Bayesian Neural Networks

Yuhan Xue

March 9, 2010
Motivation

- Neural networks have been developed mainly by the machine learning community
- Statisticians often consider them to be black boxes not based on a probability model
- **How neural networks can be applied into nonparametrics regression and classification modeling**

My talk is based on *Bayesian Nonparametrics via Neural Networks* by Professor Herbie Lee.
Outline

Bayesian Non Parametrics Model for Neural Networks
- BNP Regression using Local Methods
- BNP Regression using Basis Functions
- Neural Networks: Basic Model

Nonparamatric Multivariate regression

Nonparamatric Classification
- Example: Fisher’s Iris Data

Modeling Issues
- Choosing Priors
- Bayesian Model Selection

Conclusions
BNP Regression using Local Methods

- Spline
- Generalized Additive Model (GAM)

\[ y_i = \sum_{j=1}^{r} f_j(x_{ij}) + \epsilon_i \]

local smoothed function for each variable without interaction effects

\[ y = f_1(x_1) + f_2(x_2) + ... + \epsilon \] where \( f_j \) is a one-dim spline.

- Projection Pursuit Regression (PPR)

\[ y_i = \sum_{j=1}^{r} f_j(\beta^t x_i) + \epsilon_i \]

rotation of the axes
BNP Regression using Basis Functions

- **General form**

\[ y_i = \sum_{j=0}^{k} \beta_j f_j(x_i) + \epsilon_i \]

- **Polynomial basis functions**

\[ f_j(x) = x^j \quad \text{i.e.} \quad \{1, x, x^2, x^3, x^4, \ldots\} \]

- **Logistic basis functions**

\[ y_i = \sum_{j=0}^{k} \beta_j \psi(x_i) + \epsilon_i, \text{ where } \psi(x) = \frac{1}{1 + \exp\{-x\}} \]

- **Gaussian densities**

\[ y_i = \sum_{j=0}^{k} \beta_j \phi\left(\frac{x_i - \mu_j}{\sigma_j}\right) + \epsilon_i, \text{ where } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \]
Neural Networks: Basic Model

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \psi(\gamma_j x_i) + \epsilon_i \]

- Recall Projection Pursuit Regression (PPR):
  \[ y_i = \sum_{j=1}^{r} f_j(\beta^t x_i) + \epsilon_i \]

- Recall logistic basis functions:
  \[ y_i = \sum_{j=0}^{k} \beta_j \psi(x_i) + \epsilon_i \]

- The basic neural network model = PPR + logistic basis functions
Neural Networks: Basic Model

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \frac{1}{1 + \exp \left\{ -\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih} \right\}} + \epsilon_i \]
Example:
- Single input
- Single output
- 2 Hidden Nodes
Neural Networks: Basic Model (Continued)

The fitted curve:

\[ y = 4 - \frac{10.58}{1.0 + \exp(21.75 - 0.19 \times x)} + \frac{13.12}{1.0 + \exp(19.60 - 0.07 \times x)} \]
Neural Networks: Basic Model (Continued)

\[ y = 4 - \frac{10.58}{1.0 + \exp(21.75 - 0.19 \times x)} + \frac{13.12}{1.0 + \exp(19.60 - 0.07 \times x)} \]

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \psi(\gamma_j^t x_i) + \epsilon_i \]

- \( \beta_0 \) Overall location parameter for \( y \)
- \( \beta_j \) Overall scale factor for \( y \)
- \( \gamma_0 \) Center of the logistic function occurs at \(-\frac{\gamma_0}{\gamma_1}\)
- \( \gamma_j \) For \( \gamma_1 \) larger, \( y \) changes at a steeper rate at the neighborhood of the center.
Outline

Bayesian Non Parametrics Model for Neural Networks
  BNP Regression using Local Methods
  BNP Regression using Basis Functions
  Neural Networks: Basic Model

Nonparamatric Multivariate regression

Nonparamatric Classification
  Example: Fisher’s Iris Data

Modeling Issues
  Choosing Priors
  Bayesian Model Selection

Conclusions
Multivariate regression: Basic Setup

- From

\[ y_{ig} = \beta_0 g + \sum_{j=1}^{k} \beta_{jg} \psi_j(\gamma_j^t x_i) + \epsilon_{ig} \]

And

\[ \psi_j(\gamma_j^t x_i) = \frac{1}{1 + \exp \left\{ -\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih} \right\}} \]

The multivariate regression model in Neural Networks:

\[ y_{ig} = \beta_0 g + \sum_{j=1}^{k} \beta_{jg} \frac{1}{1 + \exp \left\{ -\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih} \right\}} + \epsilon_{ig} \]

\[ \epsilon_{ig} \sim N(0, \sigma^2) \]

- Each dimension of the output is modeled as a **different** linear combination of the **same** basis functions.
From

\[ y_{ig} = \beta_{0g} + \sum_{j=1}^{k} \beta_{jg} \frac{1}{1 + \exp\{-\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jhx_{ih}}\}} + \epsilon_{ig} \]
Outline

Bayesian Non Parametrics Model for Neural Networks
  BNP Regression using Local Methods
  BNP Regression using Basis Functions
  Neural Networks: Basic Model

Nonparamatric Multivariate regression

Nonparamatric Classification
  Example: Fisher’s Iris Data

Modeling Issues
  Choosing Priors
  Bayesian Model Selection

Conclusions
Classification: Basic setup

An extension of multinomial likelihood approach

- $q$ categories $\{1, \ldots, q\}$

$$y_{ig} = \begin{cases} 
1, & \text{if } g = y_i \\
0, & \text{otherwise} 
\end{cases}$$

