Distribution Matching for Transduction

Authors

Novi Quadrianto[1] | James Petterson[1] | Alex J. Smola[2]

1: SML, NICTA & RSISE, ANU | 2: Yahoo! Research

Abstract

- Many transductive inference algorithms assume that distributions over training and test estimates should be related, e.g. by providing a large margin of separation on both sets.
- We use this idea to design a transduction algorithm which can be used without modification for classification , regression , and structured estimation .
- At its heart we exploit the fact that for a good learner the distributions over the outputs on training and test sets should match.
- This is a classical two-sample problem which can be solved efficiently in its most general form by using distance measures in Hilbert Space.
- Further, our approach is scalable and can be easily used with online optimization algorithms.

Two-sample Problem

The two-sample problem

- Let p and p' be distributions defined on a domain $\mathcal{X}$. Given observations $X := \{x_1, \ldots, x_m\}$ and $X' := \{x'_1, \ldots, x'_n\}$, drawn i.i.d from p and p' respectively, is $p \neq p'$?

Maximum Mean Discrepancy (MMD; Gretton et al. 2008)

Denote $\mu[p] := \mathbf{E}_{x \sim p(x)}[k(x, \cdot)]$, then

$$\mathrm{MMD}[p, p'] = \left\|\mu[p] - \mu[p']\right\|^2_{\mathcal{H}}$$

Empirical estimate of MMD

$$\mathrm{MMD}[X, X'] = \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, x'_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x'_i, x'_j)\right]^{\frac{1}{2}}$$

One of the advantages of MMD is

- Computing MMD is simple: only the kernel matrices K and L are needed.

Distribution Matching for Transduction

Standard Supervised Learning

Given a training set $\mathcal{D}$ comprising m labeled samples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, design an estimator which minimizes

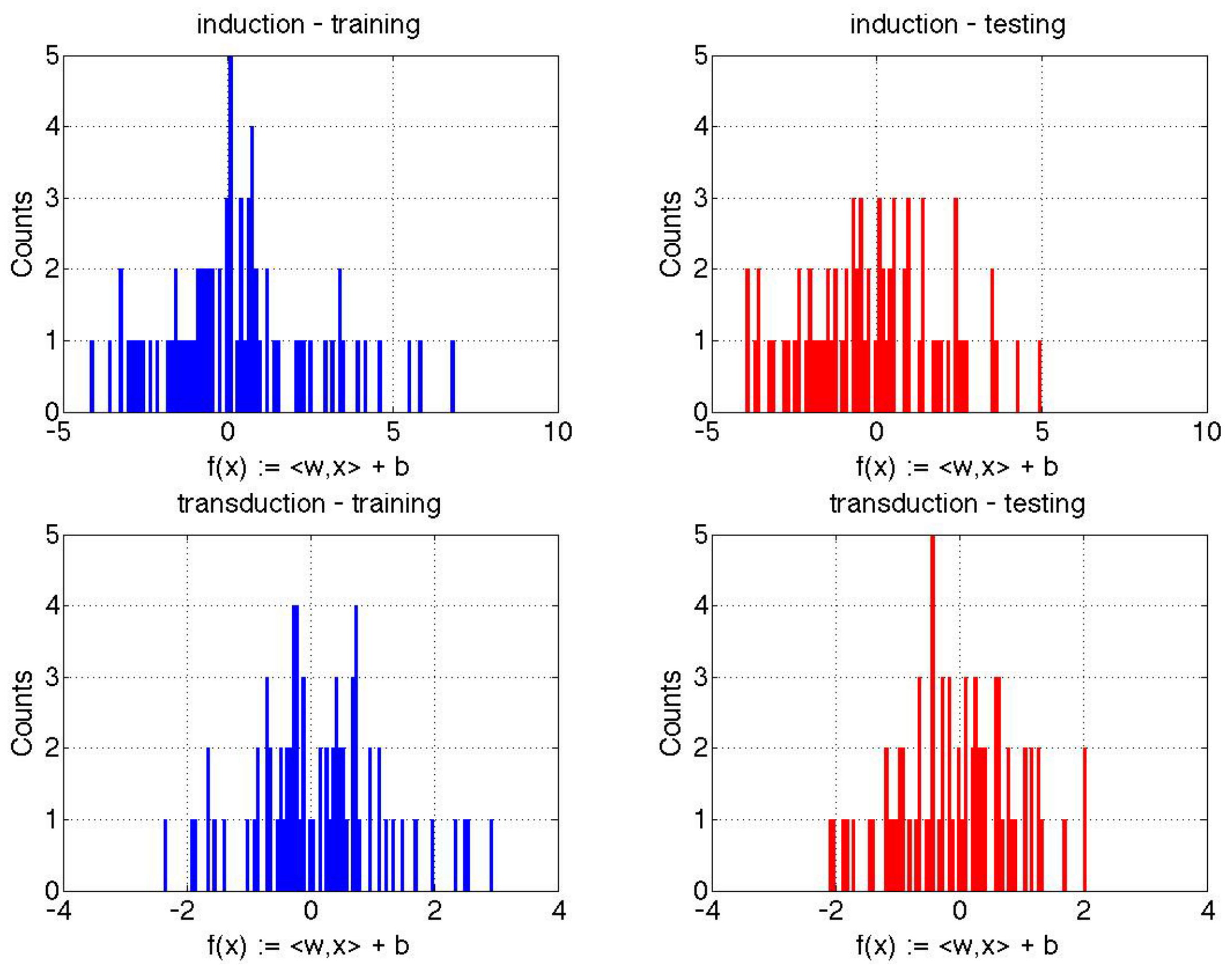
- Regularized risk functional
  $R_{\mathrm{reg}}[f, X, Y] := \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f) + \lambda\Omega[f]$
- Or, log-posterior probability
  $\log p(f|X, Y) = \sum_{i=1}^{m} \log p(y_i|x_i, f) + \log p(f) + \text{const.}$

