Rates for Inductive Learning of Compositional Models

Adrian Barbu
Department of Statistics
Florida State University
Joint work with Song-Chun Zhu and Maria Pavlovskaia (UCLA)
Bernoulli Noise

- Appears for thresholded responses of
  - Gabor filters
  - Learned part detectors
Bernoulli Noise

We will focus on the following simplified setup:

- The parts to be learned are rigid
- Bernoulli noise in the terminal nodes
  - Foreground noise probability $p$ to switch from 1 to 0 (due to occlusion, detector failure, etc)
  - Background noise probability $q$ to switch from 0 to 1 (due to clutter)
The AND/OR Graph

The AND/OR graph (AOG) is

- a hierarchical representation
- used to represent objects through intermediary concepts such as parts
- the basis of the generative image grammar (Zhu and Mumford, 2006)

- AND nodes = composition out of parts
- OR nodes = alternate configurations (e.g. deformations)
The AND-OR Graph

- Defined on $\Omega = \{0, 1\}^n$
  - The space of thresholded filter responses

- Is a Boolean function
  $$g : \Omega \rightarrow \{0, 1\}$$
  obtained by composition of AND and OR boolean functions

- Can be represented as a graph with AND and OR nodes

- Other AOG formulations:
  - Bernoulli AOG
  - Real AOG
AND Node

Composition of a concept from its parts

Example

- Dog face
  - Eyes, ears, nose, mouth ...
- Dog Ears of type A
  - Sketch type 5 at position (2,0)
  - Sketch type 8 at position (1,2)
  - ...

elements (alphabet)
OR Node

- Alternative representations
- Example
  - Dog head
    - Side view
    - Frontal view
    - Back view
  - Dog Ears
    - Type A
    - Type B
AOG parameters

- Maximum depth $d$
  - Usually at most 4
- Maximum branching numbers $b_a, b_o$ for AND/OR nodes respectively
  - $b_a$ usually less than 5
  - $b_o$ usually less than 7
- Number of terminal nodes $n$, $\Omega = \{0, 1\}^n$
- Let
  $$\mathcal{H}(d, b_a, b_o, n) \subset \{0, 1\}^\Omega$$

the space of AOGs with
- max depth $d$
- max branching numbers $b_a, b_o$
- $n$ terminal nodes
Example: Dog AOG

- Depth $d=2$
- Branching numbers $b_a=7$, $b_o=2$
- Number of terminal nodes $n=15 \times 15 \times 18 = 4050$
The AND-OR Graph

- Object composed of parts with different possible appearances

Samples from the dog AOG
Synthetic Bernoulli Data

- Samples from dog AOG corrupted by Bernoulli noise
  - Switching probability $q$
**Concept**

- Given instance space $\Omega$
- A **concept** is a subset $C \subset \Omega$

- Can also be represented as a target function $f: \Omega \rightarrow \{0, 1\}$

- There are equivalent representations

\[
C = \Omega_f = \{x \in \Omega, f(x) = 1\}
\]
Concept Learning Error

The true error $\text{err}_\mu(h, C)$ of hypothesis $h$ with respect to concept $C$ and distribution $\mu$ is the probability that $h$ will misclassify an instance drawn at random from $\mu$

$$
\text{err}_\mu(h, C) = \mu(C \triangle \Omega_h)
$$
Capacity of AOG

- $\mathcal{H}(d, b_a, b_o, n) \subset \{0, 1\}^\Omega$ is a finite space
- From Haussler’s Theorem
  
  \[ m \geq \frac{1}{\epsilon} (\ln |\mathcal{H}| + \ln \frac{1}{\delta}) \]

  examples are sufficient for any consistent hypothesis $h$ to have $err_\mu(h, C) < \epsilon$ with probability $1-\delta$
- Define the capacity as
  
  \[ C(d, b_a, b_o, n) = \ln |\mathcal{H}(d, b_a, b_o, n)| \]
- We have the bound
  
  \[ C(d, b_a, b_o, n) \leq (b_a b_o)^d \ln n \]
Example: 50-DNF

- 18 types of sketches on a 15x15 grid
- Totally $n = 15 \times 15 \times 18 = 4050$

$$\Omega = \{0, 1\}^{4050}$$

- Assume at most 50 sketches present
- There are $\sim 4050^{50}$ templates with 50 sketches
- $k$-DNF space size is about $2^{4050^{50}}$
- Capacity is $\sim 10^{180}$
- Too large to be practical
Example: $C(2, 5, 5, 4050)$

- Same setup $\Omega = \{0, 1\}^{4050}$
- Space of AOG $\mathcal{H}(2, 5, 5, 4050)$
- Max depth 2, max branching number 5
- Capacity is $C(2, 5, 5, 4050) < 25^2 \ln 4050 \approx 5192$

So $m \geq \frac{1}{\epsilon} (5192 + \ln \frac{1}{\delta}) \approx 5200/\epsilon$

examples are sufficient for any hypothesis consistent with the training examples to have $err_\mu(h, C) < \epsilon$ with 99.9% probability
Capacity of AOG with Localized Parts

- Consider the subspace

\[ \mathcal{H}(d, b_a, b_o, n, l) \subseteq \mathcal{H}(d, b_a, b_o, n) \]

where the first level parts are localized:

- First terminal node can be anywhere
- The other terminal nodes of the part are chosen as one of the \( l \) nodes close to the first one

