A Case Study with nn4pse: An R Package for Propensity Score Estimation with Neural Networks

Bryan Keller & Zhuangzhuang Han
Teachers College, Columbia University
Measurement, Evaluation, & Statistics

BACKGROUND & CONTEXT

Introduction
While logistic regression remains the most frequently used method for estimating propensity scores in practice, there has been a burgeoning interest in data mining alternatives that can handle interactions and nonlinear relationships in their naive implementations. Neural networks is one such method which has performed well in simulation studies (Setoguchi et al., 2008; Keller et al., 2013).

Purpose
• to examine the use of neural networks for propensity score estimation in the context of a real-data example as compared with main-effects only logistic regression and a customized logistic regression model that includes several two-way interactions;
• to demonstrate the functionality of the nn4pse R package (Keller & Han, in preparation), which implements neural networks for propensity score estimation.

The Data
- The 34 covariates we include in our study were selected by Morgan et al. (2010) based on theory and previous results in the literature that linked them to special education placement.
- A child was defined as a recipient of special education services based on his or her special education status gathered from school administrative records from the spring of 2002.
- The outcomes of interest are the IRT scaled math and reading achievement test scores.

STATISTICAL MODEL/METHODOLOGY

Neural Networks
- We estimate propensity scores using
  a) main-effects only logistic regression,
  b) logistic regression with four two-way interaction terms, c) neural networks.
- Single-layer feed-forward neural networks consist of an input layer of observed variables, a hidden layer of M hidden units, and an output layer that contains a single unit for binary classification in propensity score estimation.
- Weight decay is a technique that imposes penalties on large weights, as in ridge regression, thereby preventing overfitting.
- Two tuning parameters: the number of hidden units (M) and the weight decay value λ.

nn4pse
- Automatic pre-processing of data for neural networks.
- Options to select the tuning parameters based on cross-validated prediction accuracy or average balance on observed covariates and their two-way interactions.
- Options for removing non-overlapping data based on a caliper.
- Choice of estimating the ATE or the ATT by PS weighting.

Definitions
- The standardized mean difference for covariate X (Cohen’s d): 
  \[ d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_p} \]
- The variance ratio for covariate X: 
  \[ \frac{\sigma_2^2}{\sigma_1^2} \]
- The average standardized absolute mean difference (ASAMD): 
  \[ ASAMD = \frac{1}{p} \sum_{i=1}^{p} |d_i| \]
- The average absolute variance ratio deviation from 1 (AAVRI1): 
  \[ AAVRI1 = \frac{1}{p} \sum_{i=1}^{p} |v_i - 1| \]

RESULTS

Balance Training Plot (Output from nn4pse)

Balance Training Plot (Output from nn4pse)

Fig. 1. Covariate balance averaged over 5 neural network runs per weight decay value. The weight decay value associated with the best balance was λ = .52.

Summary of Balance Measures
- As measured by the ASAMD and AAVR1, propensity scores estimated by neural networks achieved uniformly better balance than logistic regression with only main-effects terms.
- The second logistic model, which included 4 two-way interaction terms selected by Morgan et al. (2010) improved balance on both first- and second-order terms.
- The best balance on second-order terms was achieved by neural networks, which is consistent with our expectation.

CONCLUSION & RECOMMENDATIONS

• If an analyst must choose an automated procedure for propensity score estimation, we recommend neural networks over main-effects-only logistic regression.
• The customized logistic model that included two-way interactions achieved better balance than neural networks on first-order terms, but not on second-order terms.
• An advantage of using neural networks for propensity score estimation is that they automatically handle interactions and nonlinear features.
• A topic for future research is how the number of covariates affects the performance of neural networks for propensity score estimation.

REFERENCE


ASAMDS and AAVRIs Between Special Education Groups Before and After Propensity Score Weighting

<table>
<thead>
<tr>
<th></th>
<th>Unadj.</th>
<th>Logistic: Main Effects</th>
<th>Logistic: Interactions</th>
<th>Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASAMD</td>
<td>AAVR1</td>
<td>ASAMD</td>
<td>AAVR1</td>
</tr>
<tr>
<td>First-Order</td>
<td>0.328</td>
<td>0.385</td>
<td>0.022</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>0.008</td>
<td>0.018</td>
<td>0.092</td>
</tr>
<tr>
<td>Second-Order</td>
<td>0.282</td>
<td>0.476</td>
<td>0.022</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>0.021</td>
<td>0.093</td>
<td>0.019</td>
<td>0.101</td>
</tr>
</tbody>
</table>