Introduction, Notation, and Problem Statement

Description

This document will outline the goals of the work to follow, including descriptions of the appraoches we will employ to demonstrate the superior accuracy of Probabilistic Scoring in the classification of PHQ-9 administrations.

Problem Statement

The main goals of this analysis are:

- demonstrate that Probabilistic Scoring is superior in accuracy to conventional linear scoring
- show that the accuracy of Probabilistic Scoring is a function of sample size, and that larger sample sizes (larger training data sets) correspond to better classification accuracy
- Show that Probabilistic Scoring converges to the best estimate of "Baseline Truth" in larger samples (bigger training sets).

Implementation

The analysis and investigation of Probabilistic Scoring will begin with a thorough investigation of mehtods:

- Attempt to replicate stated results so that the defined method is completely understood.
- Use cross-validation to partition data into various sizes of testing and training sets and implement replicated method across CV data.

We can investigate the convergence point of the accuracy of each method as it pertains to "Baseline Truth" by examining the behavior of the prediction accuracy as the sample size increases.

Notation

We will use the following generalized notation in order to clarify our communications.

- Indices
  - $h =$ Classification Group Index $\quad h = 1, \ldots, H$
    - The classification group is the result of the PHQ-9. It quantifies the extent to which a test-taker is at risk of depression
  - $i =$ Subject/Observation Number $\quad i = 1, \ldots, I$
    - The subject/observation number will be used as a unique mapping to each row of the data
  - $j =$ Question Number $\quad j = 1, 2, \ldots, 9$
    - The question number will be used for mapping response data back to the question of origin
- Random Variables
  - $K_{ij} =$ Response of Subect i on question j $\quad K_{ij} \in \{0, 1, 2, 3\}$
  - $M_i =$ Sum of question responses for subject i $\quad M_i \in \{0, 1, \ldots, 27\}$
    - $M_i = \sum_{j=1}^{9} K_{ij}$
  - $C_h = \{i | \gamma_h \leq M_i < \gamma_{h+1}\} =$ Level-set Classification Outcome Spaces
    - $\gamma_h \in \{\gamma_1, \gamma_2, \ldots, \gamma_H, \gamma_{H+1}\}$ are real numbers defining a partition of $\{0, 1, \ldots, 27\}$
- We say that the probability of subject i being classified into Classification group C_h after answering K_{ij} to question j is given by:

$$P\left(C_h = c_h \mid K_{ij} = k_{ij}\right) = \frac{P\left(K_{ij} = k_{ij} \mid C_h = c_h\right)\ P\left(C_h = c_h\right)}{P\left(K_{ij} = k_{ij}\right)}$$

Note that:

$$P\left(C_h = c_h\right) = \frac{\text{no. of people in Ch group}}{\text{Total no. of people}}$$

$$P\left(K_{ij} = k_{ij}\right) = \frac{\text{no. of people answer kij to quest. j}}{\text{Total no. question responses to quest. j}}$$

$$P\left(K_{ij} = k_{ij} | C_h = c_h\right) = \frac{\text{no. of people in Ch group w/kij}}{\text{no. of people in Ch group}}$$

Code Appendix

Script Dependencies

Package Dependencies

Working Directory

Load Data

Begin Script

[1]

References

[1] Centers for Disease Control, Prevention, et al., *New cdc report: More than 100 million americans have diabetes or prediabetes. july 18, 2017*, 2017.

End Script