- Likelihood

$$f(y|p) = \prod_{i=1}^{n} f(y_i|p_{i1}, \cdots, p_{iq})$$

$$\propto \prod_{i=1}^{n} (p_{i1})^{y_{i1}} \cdots (p_{iq})^{y_{iq}}$$

For example, if we have three categories, then $q=3$

$$y_{11} = 1 \quad y_{12} = 0 \quad y_{13} = 0$$
$$y_{21} = 0 \quad y_{22} = 0 \quad y_{23} = 1$$
$$\vdots$$
$$y_{i1} = 0 \quad y_{i2} = 1 \quad y_{i3} = 0$$
$$\vdots$$
Classification: Basic setup (Continued)

\[ p_{ig} = \frac{\exp \{ w_{ig} \}}{\sum_{g=1}^{q} \exp \{ w_{ig} \}} \]

\[ w_{ig} = \beta_0g + \sum_{j=1}^{k} \beta_{jg} \psi_j(\gamma_j^t x_i) \]

\[ \psi_j(\gamma_j^t x_i) = \frac{1}{1 + \exp \{ -\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih} \}} \]

\( i = 1, \ldots, n \) Sample size

\( h = 1, \ldots, r \) Number of imput variables

\( j = 1, \ldots, k \) Number of hidden nodes

\( g = 1, \ldots, q \) Number of output variables, i.e. categories

Classification Rule: \( \hat{g}_i = \arg \max w_{ig} \)
Classification: Basic setup (Continued)

- $p_{ig} = \frac{\exp\{w_{ig}\}}{\sum_{g=1}^{q} \exp\{w_{ig}\}}$

- $w_{ig} = \beta_{0g} + \sum_{j=1}^{k} \beta_{jg} \psi_j(\gamma_j^t x_i)$

- $\psi_j(\gamma_j^t x_i) = \frac{1}{1 + \exp\{-\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih}\}}$
Classification: Fisher’s Iris Data

For simplicity, in this example I am using

- Classification between 2 types
  - Setosa vs. Versicolor
  - Versicolor vs. Virginica

- 2 input variables
  - Sepal Width
  - Petal Width

- Number of hidden nodes
  - 2 nodes
  - 4 nodes

- Maximum Likelihood
Classification: Fisher’s Iris Data
Setosa vs. Versicolor

- Setosa vs. Versicolor
- 2 hidden nodes
- 4 hidden nodes
Classification: Fisher’s Iris Data
Versicolor vs. Virginica

- Versicolor vs. Virginica
- 2 hidden nodes
- 4 hidden nodes
Outline

Bayesian Non Parametrics Model for Neural Networks
  BNP Regression using Local Methods
  BNP Regression using Basis Functions
  Neural Networks: Basic Model

Nonparamatric Multivariate regression

Nonparamatric Classification
  Example: Fisher’s Iris Data

Modeling Issues
  Choosing Priors
  Bayesian Model Selection

Conclusions
Choosing Priors: Parameter Lack Interpretability

Parameter may lack interpretability, so it is hard to determine a reasonable prior belief.
Example:
Choosing Priors: Hierarchical Priors

\[
\begin{align*}
y_i &\sim N \left( \sum_{j=0}^{k} \beta_j \psi(\gamma_j^t x_i, \sigma^2) \right) \\
\beta_j &\sim N(\mu_\beta, \sigma^2_\beta) \\
\gamma_j &\sim N_p(\mu_\gamma, \Sigma_\gamma) \\
\sigma^2 &\sim \Gamma^{-1}(s, S) \\
\mu_\beta &\sim N(a_\beta, A_\beta) \\
\mu_\gamma &\sim N(a_\gamma, A_\gamma) \\
\sigma^2_\beta &\sim \Gamma^{-1}(c_\beta, C_\beta) \\
\Sigma_\gamma &\sim \text{Wish}^{-1}(c_\gamma, (c_\gamma C_\gamma)^{-1})
\end{align*}
\]
Choosing Priors: Comparison between Priors

- Problems with Improper Priors
- Priors in Consideration of Overfitting
  - Weight decay
  - Shrinkage Priors
- Example comparing Priors
Bayesian Model Selection

Recall the Basic Model

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \psi(\gamma_j^t x_i) + \epsilon_i \]

Bayesian model selection

- The covariates to be included. i.e. the optimal \( x \)
- The number of hidden nodes. i.e. the optimal \( k \)
- Modeling vs Prediction
- Model Averaging
Bayesian Model Selection Criteria

**Bayes Factors**

\[
\frac{P(M_1|y)}{P(M_2|y)} = \frac{P(M_1)}{P(M_2)} \frac{P(y|M_1)}{P(y|M_2)}
\]

**BIC**

\[
BIC_i = \ln f(y|\hat{\theta}, M_i) - \frac{1}{2} d_i \ln n
\]

**Log Scores**

\[
LS_{CV} = \frac{1}{n} \sum_{j=1}^{n} \ln p(y_j|y_{-j}, M_i)
\]

\[
LS_{FS} = \frac{1}{n} \sum_{j=1}^{n} \ln p(y_j|y, M_i)
\]

**DIC**

\[
DIC(M_i|y) = D(\hat{\theta}) + 2\hat{p}_{Di}
\]
Outline

Bayesian Non Parametrics Model for Neural Networks
   BNP Regression using Local Methods
   BNP Regression using Basis Functions
   Neural Networks: Basic Model

Nonparamatric Multivariate regression

Nonparamatric Classification
   Example: Fisher’s Iris Data

Modeling Issues
   Choosing Priors
   Bayesian Model Selection

Conclusions
Conclusions

Advantages

▷ Flexibility
▷ High-dimensional
▷ Good track record

Disadvantages

▷ Complexity
▷ Lack of interpretability
▷ Difficult to specify prior information