# CS289 Initial Report: Methods for handling missing and categorical data for classification with neural networks

Jason Poulos[†] Rafael Valle[‡]

November 4, 2015

## 1 Motivation

In our project, we plan to investigate techniques for handling missing data and encoding categorical data such that it is appropriate to neural network classifiers.

Missing data is a common problem in survey data in various domains. Several techniques for data imputation (replace missing values with plausible ones) and direct estimation (all missing data is analyzed using a maximum likelihood approach) have been developed.[De Leeuw et al., 2003]

In the case of categorical variables, which by definition have no direct representation or computation scheme of the distance between its values, decision trees can be useful because they do not require distance metrics. However, their training process is slow given a large enough dataset and they might not be suitable for problems where the decision boundary between classes described by a second-order polynomial[1], for example.[Fayyad et al., 1996]

[†]poulos@berkeley.edu.
[‡]rafaelvalle@berkeley.com.
[1]We note, however, that a property test can be as complex as the data at hand.

# 2 Data

## 2.1 Benchmarks

We plan to experiment with the Adult dataset from the UCI Machine Learning Repository [Lichman, 2013]. The dataset has 48,842 samples (train = $32,561$ and test = $1,6281$). The dataset contains 14 features: 6 continuous and 8 categorical. The prediction task is to determine whether a person makes over $50,000 a year; 24% of individuals in the training data make more than this amount.

Table 1 shows the test error rates obtained by the data set donor [Kohavi, 1996]. All error rates were obtained after removing samples with missing values. The error rate to beat is 14.05%.

Given the results of our experiments and if time permits, we may move to a much larger dataset, such as the 1940 full–count U.S. Census file [NAPP, 2008, Ruggles et al., 2010]. The 1940 Census has about 100 million samples and 100 features.

| **Algorithm** | **Error** |
|---|---|
| 1 C4.5 | 15.54 |
| 2 C4.5-auto | 14.46 |
| 3 C4.5 rules | 14.94 |
| 4 Voted ID3 (0.6) | 15.64 |
| 5 Voted ID3 (0.8) | 16.47 |
| 6 T2 | 16.84 |
| 7 1R | 19.54 |
| 8 NBTree | 14.10 |
| 9 CN2 | 16.00 |
| 10 HOODG | 14.82 |
| 11 FSS Naive Bayes | 14.05 |
| 12 IDTM (Decision table) | 14.46 |
| 13 Naive-Bayes | 16.12 |
| 14 Nearest-neighbor (1) | 21.42 |

Table 1: Test set error rates on Adult dataset for various algorithms, obtained after removal of samples with missing values and using the original train/test split. Source: Lichman [2013].

## 2.2 Patterns of missing values

The Adult dataset has 3,620 (7.4%) samples containing missing values. Missing values occur in three of the categorical features: *Work class*, *Occupation*, and *Native country*. It is unlikely that these values are missing completely at random (MCAR); it is more likely, and much less desirable that the values are not missing at random (MNAR). Since these data originate from a survey, the missing values may be due to respondents unwilling or unable

to provide an answer.

Uncovering patterns of missing values in the dataset will help select strategies for imputing missing values. The histogram (left) in Figure 1 shows *Work class* and *Occupation* each have about 5.6% of missing values, and *Native country* has about 1.7% missing values. The aggregation plot (right) shows 5.5% of samples are missing values for both *Work class* and *Occupation*. Less than 2% of samples are missing just *Native country* and less than 1% are missing all three features.

Figure 2 shows the frequency of observed categories and missing values for *Work class* and *Occupation*. Each stacked column shows the proportion of missing values in the other feature and *Native country* for each category. The plot shows the missing values are not MCAR: for instance, individuals working in the private sector are more likely to have missing values than those individuals in other work classes. However, missing values tend to be evenly distributed across occupational categories.

