A note on the instability and degeneracy of deep learning models

Andee Kaplan

Iowa State University
ajkaplan@iastate.edu

June 22, 2017

Slides available at http://bit.ly/kaplan-private

Joint work with D. Nordman and S. Vardeman

Introduction

- A probability model exhibits *instability* if small changes in a data outcome result in large changes in probability
- Model *degeneracy* implies placing all probability on a small portion of the sample space

Goal: Quantify instability for general probability models defined on sequences of observations, where each sequence of length N has a finite number of possible outcomes

Notation

- $\boldsymbol{X} = (X_1, \dots, X_N)$ a set of discrete random variables with a finite sample space, $\mathcal{X}^N$
- For each N, $P_{\boldsymbol{\theta}_N}$ is a probability model on $\mathcal{X}^N$

FSFS models

Finitely Supported Finite Sequence (FSFS) model class

A series $P_{\boldsymbol{\theta}_N}$ of probability models, indexed by a generic sequence of parameters $\boldsymbol{\theta}_N$, to describe data of each length $N \geq 1$ with model support of $P_{\boldsymbol{\theta}_N}$ equaling the (finite) sample space $\mathcal{X}^N$.

- The size and structure of such parameters $\boldsymbol{\theta}_N$ are without restriction
- Natural cases include $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$ for some arbitrary integer-valued function $q(\cdot) \geq 1$

Discrete exponential family models

Exponential family model for $\boldsymbol{X}$ with pmf of the form

$$p_{N,\lambda}(\boldsymbol{x}) = \exp\left[\boldsymbol{\eta}^T(\boldsymbol{\lambda})\boldsymbol{g}_N(\boldsymbol{x}) - \psi(\boldsymbol{\lambda})\right], \quad \boldsymbol{x} \in \mathcal{X}^N,$$

for fixed positive dimensions of the parameter, $\boldsymbol{\lambda} \in \Lambda \subset \mathbb{R}^k$ and natural parameter $\boldsymbol{\eta} : \mathbb{R}^k \mapsto \mathbb{R}^L$ spaces, $\boldsymbol{g}_N : \mathcal{X}^N \mapsto \mathbb{R}^L$ a vector of sufficient statistics,

$$\psi(\boldsymbol{\lambda}) = \log \sum_{\boldsymbol{x} \in \mathcal{X}^N} \exp\left[\boldsymbol{\eta}^T(\boldsymbol{\lambda})\boldsymbol{g}_N(\boldsymbol{x})\right], \qquad \boldsymbol{\lambda} \in \Lambda,$$

the normalizing function, and $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^k : \psi(\boldsymbol{\lambda}) < \infty, k \leq q(N)\}$ is the parameter space.

Discrete exponential family models (cont'd)

- Such models arise with
  - Spatial data on a lattice (Besag 1974)
  - Network data (Wasserman and Faust 1994; Handcock 2003)
  - Binomial sampling with N iid Bernoulli random variables
- These models are special cases of the FSFS models
- $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \equiv p_{N,\boldsymbol{\lambda}_N}(\boldsymbol{x})$ with $\boldsymbol{\theta}_N = \boldsymbol{\lambda}_N$ a sequence of elements of $\Lambda \subset \mathbb{R}^k$
- $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}^N$
- The dimension of the parameter $\boldsymbol{\theta}_N$ is the same for each N (k)
- Schweinberger (2011) considered *instability* in such exponential models

Restricted Boltzmann machines

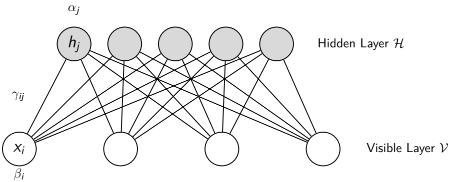

Figure 1: An example restricted Boltzmann machine (RBM). Hidden nodes are indicated by gray filled circles and the visible nodes indicated by unfilled circles.

- $\mathcal{X} = \{-1, 1\}$
- $\boldsymbol{X} = (X_1, \dots, X_N)$: N random variables for visibles with support $\mathcal{X}^N$
- $\boldsymbol{H} = (H_1, \dots, H_{N_H})$: N_H random variables for hiddens with support $\mathcal{X}^{N_H}$
- Parameters $\boldsymbol{\alpha} \in \mathbb{R}^{N_H}$, $\boldsymbol{\beta} \in \mathbb{R}^N$, Γ a matrix of size $N_H \times N$ ($\boldsymbol{\theta}_N = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Gamma) \in \Theta_N \subset \mathbb{R}^{q(N)}$ with $q(N) = N + N_H + N * N_H$)

Joint pmf:

$$P_{\boldsymbol{\theta}_N}(\tilde{\boldsymbol{x}}) = \exp\left[\boldsymbol{\alpha}^T \boldsymbol{h} + \boldsymbol{\beta}^T \boldsymbol{x} + \boldsymbol{h}^T \Gamma \boldsymbol{x} - \psi(\boldsymbol{\theta}_N)\right], \quad \tilde{\boldsymbol{x}} = (\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{X}^{N+N_H}$$

Restricted Boltzmann machines (cont'd)

- The pmf for the visible variables $X_1, \ldots, X_N$ follows from marginalization:

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) = \sum_{\boldsymbol{h} \in \mathcal{X}^{N_H}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}), \qquad \boldsymbol{x} \in \mathcal{X}^N.$$

- Size of $\boldsymbol{\theta}_N$, $q(N)$, increases as a function of sample dimension N
- Can choose the number N_H of hidden variables to change with N (potentially increase)
- The RBM model specification for visibles is a FSFS model
- Models formed by marginalizing a base FSFS model (e.g., a type of exponential family model) is again a FSFS model class

Deep learning

Instability results

Implications

Thank you

Questions?

- Slides – http://bit.ly/kaplan-private
- Contact
  - Email – ajkaplan@iastate.edu
  - Twitter – http://twitter.com/andeekaplan
  - GitHub – http://github.com/andeek

References I

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 192–236.

Handcock, Mark S. 2003. "Assessing Degeneracy in Statistical Models of Social Networks." Center for Statistics; the Social Sciences, University of Washington. `http://www.csss.washington.edu/`.

Schweinberger, Michael. 2011. "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families." *Journal of the American Statistical Association* 106 (496). Taylor & Francis: 1361–70.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge: Cambridge University Press.