David, UAI 2009

Characterizing Single Linkage clustering

Given a dissimilarity measure, d over some domain set X, we define the *d-induced path distance*, P_d, by setting, for all $x, y \in X$,

$$P_d(x,y) = \min_{q \in P_{x,y}} \max_{i<|q|} d(q(i), q(i+1))$$

In other words, we find the path from x to y, which has the smallest longest jump in it.



Characterization of Single Linkage

Theorem [BD- Bosagh Zadeh]

Single-Linkage is the *only clustering function*

satisfying:

k-Richness,

Order-Consistency

and

is determined by the Path Distance

Characterizing Linkage Based clusterings

- Given (X,d) define an induced dissimilarity over subsets of X, $\hat{d}$
- Let $F_0(X,d) = \{\{x\} : x \in X\}$
- Construct $F_{n+1}(X,d)$ from $F_n(X,d)$ by merging the two $\hat{d}$ –closest clusters of $SL_n(X,d)$

Properties of Linkage-Based clustering

- The Refinement property:

  For all $k' < k$, for every $C \in F(X, d, k)$

  there exist $C' \in F(X, d, k')$ such that $C \subseteq C'$

- The Locality property:

  For every $S \subseteq F(X, d, k)$,

  $$F(US, d, |S|) = S$$

Characterizing Linkage-Based clusterings

- Theorem [Ackerman, BD, Loker]:

  A clustering function can be

  defined as a linkage-based clustering if and only if it satisfies the Refinement and the Locality properties.

Note that most of the fundamental questions raised remain open ....

- Can clustering be given an *formal* and *general* definition?
- What is a "good" clustering?
- Can we distinguish "clusterable" from "structureless" data?
- Can we distinguish meaningful clustering from random structure?
- Given a clustering task, how should a user choose a suitable clustering algorithm?

Some further important new research directions

Privacy-preserving learning (or data mining):

Come up with a protocol for interaction between a data-base curator and a learner so that:

- The interaction allows significant learning
- The interaction provably does not leak any "private" information.
- The interaction does not ask the data curator to perform any non-standard computations.

Another practice-driven challenge
Learning with teachers of varying expertise

In the context of developing automated medical image analysis tools, a key obstacle is the cost of expert classification of images.

When/how can we use less-reliable, cheaper, expert supervision for learning to classify?