- In this case we have

\[ C(d, b_a, b_o, n, l) \leq b_a^{d-1} b_o^d \ln(n l^{b_a-1}) \]
Example: $C(2, 5, 5, 4050, 450)$

- Same setup $\Omega = \{0, 1\}^{4050}$
- Space of AOG $\mathcal{H}(2, 5, 5, 4050, 450)$
- Max depth 2, max branching number 5
- Locality in a 5x5 window ($l = 5 \times 5 \times 18 = 450$)

- Capacity is
  
  $C(2, 5, 5, 4050, 450) < 5 \cdot 5^2 \ln 4050 \cdot 450^4 \approx 4093$

- Reduction from 5192
Supervised Learning AOG

- **Supervised setup:**
  - Known And/OR Graph structure
  - Object and parts are delineated in images
    - E.g. by bounding boxes
  - Part appearance (OR branch) is not known

- **Need to learn:**
  - Part appearance models
    - OR templates and weights
    - Noise level
Two Step EM

EM for mixture of Bernoulli templates [Barbu et al, 2013]

- Similar to EM of Mixture of Gaussians [Dasgupta, 2000]

Say we want \( k \) clusters in \( \{0,1\}^n \)

We will start with \( l \sim O(k \ln k) \) clusters

Two Step EM Algorithm

1. Initialize \( \mu_i, i=1,\ldots,l, \) as random data points, \( w_i = 1/l, \)
   \[
   \sigma = \min_{i,j} \| \mu_i - \mu_j \|_1
   \]

2. One EM step

3. Pruning Step

4. One EM Step
Two Step EM

Pruning step:

1. Remove all clusters with $w_i < 1/4l$
2. Selected $k$ centers furthest from each other
   1. Add one random $\mu_i$ to $S$
   2. For $j=1$ to $k-1$
      Add to $S$ the center with maximum distance $d(\mu_i, S)$

\[
d(\mu_i, S) = \min_{j \in S} \|\mu_i - \mu_j\|_1
\]
Theoretical Guarantees

- Under certain conditions C1-C3

**Theorem 1.** If \( m \) examples are generated from a mixture of \( k \) Bernoulli templates under Bernoulli noise of level \( q \) and \( w_i > w_{\text{min}} \) for all \( i \). Let \( \epsilon, \delta \in (0, 1) \). If conditions C1 – C3 hold and in addition the following conditions hold

1. The initial number of clusters is \( l = \frac{4}{w_{\text{min}}} \ln \frac{2}{\delta w_{\text{min}}} \).

2. The number of examples is \( m \geq \frac{8}{w_{\text{min}}} \ln \frac{12k}{\delta} \).

3. The separation is \( c > \frac{4}{nB} \ln \frac{5n}{\epsilon w_{\text{min}}} \).

4. The dimension is \( n > \max \left( \frac{3}{\min(c, 0.5)E^2} \ln \frac{12(m + 1)^2}{\delta}, \frac{6k}{\delta} \right) \).

Then with probability at least \( 1 - \delta \), the estimated templates after the round 2 of EM satisfy:

\[
\|T_i^{(2)} - P_i\|_1 \leq \|\text{mean}(S_i) - P_i\|_1 + \epsilon q
\]
Noise Tolerant Parts

- Part learned using Two-Step EM:
  - Mixture centers $T_i$
  - Mixture weights $w_i$
  - Noise level $\hat{q}$

- Obtain noise tolerant part model:

$$p(x) = (1 - \hat{q})^d \sum_{i=1}^{k} w_i \left( \hat{q} / (1 - \hat{q}) \right) \|x - T_i\|_1$$

- Detection: compare $p(x)$ with a threshold
- For one mixture center, same as comparing $\|x - T\|_1$ with a threshold
Noise Tolerant Parts

For a single mixture center, part of size $d$ and threshold $k$:

- Probability of missing the part:
  
  $$p_{10} = 1 - \sum_{i=0}^{k} \binom{d}{i} q^i (1 - q)^{d-i}$$

- Probability of a false positive
  
  - assuming empty background and all 1 template

  $$p_{01} = \sum_{i=0}^{k} \binom{d}{i} q^{d-i} (1 - q)^i$$

- Example: $d=9$, $q=0.1$, then $p_{10} = p_{01} < 0.001$. 
Supervised Learning AOG

Recursive Graph Learning

- Learn bottom level parts first with two-step EM
- Detect the learned parts in images
  - Obtain a cleaner image
- Learn next level of the graph using two-step EM
Part Sharing Experiment

Setup:

- Dog AOG data with Bernoulli noise
- 13 Noise tolerant parts
  - previously learned from data coming from other objects (cat, rabbit, lion, etc)

- Two learning scenarios
  - Learn the dog AOG from the 13 parts
  - Learn the dog AOG directly from image data
    - Learn parts with two-step EM first
    - Learn AOG from parts
Part Sharing Experiment

Conclusion:

- Learning from parts is easier than learning from images
- Part sharing helps
Conclusions

- Capacity of AOG space is much smaller than k-CNF or k-DNF
  - Much fewer examples needed for training
  - Using part locality helps

- Learning OR components using two-step EM works
  - Has theoretical guarantees when
    - OR components are clearly different from each other
    - Noise is not very large
    - Dimensionality is large enough
    - Sufficiently many examples

- Part sharing improves learning performance