Understanding Deep Learning

Many researchers in Computer Science, Machine Learning, Engineering and Statistics are trying to understand the spectacular performance of DNN in certain problem areas.

Unhelpful Theory

Understanding Deep Learning requires rethinking generalization
Zhang, Bengio, Hardt, Recht, Vinyals (2016-2017, Arxiv)

- They use deep CNNs trained on image data, which achieve zero training error, and near zero generalization error.
- They then randomize labels, or add noise to images in various degrees. Training error still zero, but generalization error varies from terrible to poor.
- This flies in the face of many *theories* based on VC dimension, Rademacher complexity and the like, which give bounds on generalization error in terms of training error and capacity.
- We complain about these bounds in ESL, and have never thought them very useful in practice.

Helpful Theory

Cover and Hart (1967), my wording: *The generalization error of NN classification is asymptotically no worse than twice the Bayes error.*
along with a very nice paper pointed to me by Jason Lee:
DNN or k-NN: That is the generalize vs Memorize question
Cohen, Sapiro and Giryes (preprint Arxiv, 2018)

- They demonstrate empirically and convincingly that a DNN derives an embedding (feature space) in which k-NN and the rest of the network get practically identical performance.
- So k-NN and various levels of Bayes error explain the training - test error "dilemma" in the first paper.

Unifying Theory

Many useful techniques are developed in application areas, and often more than one. For me theoretical analysis that connects them, explains them better, sheds light on their performance is always very gratifying.

"Presence Only" Data in Ecology

Species are found at geographical locations. Features are measured there. Features can be measured on geographical grid via satellite imagery (called "background data")

Two warring factions in ecology:

1. Naive logistic regression crowd: treat background data as zeros, and fit logistic regression.
2. "Maxent" crowd: fit tilted exponential density models for presence data, with tilters a function of features

1. easy to use, but thought to be "wrong"
2. felt to be useful, but hard to understand and not well motivated (but nice software, so used a lot)

Method 3. — the unifier

3. Inhomogeneous Poisson Processes. The Poisson rate $\lambda(x)$, depending on features, accounts for the emission of observed species. Seems very natural for the problem. (Warton and Shepherd 2010)

Guess what? Method 3. is equivalent to method 2., and for large number of background points, is equivalent to method 1. (Rennie and Warton 2013, Fithian and H, 2013)

Many more examples like this.