Transductive Learning via Distribution Matching

Given the labeled training set $\mathcal{D}$ and a test set comprising m' unlabeled samples $\{x_1, \ldots, x'_m\}$. Denote the training risk term as $R_{\mathrm{train}}[f, X, Y]$. Further, denote by $f(X) := \{f(x_1), \ldots, f(x_m)\}$ and by $f(X') := \left\{f(x'_1), \ldots, f(x'_{m'})\right\}$ the applications of our estimator to training and test set respectively. The objective function for a transductive inference is then

$$R_{\mathrm{train}}[f, X, Y] + \gamma D(f(X), f(X')) \text{ for some } \gamma > 0$$

In the above, $D(f(X), f(X'))$ denotes the distance between the two distributions $f(X)$ and $f(X')$. We choose $D(f(X), f(X'))$ to be $\mathrm{MMD}[f(X), f(X')]$.



Optimization

Online Approximation

The empirical estimate of MMD can be approximated by

$$\hat{D} := \frac{1}{m}\sum_{i=1}^{m} D_i \text{ where}$$

$$D_i := [k(f(x_i), f(x_{i+1})) - k(f(x_i), f(x'_{i+1})) - k(f(x_{i+1}), f(x'_i)) + k(f(x'_i), f(x'_{i+1}))]$$

Stochastic Gradient Descent

The streaming transductive objective function is now

$$\begin{aligned} & \bar{l}(x_i, x_{i+1}, y_i, y_{i+1}, x'_i, x'_{i+1}, f) \\ := \; & l(x_i, y_i, f) + l(x_{i+1}, y_{i+1}, f) + 2\lambda\Omega[f] + \\ & \gamma[k(f(x_i), f(x_{i+1})) - k(f(x_i), f(x'_{i+1})) - k(f(x_{i+1}), f(x'_i)) + k(f(x'_i), f(x'_{i+1}))] \end{aligned}$$

Algorithm

```
**Input:** Convex set $A$, objective function $\bar{l}$
Initialize $w = 0$
**for** $t = 1$ to $N$ **do**
  Sample $(x_i, y_i), (x_{i+1}, y_{i+1}) \sim p(x, y)$ and $x'_i, x'_{i+1} \sim p(x)$
  Update $w \leftarrow w - \eta_t \partial_w \bar{l}(x_i, x_{i+1}, y_i, y_{i+1}, x'_i, x'_{i+1}, f)$ where $f(x) = \langle\phi(x), w\rangle$
  Project $w$ onto $A$ via $w \leftarrow \mathrm{argmin}_{\bar{w} \in A} \|w - \bar{w}\|$.
**end for**
```

