CS289 Initial Report: Methods of handling missing data for classification

Rafael Valle[†] Jason Poulos[‡]

October 30, 2015

1 Motivation

Methods of handling missing data for neural networks classification model

Given that we plan to use NNets for classification on the Adult dataset, we must handle missing data. This is not necessary for other ML models such as random forest, decision trees, etc.

2 Data

2.1 Adult data set

2.2 Benchmarks

2.3 Exploratory data analysis

3 Methods

3.1 Techniques for handling missing data

1. Basic Statistics : Replace the missing data with the mean or median of the feature vector.

[†] `rafaelvalle@berkeley.com`.

[‡] `poulos@berkeley.edu`.

2. One-hot : Create an indicator variable to indicate whether or not the feature has missing data.

3. . Nearest Neighbor Imputation : Recursively compute the K-Nearest Neighbors of the observation with missing data and assign the mean or median of the K-neighbors to the missing data.

4. . Regression : Recursively train a regression model to predict the feature with missing data.

5. Factor analysis : Perform eigen decomposition on the design matrix, project the design matrix onto the first two eigen vectors and replace the missing values by the values that might be given by the projected design matrix.

6. Find other features with distribution similar to the feature containing missing data and use this information (e.g. correlation) to fill in in the missing data. But then, if two features are highly correlated, it might be better to remove one of them.

3.2 Neural networks for classification

4 Anticipated results

Appendix