Mathematical Aspects of Deep Learning

Mathematics of Deep Learning: Lecture 7 – Recovering Tree Models

Transcribed by Paxton Turner (edited by Asad Lodhia, Elchanan Mossel and Matthew Brennan)

Tree Reconstruction II

We prove Claim 2 which was left over from last time.

Lemma. *Fix a constant $\ell \in \mathbb{N}$. Given a d-ary tree with percolation parameter $\lambda < 1$ which is sufficiently large, then*

$$\mathbb{P}_\lambda \left[\text{root is in an } \ell - \text{diluted } (d^\ell - 1)\text{-ary open tree}\right] \geq 1 - \epsilon$$

Proof. Let

$$f(p) := \mathbb{P}[\text{Bin}(d^\ell, p) \geq d^\ell - 1],$$

and note that f is monotone in p. We see that $f(1) = 1$, and $f(p) = p^{d^\ell} + d^\ell(1-p)p^{d^\ell - 1}$. Also

$$f'(1) = d^\ell + d^\ell(d^\ell - 1 - d^\ell) = 0.$$

Thus taking p^* sufficiently close to 1 yields that $f(p) > p$ for all $p^* < p < 1$. Also choose p^* to satisfy $p^* > 1 - \epsilon$. Now choose $\lambda < 1$ such that

$$q = \mathbb{P}[\text{root is connected to all } d^\ell \text{ nodes in the } \ell \text{ level}] \geq p^*/f(p^*).$$

Now we apply a recursive argument. Let

$$p_i = \mathbb{P}[\text{root is in an } \ell - \text{diluted } (d^\ell - 1)\text{-ary open tree for } i \cdot \ell \text{ levels}].$$

We will show by induction that $p_r \geq p^*$ for all r, which implies the lemma. Note that $p_0 = 1$ and

$$p_{r+1} \geq q \cdot \mathbb{P}\left[\text{Bin}(d^\ell, p_r) \geq d^\ell - 1\right] = \frac{p^* f(p_r)}{f(p^*)} \geq p^*$$

since f is monotone and $p_r \geq p^*$ by the induction hypothesis. The lemma is proved. ■

Recovering the Tree Structure

We begin with an ℓ-level full binary tree and the size of the alphabet to be $q = \infty$. We generate k independent samples of the broadcasting procedure with parameter λ on the tree.

In contrast to the last lecture, today, we do not know the *tree structure* , *i.e.* we only receive the colors of the nodes and do not know which nodes are siblings, cousins, etc. Our goal is to determine how large k must be to recover the tree structure.

In practice, these trees are phylogenetic trees, and the information we have are the DNA/amino acid sequences. Thus, usually $q = 4$, or $q = 20$.

Question : How large should k, ℓ, λ be so that we recover the tree correctly as $\ell \to \infty$?

1. Recovering Distances Using Correlations

If u, v are graph distance $2r$ apart, then

$$\mathbb{P}[\sigma(u) = \sigma(v)] = \lambda^{2r}.$$

How large should k be so that all distances can be computed accurately?

) $= 2\ell - 2$, then taking a union bound gives

$$\mathbb{P}[\exists \text{ sample with } \sigma(v) = \sigma(u')] \leq 1/2.$$

Then we can't say if $d(u, v) = 2\ell$ or $d(u, v) = 2\ell - 2$. If we use concentration, then

$$k \geq \log(2^{\ell}) \left(\frac{1}{\lambda^{\ell}} \right)^2 \log(1/\delta)$$

allows us to compute distances correctly with probability $\geq 1 - \delta$. One can show this with Chernoff bounds, and the details are omitted here. If we compute all of the distances between the leaves correctly, then we trivially can recover the tree structure. However, as noted above, this requires at least $k \geq c(1/\lambda)^{2\ell} = |T|^{\beta}$ samples.

2. Recursive "Deep" Algorithms for $q = \infty$

Computing all of the distances between the leaves seems like too strong of a requirement. This raises the question: Is there a deep algorithm that tries to infer the labels at the internal nodes sample by sample that can recover the tree with fewer samples? We will outline such an algorithm when $2\lambda > 1$.

After detecting siblings, we want to understand the probability of being able to recover colors at their parents sample by sample. Consider first a small example. Suppose we have a 2-level binary tree with leaves labeled. If there is a 1 in both the left and right branches, then with probability 1, the root is colored 1. However, if no color appears in both branches, we mark the node with a question mark ?. Returning to our original problem, we can first apply this procedure to identify siblings among the leaves and the labels of their parents sample by sample, and then apply this to the nodes at level $\ell - 1$, and so on.

