Abstract

- Proliferation of machine learning algorithms in diverse domains
- Necessitates working with non-explicit features
- Notions of distance/similarity more natural than hand-coded features
- Co-authorship graphs, Earth-mover’s distance
- Typically end up with non-PSD similarity measures (kernels)
- Goal: a model of learning with arbitrary similarity measures

Our Contributions

- Develop a general notion of goodness for similarity measures
- Propose algorithms that make optimal use of any such measure
- Provide classifiers with provable error bounds

What is a good similarity function?

- Suitability of a similarity function to a given classification problem
- Points with same label should be more similar than dissimilarly labeled points
- Yield classifiers with bounded generalization error

Theorem 2 guarantees allow use of any Lipschitz loss function—hinge, logit, quadratic ...

Examples of Transfer functions

- Identity function
- Sign function
- Ramp function
- Sigmoid function

This model encompasses BBS and DBOOST
- BBS uses identity, DBOOST uses sign as transfer function

Generalization Guarantee

- Modify Definition 1 to include a loss function $L$: require $L(f) = \mathbb{E}_{x,y \sim D} [L(f(x), y)] \leq \epsilon$
- Definition 1 can be shown to use the loss function $L(x) = \mathbb{E}_{f \sim \mathcal{F}}[\ell(x, f(x))]$

Theorem 2. (Uniform Convergence Bound)

- If $f$ is a set of transfer functions with an $\epsilon$-net with respect to infinity norm at scale $r^{-1}$ of size almost $N \{f \colon \|f\|_\infty \leq \epsilon \}$, then for any $\gamma > 0$, $n = O(N r^2)$ random landmark pairs suffice to output a classifier with expected loss less than $\epsilon + \gamma$ with probability $1 - \delta$.

- Guarantees that suitability of $f$ will be evident in $\mathbb{R}^d$ for all $f \in \mathcal{F}$ with a single $\ell$
- Validates the use of ERM style algorithms to select a good $f$ from $\mathcal{F}$
- e.g. possible to tune parameter $\sigma$ in sigmoid transfer function

Landmark Selection

- On small datasets, choice of transfer function can lead to overfitting
- DSELECT: heuristic for landmark selection that improves performance
  - If landmarks clumped together then all training points get same embedding
  - Need to promote diversity among landmark points
  - Incrementally select landmark points in a greedy manner
  - At each step choose a point that is least similar to already chosen points
  - Form pairs out of these points later on to get landmark pairs

Experimental Results : UCI Benchmark Datasets

- Average dataset size 1320: validation can be performed without overfitting
- DSELECT does not help on large datasets: FTUNE alone performs well