Special Cases

- Mean matching for classification/class balancing constraint (Joachims 1999)

$$\mu[f(X)] = \frac{1}{m}\sum_{i=1}^{m}\langle f(x_i), \cdot\rangle = \frac{1}{m'}\sum_{i=1}^{m'}\left\langle f(x'_i), \cdot\right\rangle = \mu[f(X')].$$

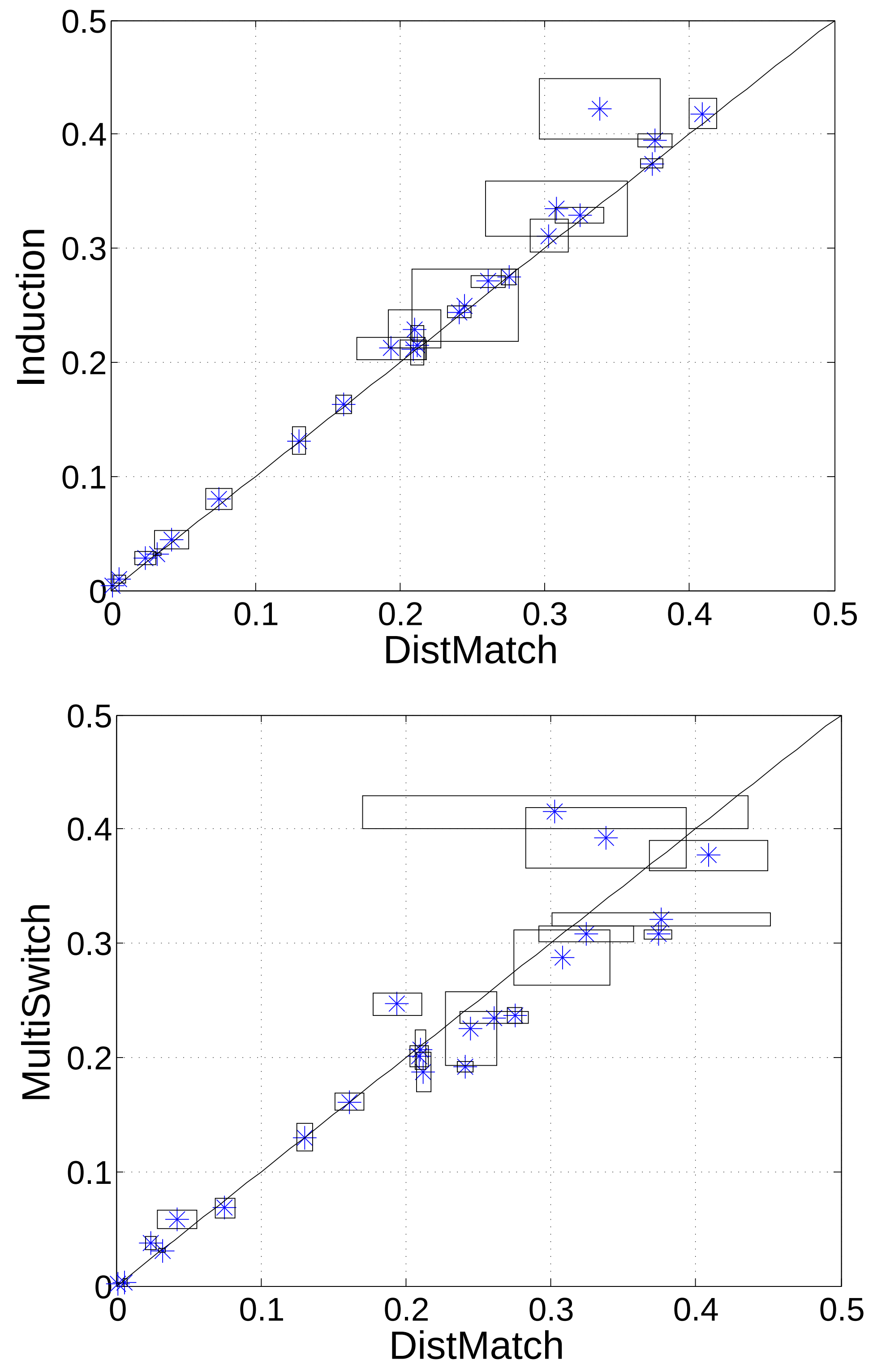
- Distribution matching for classification (Gärtner et al. 2006)
- Distribution matching for regression (Le et al. 2006)