Claim. *If $2\lambda > 1$, then there exists $C(\lambda) > 0$ such that*

$$\mathbb{P}[? \text{ recovered at level} r] \leq 1 - C(\lambda).$$

Proof. Note that an internal node is not a ? if its label survives in its left and

$$\begin{aligned}\mathbb{P}[\text{a node at level } r \text{ is not ?}] &\geq \mathbb{P}[\text{root connected to 2 children}] \\ &= \mathbb{P}[\text{branching process survives}]^2 \\ &= C(\lambda) > 0\end{aligned}$$

if $2\lambda > 1$. ■

How many samples are needed to identify all siblings correctly?

$$k = \frac{(1-\lambda)^4}{\lambda} \log\left(\frac{2^\ell}{\delta}\right)$$

suffices for error $\leq \delta$. It turns out that there is a similar bound for recovering the tree if $2\lambda > 1$,. As observed above, if we purely used distance methods, then we require $k \geq c(1/\lambda)^{2\ell} = |T|^\beta$. However, if $2\lambda > 1$, then our "deep" algorithm requires only $k = \log(|T|)$ samples.

If $2\lambda < 1$, then what is true? The goal is to show you can't distinguish between the uniform distribution. Heuristically, we must estimate the probability of getting all question marks ? for the lower bound.

Lower bound proof idea: Start with four trees having roots a, b, c, and d. We want to couple the broadcasting procedure on these trees so that they produce the same data. In such a coupling, we can show

$$\mathbb{P}[\text{generating different samples}] \leq 4(2\lambda)^\ell.$$

Thus,

$$\mathbb{P}[\text{coupling succeeds}] \geq 1 - 4k(2\lambda)^\ell.$$

Thus if we want recover the tree correctly with probability $\geq 1/\delta$, we need $k \geq c(2\lambda)^\ell = \Omega(|T|^\ell)$ samples (if $\lambda < 1$).

Remark. If $2\lambda < 1$, distance methods do not give optimal sample complexity.

We make the following general conjecture.

Conjecture : Recursive methods should give optimal complexity.

3. A Recursive Algorithm for $q = 2$

Next we consider a new model. Now we have a full binary tree, and on all edges, we copy with probability η. Our alphabet now has size $q = 2$, and in particular, we set the "letters" to be $\{-1, 1\}$.

Using distance methods, we cannot distinguish between

$$\mathbb{E}[\sigma(u)\sigma(v)] = \eta^{2\ell}, \quad \mathbb{E}[\sigma(v')\sigma(v)] = \eta^{2\ell-2}$$

The lower bound for distance methods here is $k \geq c/\eta^{2\ell}$, and this is more or less tight. As before, we will instead use a recursive approach: first we identify siblings, and then estimate colors of the parents. To identify parents, use the correlations

$$\mathbb{E}[\sigma(v)\sigma(v')] = \eta^2$$

if v and v' are siblings. We use correlations to recover distances up to a constant i levels above. And this requires $\approx \log(2^\ell)$ samples to concentrate. Then we can use majority vote to recover the color of the parent. By a similar computation to the 2nd moment method from last lecture, if $2\eta^2 > 1$,

$$\mathbb{P}[\text{Maj}(x_1, \ldots, x_{2^r}) = x_{\text{root}}] \geq \frac{1}{2} + \epsilon$$

for all r. Thus we define

$$\hat{\sigma}_i = \text{Maj}\left(\sigma_i(u) : u \text{ leaf below } v\right).$$

Note that

$$\mathbb{E}[\sigma_{\text{root}}\hat{\sigma}_{\text{root}}] = \rho > 0 \Rightarrow \mathbb{E}[\hat{\sigma}_i(v)\hat{\sigma}_i(w)] = \eta^{2d(v,w)}\rho^2.$$

Results If $2\eta^2 > 1$, we can recover the tree with $k = \log(|T|)$ samples. Otherwise, $k \geq |T|^{\beta}$. Compare this to the thresholds $d\eta^2 > 1$ for count reconstruction and $d\lambda_2^2 > 1$ for general broadcasting.

4. Four point method for detecting distances

We define a new twist on the previous model. The parameter η is now non-uniform, but we still require it to satisfy $2\eta^2(e) \geq 1 + \epsilon$ for all edges e. In this general set-up, one has to be more careful with detecting distances/siblings.

Four point method : This is a trick from phylogeny. For the purposes of detecting distances, there are 3 possible 2 level binary trees. The neighbors are either $\{(A, B), (C, D)\}$, $\{(A, C), (B, D)\}$, and $\{(A, D), (B, C)\}$. We use pairwise distances to determine which tree it is. Set

$$Dist(u, v) = \log\left(\frac{1}{\mathbb{E}[\sigma_u \sigma_v]}\right).$$

Claim. *The correct pairing of A, B, C and D is the one for which*

$$Dist(X, Y) + Dist(W, V)$$

is minimal, where (X, Y, W, V) is some permutation of (A, B, C, D).

Now to determine siblings of v, find all vertices with

$$|Dist(u, v)| \leq 3\log(1/\sqrt{2}),$$

and apply the four point method on these vertices.

One can also run into an issue when estimating the colors of parents. To solve this, use an ℓ level recursive estimator.

Theorem. *If $2\eta^2 > 1$, then there exists $\ell(\eta)$ such that using an ℓ-level recursive majority gives estimator $\hat{\sigma}$ such that*

$$\min_{\pi} \mathbb{P}[\pi\sigma = \hat{\sigma}] \geq 1/2 + \epsilon.$$

In the next lecture, we discuss hierarchical generative models.

elmos / July 6, 2017