

Correcting Sample Selection Bias by Unlabeled Data

J. Huang, A. Smola, A. Gretton, K. Borgwardt and B. Scholkopf

❖ The problem

- ❖ we consider the learning problem that the training samples are drawn from $\Pr(x, y)$ which is different from the distribution $\Pr'(x, y)$ that test data are from
- ❖ data is collected in a biased manner while test is usually performed over a more general target population, e.g., clinic data

❖ Our solution

- ❖ use unlabeled data to direct a de-biasing procedure
- ❖ no estimation of biased densities or selection probabilities
- ❖ no assumption of knowing probabilities of different classes
- ❖ a non-parametric method, can handle high dimensional data
- ❖ efficient computation by solving a simple QP problem
- ❖ extend to many classification and regression algorithms



MAX-PLANCK-GESellschaft



Correcting Sample Selection Bias by Unlabeled Data

J. Huang, A. Smola, A. Gretton, K. Borgwardt and B. Scholkopf

❖ Motivation

- ❖ importance sampling: one can swamp distributions for risk expectations.
- ❖ minimize a reweighed empirical risk with regularizer.

❖ Theorems

- ❖ The reweighed set of observations will behave like one drawn from training distribution with effective sample size.
- ❖ with high probability, minimizing the reweighed empirical risk will also minimize an upper bound on the expected risk on the test set.

❖ Experiments

- ❖ test on various biased situations on benchmark datasets.
- ❖ cross-platform microarray classification.

