Missing Data in Cardiovascular Disease Prediction With Neural Networks

Megan Shand, Samuel Friedman, Kaan Yuksel, Puneet Batra

Motivation

UKBiobank has provided a rich multimodal dataset including imaging data (MRIs), time-series data (EKGs), genetic data, survey responses, and physical measurements for 500,000 participants. Our goal is to predict various cardiovascular diseases (CVD) from these datasets. The need to integrate these diverse types of data motivates the use of neural networks, which can be trained jointly across all data modalities and disease types, such as atrial fibrillation, coronary artery disease (CAD) and myocardial infarction.

Challenges

Missing data is a driving component of the UKBiobank surveys and physical measurements. Missingness is complex; survey answers can be absent, refused, unknown, censored, or truncated, and devices can emit errors, or impossible values. Out of thousands of variables, even features found to have above average importance for CVD prediction can have up to 95% missing values. This occurs for variables such as 'Age diabetes diagnosed', which demonstrates the need for careful consideration of the missingness. Simplistic solutions such as listwise deletion would both remove valuable signal and make training infeasible due to a lack of data.
Neural Networks cannot handle ragged tensors naively, so each variable with missing values must be dealt with explicitly.

Methods

- Hyperparameter Optimization: Bayesian Hyperparameter Optimization was used to tune architecture (number of dense layers and number of nodes) and number of features.
- Feature Selection: A Random Forest model was trained on all inputs and indicator variables of their missingness to predict CAD incidence. Missing values in the features were input as zeros. The 80 most important features from this model were used for training the Neural Networks.

BROAD INSTITUTE

Handling Missingness

- Model Missingness Explicitly: Each input value can have a separate channel that indicates if a value is missing or not.
- Impute Missing Values: Missing values can be imputed as the average value for that variable, or drawn from a standard normal distribution (since all inputs are normalized).
- Combine Discretely Missing Values: Variables that can be combined give a significant boost to performance. For example, 'Mother's age' and 'Mother's age at death' are separate variables that are mutually exclusive. Combining them with additional channels for 'Mother is alive', 'Mother is dead', and 'Otherwise missing' increases Area Under the ROC Curve (AUROC) by 6%.
- Hierarchical Survey Questions: Some missing data is due to the hierarchical structure of the survey. For example 'Number of minutes exercise per day' is only asked of participants who had non zero 'Number of days per week exercise'. Architectures that included this hierarchy (Figure 1) had only .4% higher AUROC than flat architectures for predicting CAD incidence.

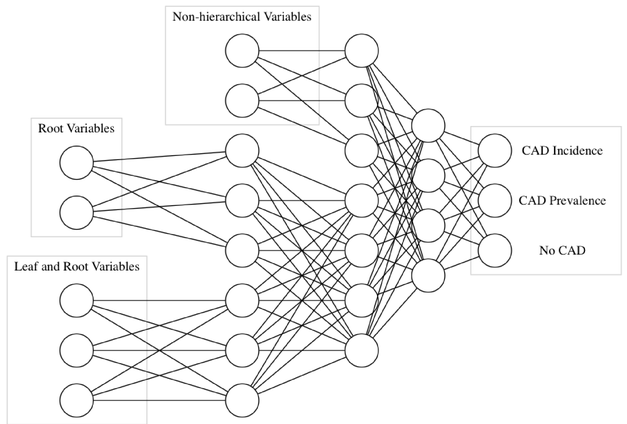

Figure 1: Example hierarchical architecture

Results

We compared models with the same architecture, trained on the same amount of data, and tested on the same held out test set. CAD incidence is the hardest and most clinically important prediction we make (see Table 1). Each model here does not use the hierarchical architecture, but does combine discretely missing values.

Missingness Channel	Imputation Method	AUROC
Yes	Normal(0,1)	.709
No	Normal(0,1)	**.713**
No	Mean	.696

Table 1: AUROC for CAD Incidence Prediction

The model without a missingness channel that imputed the missing values from a standard normal distribution outperformed those using the mean as the imputation method or modeling the missingness explicitly.

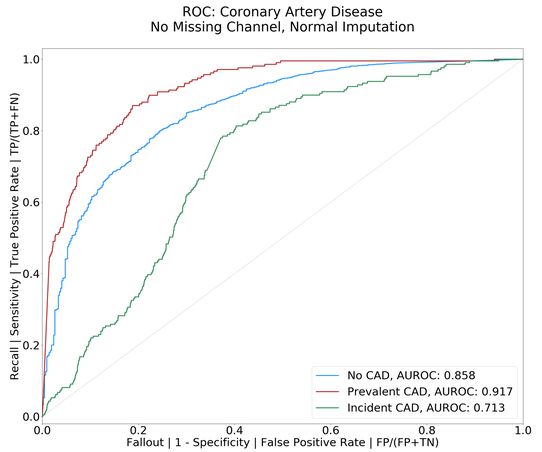

Figure 2: ROC Curves for predicting CAD with best model from Table 1

Acknowledgements

- Data used in this poster was generated by UKBiobank, for more information please visit: https://www.ukbiobank.ac.uk/