Applications

Small-Scale Classification

Binary Classification

- Dataset: 23 binary problems from UCI/LibSVM repository
- A Gaussian RBF kernel is used for the distribution matching term
- Results are averaged across 5 different runs
- Performance comparison with Multi Switch Transductive SVM (Sindhwani & Keerthi 2006)



Multiclass Classification

- Dataset: 5 multi-class problems from UCI/LibSVM repository
- Performance comparison with a Gaussian processes based transductive algorithm (Gärtner et al. 2006)
- Same experimental setup as binary experiments

dataset	m	classes	Induction	DistMatch	GPDistMatch
usps	730	10	0.143±0.021	0.125±0.019	0.140±0.034
satimage	620	6	0.190±0.052	0.186±0.037	0.212±0.034
segment	693	7	0.279±0.090	0.206±0.047	0.181±0.020
svmguide2	391	3	0.280±0.028	0.256±0.020	0.231±0.018
vehicle	423	4	0.385±0.070	0.333±0.048	0.336±0.060

Large-Scale Multi-Category Classification

- Dataset: DMOZ ontology of topics (`http://www.dmoz.org`)
- #categories: 100, #observations: (up to) $3.2 \cdot 10^6$, #features: $1.3 \cdot 10^6$

*Scaling the algorithm with respect to the **training** set size*

training / test set size	50,000	100,000	200,000	400,000	800,000	1,600,000
induction	0.365	0.362	0.337	0.299	0.300	0.268
transduction	0.344	0.326	0.330	0.288	0.263	0.250

*Scaling the algorithm with respect to the **test** set size*

test set size	100,000	200,000	400,000	800,000	1,600,000
induction	0.358	0.358	0.357	0.357	0.357
transduction	0.326	0.316	0.306	0.322	0.329

Named Entity Recognition

- Dataset: Japanese named-entity recognition from the CRF++ toolkit
- #sentences: 716 and #annotated named entities: 17
- 1D chain CRFs with first order Markov dependency between name tags.
- Distribution matching is enforced on the clique potentials joining words and labels ((x_i, y_i))

	Accuracy	Precision	Recall	F1 Score
induction	96.82	84.15	72.49	77.89
transduction	97.13	84.46	75.30	79.62

Base NP Chunking

- Dataset: CoNLL-2000 base NP chunking from the CRF++ toolkit
- #sentences: 900 and the task is to label each word indicating whether the word is outside, starts, or continues a chunk
- Same experimental setup as in named entity experiments

	Accuracy	Precision	Recall	F1 Score
induction	95.72	90.99	90.72	90.85
transduction	96.05	91.73	91.97	91.85

Summary

- We propose a transductive algorithm which is simple , scalable and applicable to classification , regression and structured estimation.
- Experiments are performed on small scale classification problems, large scale multi-category settings (involving $3.2 \cdot 10^6$ observations and 100 categories), and chunking and named entity structured prediction.

References

- A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. Technical Report 157, MPI for Biological Cybernetics, 2008.
- V. Sindhwani and S.S. Keerthi. Large scale semi-supervised linear SVMs. *SIGIR '06*, pages 477–484, New York, NY, USA, 2006.
- T. Gärtner, Q.V. Le, S. Burton, A. J. Smola, and S. V. N. Vishwanathan. Large-scale multiclass transduction. *NIPS 18*, pages 411–418, Cambride, MA, 2006.
- T. Joachims. Transductive inference for text classification using support vector machines. *ICML*, pages 200–209, 1999.
- Q.V. Le, A.J. Smola, T. Gärtner, and Y. Altun. Transductive gaussian process regression with automatic model selection. *ECML*, volume 4212 of *LNAI*. 306-317, 2006.