Demonstrating uselessness of the Covariate Shift assumption

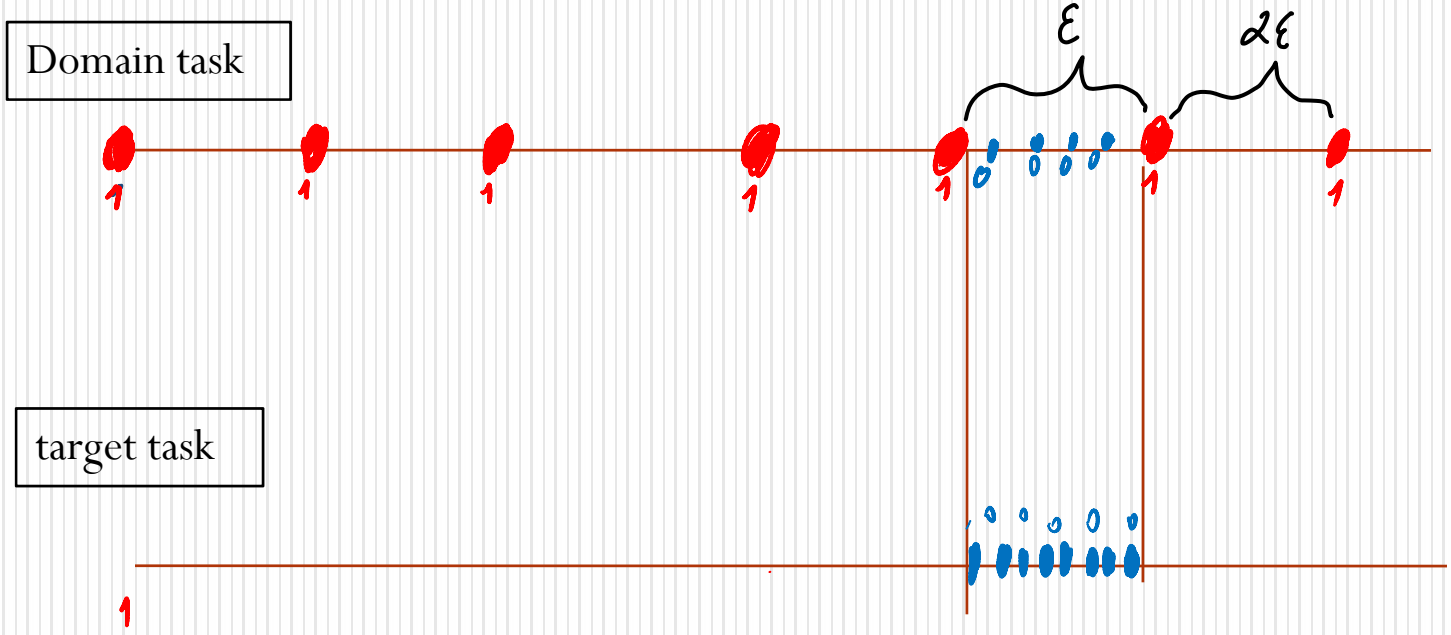
Domain task

target task

Note that here we have the covariate shift assumption + small d_H

Yet, the *All-1* h has small error on the domain task and large error on the target

Demonstrating uselessness of the Covariate Shift assumption



Note that here we have the covariate shift assumption + small λ

Yet, the al-1 h has small error on the domain task and large error on the target

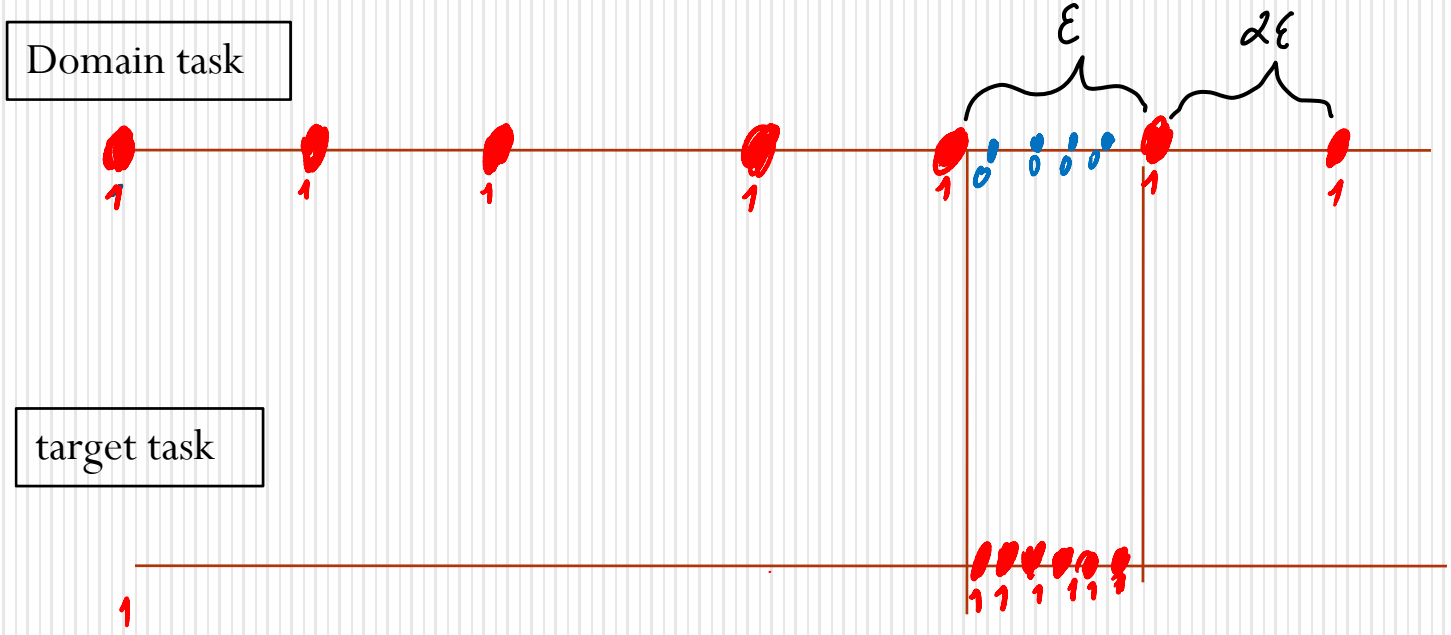
Conservative vs. Adaptive algorithms

- Can we do better than learning on the source domain and applying the SAME hypothesis to the target domain?
- We should! But how?
- The Afshani-Mohari-Mansour idea:

  Use the target unlabeled sample to reweigh the training sample. Choose h that minimizes the training error w.r.t. This reweighted training sample.

- It may fail badly!
- Q: Under what assumptions will it work?
- Q: What other types of adaptive DA may work?

Demonstrating
the failure of the AMM reweighting paradigm



Note that here , although we have small λ,
and without reweighting the learner will have zero target error,
after reweighting the learner will choose the *All-0* hypothesis
and fail badly