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 a general and broad class of 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 $\boldsymbol{\theta}_N$, for describing data of length $N \geq 1$. The model support of $P_{\boldsymbol{\theta}_N}$ equals 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

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

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

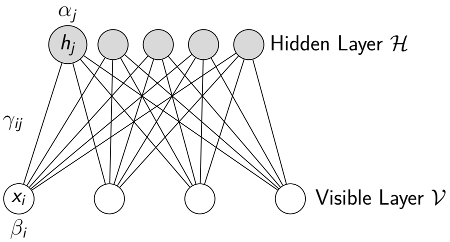
$$\psi(\boldsymbol{\lambda}) = \log \sum_{\boldsymbol{x} \in \mathcal{X}^N} \exp\left[\boldsymbol{\eta}^T(\boldsymbol{\lambda})\boldsymbol{g}_N(\boldsymbol{x})\right], \qquad \boldsymbol{\lambda} \in \Lambda,$$

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

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)
  - Standard independence models for discrete data (N iid Bernoulli 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 k of the parameter $\boldsymbol{\theta}_N$ is the same for each N
- Schweinberger (2011) considered *instability* in such exponential models

Restricted Boltzmann machines



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

Two models with "deep architecture" that contain multiple hidden layers in addition to a visible layer of data

1. Deep Boltzmann machine (DBM)
    - Stacked RBMs with conditional dependence between neighboring layers.
    - The probability mass function for $X_1, \dots, X_N$ follows from marginalization of the joint pmf
2. Deep belief network (DBN)
    - **Similar** to a DBM: Multiple layers of latent random variables stacked in a deep architecture with no conditional dependence within layers
    - **Difference**: all but the last stacked layer in a DBN are Bayesian networks (see Pearl 1985)

$q(N)$ is dependent on the dimension of the visibles
$\Rightarrow$ visible DBM and DBN model specifications are both FSFS models

S-instability

S-unstable FSFS models

Let $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$ be a sequence of FSFS model parameters where the size of the model $q(N)$ is a function of the number of random variables N. A FSFS model formulation is *Schweinberger-unstable* or *S-unstable* if, as the number of variables increase ($N \to \infty$),

$$\lim_{N\to\infty} \frac{1}{N}\text{ELPR}(\boldsymbol{\theta}_N) \equiv \lim_{N\to\infty} \frac{1}{N} \log \left[\frac{\max\limits_{(x_1,\dots,x_N)\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(x_1,\dots,x_N)}{\min\limits_{(x_1,\dots,x_N)\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(x_1,\dots,x_N)} \right] = \infty.$$

This generalizes "unstable" from Schweinberger (2011) by allowing

1. non-exponential family models and
2. an increasing number of parameters

Differs in form but matches Schweinberger (2011) for exponential models

Consequences of S-instability

Small changes in data can lead to overly-sensitive changes in probability. Let

$$\Delta(\boldsymbol{\theta}_N) \equiv \max\left\{\log \frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^*)} : \boldsymbol{x} \ \& \ \boldsymbol{x}^* \in \mathcal{X}^N \text{ differ in exactly 1 component}\right\},$$

Proposition 1

For an integer $N \geq 1$ and a given $C > 0$, if

$$\frac{1}{N}\text{ELPR}_N(\boldsymbol{\theta}_N) > C,$$

then

$$\Delta_N(\boldsymbol{\theta}_N) > C.$$

If the scaled ELPR is large, then the FSFS model can exhibit large changes in probability for small differences in the data configuration

Tie to degeneracy

Define a ϵ-modal set

$$M_{\epsilon,\boldsymbol{\theta}_N} \equiv \left\{ \boldsymbol{x} \in \mathcal{X}^N : \log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > (1-\epsilon) \max_{\boldsymbol{x}^* \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^*) + \epsilon \min_{\boldsymbol{x}^* \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^*) \right\}$$

of possible outcomes, for a given $0 < \epsilon < 1$.

Proposition 2

For an S-unstable FSFS model and for any given $0 < \epsilon < 1$,

$$P_{\boldsymbol{\theta}_N}\left((x_1, \dots, x_N) \in M_{\epsilon,\boldsymbol{\theta}_N}\right) \to 1 \text{ as } N \to \infty.$$

In S-unstable FSFS models, all probability in the model formulation with a large number of random variables will concentrate mass on an ϵ-mode set for any arbitrarily small ϵ (potentially small set of outcomes with most probability)

Implications

For a large class of models, including "deep learning" models, we have

1. Developed a formal definition of instability 2. Shown potential

consequences of instability (degeneracy)

Models that fall within the definition of a FSFS model should be used with **caution** to ensure that the effects of instability are not experienced

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/`.

Pearl, Judea. 1985. "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning." UCLA Computer Science Department.

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.