Learning with large examples: A short survey

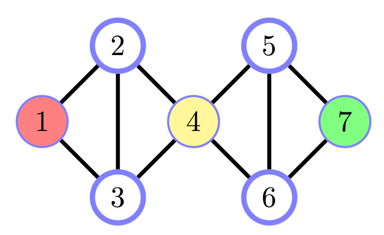
Sanyam Kapoor (sk6876@nyu.edu), Chhavi Yadav (cy1235@nyu.edu)
December 19, 2017
Courant Institute, NYU

Motivation

- Most data has implicit or explicit internal structure
- *Structured Prediction* - learning structures from input examples
- e.g. A large satellite image for segmentation, POS tagging in large corpus or analysis of large social networks (note exponential output space)
- Our focus - *Can we generalize from a single example with large internal structure?*
- Analysis and critique of two approaches to obtain generalization bounds for such scenarios

Graph-based Hypotheses $\mathcal{H}$

Figure 1: A Markov Network, e.g. $1 \perp\!\!\!\perp 7|4$ [Barber, 2012]



- Use Markov random fields $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$
- We define a graph based hypothesis for a graph G as

$$\mathbb{P}_{G,h}(Y_i|\mathbf{X}, \mathbf{Y} \setminus Y_i) = \mathbb{P}_{G,h}(Y_i|\mathbf{X}_{\mathcal{N}^d(i)}, \mathbf{Y}_{\mathcal{N}^d(i)}) \tag{1}$$

where $\mathbf{Y} \triangleq h_G(\mathbf{X})$ and $\mathcal{N}^d(i)$ are the neighbors

Modelling the problem

- Graphs are a natural way of modelling dependencies
- [London et al., 2013] proposes $\mathbf{\Theta} \in \mathbb{R}^{n \times n}$ matrix to model the *decay of dependence* over the graph for some ordering of $\mathbf{Z}$

$$\eta_{i,j} = \sup_{i} \|\mathbb{P}(\mathbf{Z}_{j:n}|\mathbf{Z}_{1:i-1}, z_i) - \mathbb{P}(\mathbf{Z}_{j:n}|\mathbf{Z}_{1:i-1}, z_i')\|_{TV} \qquad (2)$$

- We aim to see how empirical risk concentrates around the true risk (the *generalization error bound*)
- Key considerations - Weak dependence, β-*stability*, (M, λ)-admissibility, Rademacher Complexity ($\mathfrak{R}$)

Main Result

- $\sup_{h \in \mathcal{H}} \|h(\mathbf{z}) - h(\mathbf{z}')\|_1 \leq \beta$ (*β-stability*) for hypothesis class $\mathcal{H}$
- $|\ell(y, \hat{y}) - \ell(y', \hat{y})| \leq M$ and $|\ell(y, \hat{y}) - \ell(y, \hat{y}')| \leq \lambda \|\hat{y} - \hat{y}'\|_1$ ((M, λ)-admissibility) for the loss function ℓ
- Generalization bound for such graph-based Hypothesis

$$R_G(h) \leq \hat{R}_G(h, \mathbf{Z}) + 2\lambda \sum_{j=1}^{k} \mathfrak{R}_n(\mathcal{H}^j) + (M + \lambda\beta)\|\mathbf{\Theta}_n\|_\infty \sqrt{\frac{ln(1/\delta)}{2n}} \tag{3}$$

Discussion

- Intuitively, complexity of the Hypothesis set governed by the complexity of dependencies in the model
- The stability term captures how stable the predictions are, ideally we want $\beta = \mathcal{O}(1)$
- $||\mathbf{\Theta}_n||_\infty$ captures the slowest decay of dependence, again ideally we want $||\mathbf{\Theta}_n||_\infty = \mathcal{O}(1)$
- In practice, hard to qualify such constraints.
  [London et al., 2013] shows *TSMs* to obey but only with number of clique templates $|\mathcal{T}| = 1$ as $\beta = O(1)$ and $\Re_n(\mathcal{H}) = O\left(\sqrt{lg(n)/n}\right)$

A Different Approach

- PAC bounds for MLE (Maximum Likelihood Estimation)
- [He and Zhang, 2014] propose a combinatorial approach to calculate the sample complexity via a bound between the calculated estimate MLE $\hat{\theta}_n$ and the true parameter θ^*
- Key considerations - Finite Distance Dependence, Bounded Features, Minimum Curvature (Hessian Positive Definite)

Results

- Upper-bounding the number of “close-by” cliques for a given clique

$$|C(c, \lambda)| < \frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1} d^{\lambda} \tag{4}$$

- Achieve a polynomially dependent sample-complexity bound as

$$\mathbb{P}\left[||\nabla \ell^{(n)}(\theta^*)||_2 > t\right] < \frac{r^2 \phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1} d^{\lambda^*} + 1\right] \tag{5}$$

Discussion

$$n > \frac{4r^2\phi_{max}^2}{\epsilon^2\delta C_{min}^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right] \tag{6}$$

- Importance of λ^*
- C_{min} makes the Hessian more friendly for convex optimization
- Larger clique size or larger graph size would lead to a requirement of more samples
- Term of 1/n that hints at a harder learning problem

Comparisons and Challenges

- The second approach relaxes the “decay” of dependence condition and only requires the dependence to be zero after a certain graph distance.
- The choice of loss function is only L2 here whereas the former results admits a more generic problem with added unknowns
- For very large graphs, the degree of the graph might make the bound vacuous and instead the former approach may be more suitable and relevant

Future Directions

- $O(\log 1/\delta)$ v/s $O(1/\delta)$ order of dependence on the confidence, reveals a contradictory nature of both the approaches
- β-stability seemed essential for the purpose of joint inference, seems absent here. Is β-stability important in its current form?
- Number of parameters to be learnt, size of cliques, degree of the nodes.

References i

Barber, D. (2012).
***Bayesian Reasoning and Machine Learning*.**
Cambridge University Press, New York, NY, USA.

He, P. and Zhang, C. (2014).
Non-Asymptotic Analysis of Relational Learning with One Network.
In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 320–327, Reykjavik, Iceland. PMLR.

London, B., Huang, B., Taskar, B., and Getoor, L. (2013).
Collective stability in structured prediction: Generalization from one example.
In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 828–836, Atlanta, Georgia, USA. PMLR.