The sample complexity of agnostic learning under deterministic labels

Shai Ben-David and Ruth Urner

COLT 2014

June 13, 2014
A non-investigated corner in (agnostic) PAC theory

<table>
<thead>
<tr>
<th></th>
<th>Bayes in $H$</th>
<th>Agnostic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic labels</td>
<td>$\frac{\text{VCdim}(H)}{\epsilon}$</td>
<td>$?$</td>
</tr>
<tr>
<td>Probabilistic labels</td>
<td>under Tsybakov noise: $\frac{\text{VCdim}(H)}{\epsilon^{2-\alpha}}$</td>
<td>$\frac{\text{VCdim}(H)}{\epsilon^2}$</td>
</tr>
</tbody>
</table>
Why care?

- Modern machine learning tools allow the use of a very large number of features
  - This renders the labeling function deterministic (but not necessarily in class $H$)

- Fundamental problem that leads to better understanding of where the difficulty of learning does (not) come from
Results: Noise in labels is not the only source of difficulty

Deterministic labels do not imply fast rates - for most common classes (e.g., linear classifiers) learning deterministic labelings is as hard as learning with arbitrary noise.
Sample complexity is not fully determined by VC-dimension of the class

- We show that there are classes with sample complexity $\Theta\left(\frac{\text{VCdim}}{\epsilon^2}\right)$ and classes with sample complexity $\tilde{\Theta}\left(\frac{\text{VCdim}}{\epsilon}\right)$
- We provide a full characterization
ERM learning is not always optimal (in fact, for some classes, any proper learner has sub-optimal sample complexity).
Results: Unlabeled data is useful

There is a generic Semi-Supervised ERM that does achieve optimal sample rates, upon having access to unlabeled samples.
Summary

- Low noise does not imply fast rates
  - Need to identify different data properties/“easiness criteria”

- Learning rates in the deterministic labeling domain break some of the basic rules ML theory almost takes for granted
See you at the poster!