Active Learning using decision trees

June 6, 2019

1 Definitions

- **Standard definitions** Let $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{-1, +1\}$ be a space of labeled examples on which a joint distribution D is defined. Denote by (X, Y) a random variable corresponding to a single labeled example.
- **Set of Specialists** Let $\mathcal{S}$ be a set of subsets of X (specialists).
- **Conditional Bias** We define the *conditional bias* for $x \in X$ as $\eta(x) = E_D(Y|X = x)$, similarly we define bias for $A \in \mathcal{S}$ to be $\eta(A) = E_D(Y|X \in A)$
- **Transductive Framework** We are given the complete set of unlabeled instances in advance, the examples are sampled according to the marginal of D on $\mathcal{X}$. We denote the set of n unlabeled instances by $\mathbf{X} = (x_1, \ldots, x_n)$. Active learning proceeds by making queries for the label of instaces $x \in X$. The label is generated according to the true conditional bias $\eta(x)$, in other words, querying the same example multiple times can generate different labels. The goal of the algorithm is to minimize the probability of error with respect to the uniform distribution over *unlabeledSet*, i.e. to converge to the optimal bayes rule on $\mathbf{X}$ which is $\text{sign}(\eta(x))$.
- **Dense Set of Specialists** We say that $\mathcal{S}$ *are dense in* X if for any $x \in X$ there exists a sequence of specialists $A_i^x \in \mathcal{S}$, $A_1^x \supseteq A_2^x \supseteq \cdots$, denoted $S(x)$ such that $\bigcap_{i=1}^{\infty} A_i^x = \{x\}$.
- **Consistency** We say that the distribution D is consistent with the set of specialists $\mathcal{S}$ if for all $x \in X$ (can remove sets with zero-prob neighborhood). The biases for the sequence $S(x)$ converge to the bias on x:

$$\lim_{i \to \infty} \eta(A_i^x) = \eta(x)$$

- **Determined prediction** Consider a sequence $S(x)$ and a set of labeled examples. We say that a specialist in the sequence is *determined* if it has a large enough empirical bias to trigger the AKNN rule (or something like it). The empirical bias is calculated only on examples that were sampled

uniformly at random from the specialist. The smallest (highest index) specialist in $S(x)$ determines the prdiction of the sequence wrt $S(x)$.

Characterizing conditions for asymptotic consistency It remains to characterize consistency in terms of properties of the space X, in other words, replacing the convergence of concentric converging balls used in the Lesbegue theorem which we use for proving asymptotic consistency.

2 Sequences from random trees

To enable active learning we construct predictors which can agree or disagree. In this section we propose a particular construction that we call random trees (whould we call it a random forest?)

First, we replace the AKNN rule, which is expensive to compute, with partition trees which are fast. We will pay for the improved speed by worse convergence guarantees.

The trees can be constructed in many ways: KD-trees, RP-trees, C4.5, CART etc. The main condition I wish to enforce is that the diameter of the nodes goes to zero with their depth in the tree. I believe that this is enough to ensure consistency under mild requirements on the space X and the distribution D.

Another condition is that the tree construction algorithm is randomized. In other words, if the algorithm is run multiple times on the same dataset but different random seeds, a different tree is constructed. This defines a probability distribution $\mathcal{T}$ over trees. Properties of this distribution will impact the effectiveness of the active learning in ways that are yet to be characterized.

A given tree T and a given example x define a path in the tree. The tree nodes along this path define a sequence of spcialists $S^T(x)$. Selecting the tree at random $T \sim \mathcal{T}$ defines a distribution over the sequeces that converge to x. We denote the induced distribution over sequences converging to x as $\mathcal{Q}(x)$

3 Active Learning

We asssume a transductive setup. In other words

We generate N random trees, thereby defining for each point $x \in X$ N specialist sequences $S_1(x), \ldots, S_N(x)$

Our algorithm works in epochs, at each epoch we use a different distribution to sample M new query points. We use a mixed strategy to sample the M points. Specifically, we select $M/2$ points uniformly at random from $\mathbf{X}$. We sample the other $M/2$ points from a uniform distribution over the *Knuck* set (stands for "known unknown".

Knuck is defined as follows. First we identify the "uncertain centers" which are the points $x \in X$ such that the predictions of $S_1(x), \ldots, S_N(x)$ are not all the same (can probably be weakened). In other words, there are determined

specialists which give inconsistent predictions on x. Lets call those specialists the *cover* of the uncertain centers.

Knuck is defined as the union of the specialist from all of the uncertain covers.