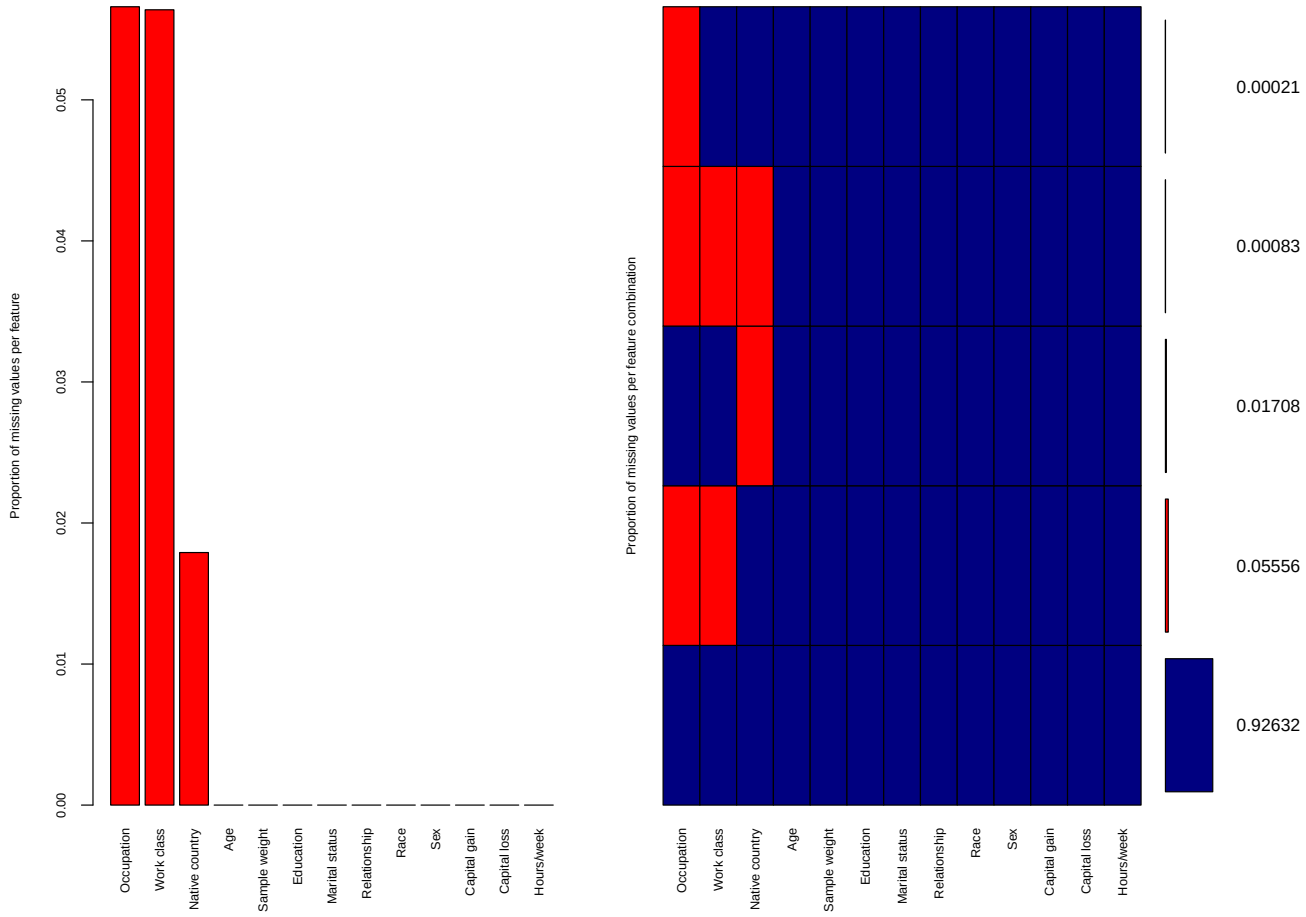


Figure 1: Histogram of proportion of missing values in each feature (Left) of Adult training set and aggregation plot of all existing combinations of missing and non-missing values in the samples (Right).

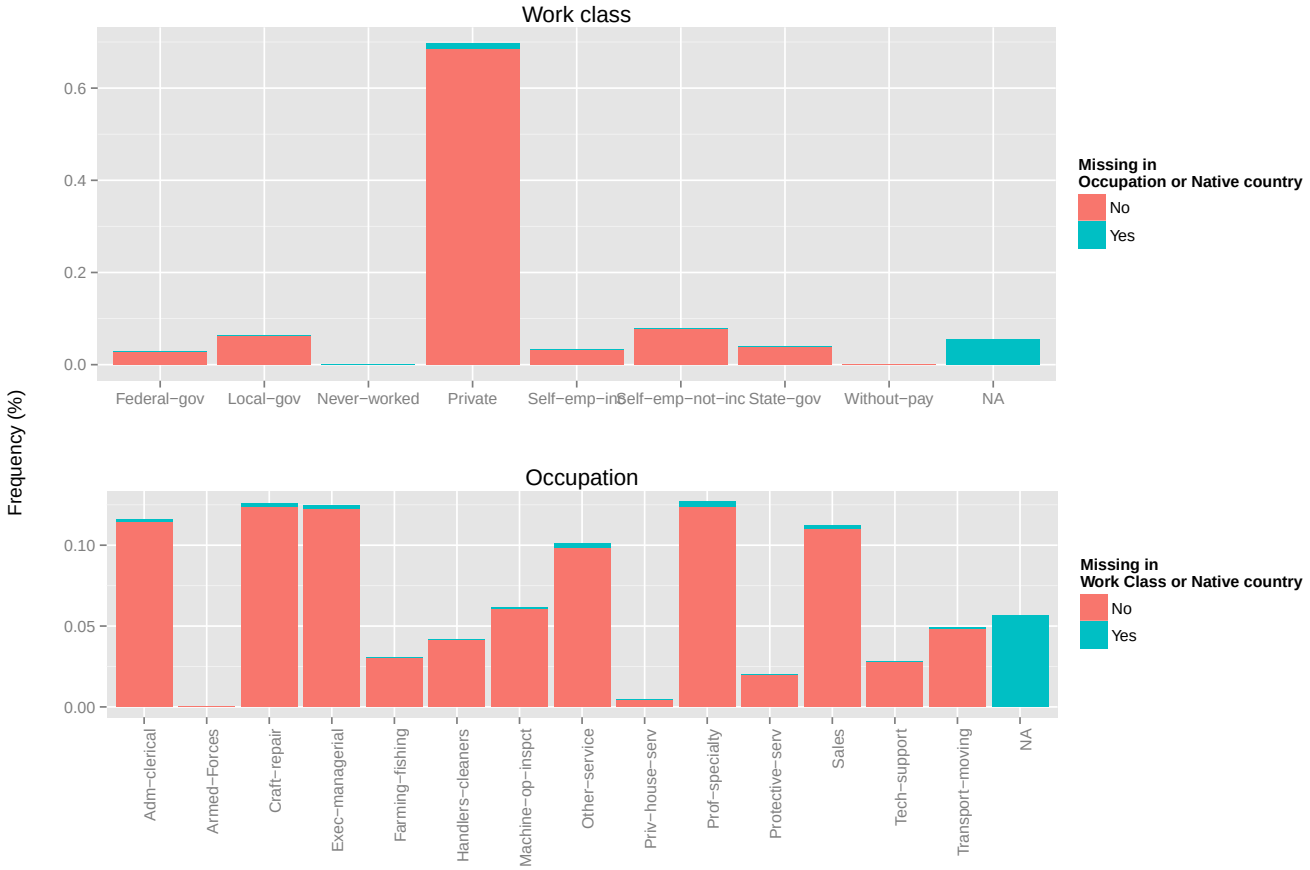


Figure 2: Barplot of proportion of observed and missing values of *Work class* and *Occupation* in Adult dataset.

# 3 Methods

## 3.1 Techniques for handling missing data

We divide imputation methods into six groups listed below[Batista and Monard, 2003]. Assuming that these techniques are easy to implement, we would like to compare their efficiency in imputing the missing values.

**Case substitution** One observation with missing data is replaced with another non-sampled observation.

**Mean or mode imputation** Replace the missing data with the mean or median of the feature vector. Since the missing variables in the ADULT dataset are all categorical, using a numerical approach directly is not appropriated.

**One-hot** Create a binary variable to indicate whether or not a specific feature is missing. This technique was suggested by Isabelle Guyon.

**Hot deck and cold deck** 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. A similar technique is used in Airbnb's fraud detection algorithm.

**Prediction Model** Train a prediction model, e.g. logistic regression or bagging, with all features to predict the missing missing value. This requires correlation amongst features to exist.

**Factor analysis** Perform some sort of factorization 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.

## 3.2 Neural networks for classification with categorical and quantitative features

Common techniques for handling categorical data in neuronal networks include encoding the categorical values into numeric values or using binary encoding. These techniques, however, have some drawbacks including unnecessarily increasing model complexity or feature dimensionality and not preserving the similarity information embedded betweeng categorical values[Hsu, 2006].

More elaborate techniques include information theoretic measures [Wang et al., 2008], training separate output units for each of the allowed combination of values of the categorical independent variables[Brouwer, 2002], and using distance hierarchies[Hsu, 2006].

# References


Gustavo EAPA Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

Roelof K Brouwer. A feed-forward network for input that is both categorical and quantitative. *Neural Networks*, 15(7):881–890, 2002.

Edith D De Leeuw, Joop Hox, Mark Huisman, et al. Prevention and treatment of item nonresponse. *Journal of Official Statistics-Stockholm*, 19(2):153–176, 2003.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

Chung-Chian Hsu. Generalizing self-organizing map for categorical data. *Neural Networks, IEEE Transactions on*, 17(2):294–304, 2006.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207. Citeseer, 1996.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

NAPP. Minnesota population center. north atlantic population project: Complete count microdata. version 2.0 [machine-readable database]. *Minneapolis, MN: Minnesota Population Center, available at* ***https://www.nappdata.org***, 2008.

S. Ruggles, T. Alexander, K. Genadek, R. Goeken, M. Schroeder, and M. Sobek. Integrated public use microdata series (ipums): Version 5.0 [machine-readable database]. *University of Minnesota, Minneapolis, available at* ***http://usa.ipums.org***, 2010.

Huanjing Wang, Guangming Xing, and Kairui Chen. Categorical data transformation methods for neural networks. In *IKE*, pages 262–266, 2008.