

# Introduction, Notation, and Problem Statement

---

## Description

This document will outline the goals of the work to follow, including descriptions of the approaches 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 methods:

- 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

21

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

22

- Indices

23

- $h$  = Classification Group Index  $h = 1, \dots, H$

24

- \* The classification group is the result of the PHQ-9. It quantifies the extent to which a test-taker is at risk of depression

26

- $i$  = Subject/Observation Number  $i = 1, \dots, I$

27

- \* The subject/observation number will be used as a unique mapping to each row of the data

28

- $j$  = Question Number  $j = 1, 2, \dots, 9$

29

- \* The question number will be used for mapping response data back to the question of origin

30

- Random Variables

31

- $K_{ij}$  = Response of Subect i on question j  $K_{ij} \in \{0, 1, 2, 3\}$

32

- $M_i$  = Sum of question responses for subject i  $M_i \in \{0, 1, \dots, 27\}$

33

- \* NOTE:  $M_i = \sum_{j=1}^9 K_{ij}$

34

- $C_h = \{i | \gamma_h \leq M_i < \gamma_{h+1}\}$  = Level-set Classification Outcome Spaces

35

- \*  $\gamma_h \in \{\gamma_1, \gamma_2, \dots, \gamma_H, \gamma_{H+1}\}$  are real numbers defining a partition of the integer outcomes in the set  $\{0, 1, \dots, 27\}$

37

- 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:

39

$$P(C_h = c_h | K_{ij} = k_{ij}) = \frac{P(K_{ij} = k_{ij} | C_h = c_h) P(C_h = c_h)}{P(K_{ij} = k_{ij})}$$

Note that:

40

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

$$P(K_{ij} = k_{ij}) = \frac{\text{no. of people answer kij to quest. j}}{I}$$

41

$$P(K_{ij} = k_{ij} | C_h = c_h) = \frac{k_{ij} \cap c_h}{c_h} = \frac{\text{no. of people answer kij to quest. j also in group Ch}}{\text{no. of people in Ch group}}$$

## Cross-Validation Data Sets

42

The goal of the cross-validation (cv) process is to understand the interaction between probabilistic scoring accuracy and training data sample size. We have been given a complete data set of  $N=2495$  observations of PHQ9 questionnaires. We will partition the full data into test and training data, calculate expected classifications on test, and perform an accuracy analysis of these classifications. We will look to compare accuracy measures as training data size increases. We define the following quantities from a provided integer-pair relation  $N.obs \sim N.set$  which quantifies the number of sets ( $N.set$ ) into which the full data set will be partitioned, each containing the same number of observations. The integers will be related according to:

43

44

45

46

47

48

49

$$N.set = \left\lfloor \frac{N}{N.obs} \right\rfloor$$

and

50

$$N.obs = \left\lfloor \frac{N}{N.set} \right\rfloor$$

$$N = 2495 \implies N.set \in \{2, 3, \dots, 2495\}$$

and

51

$$N.obs \in \left\{ \left\lfloor \frac{N}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor + 1, \dots, N - 1 \right\}$$

In this way, we can define the subset sets corresponding to each value of  $N.set$  (or  $N.obs$ ):

52

$$\Pi_{N.set} = \{\pi_1, \pi_2, \dots, \pi_{N.set-1}\} \quad \text{for } N.set = 2, 3, \dots, 2495$$

and we define our training data to be:

53

$$\Pi_{N.set}^{Train} = \bigcup_{i \in S} \pi_i$$

where

54

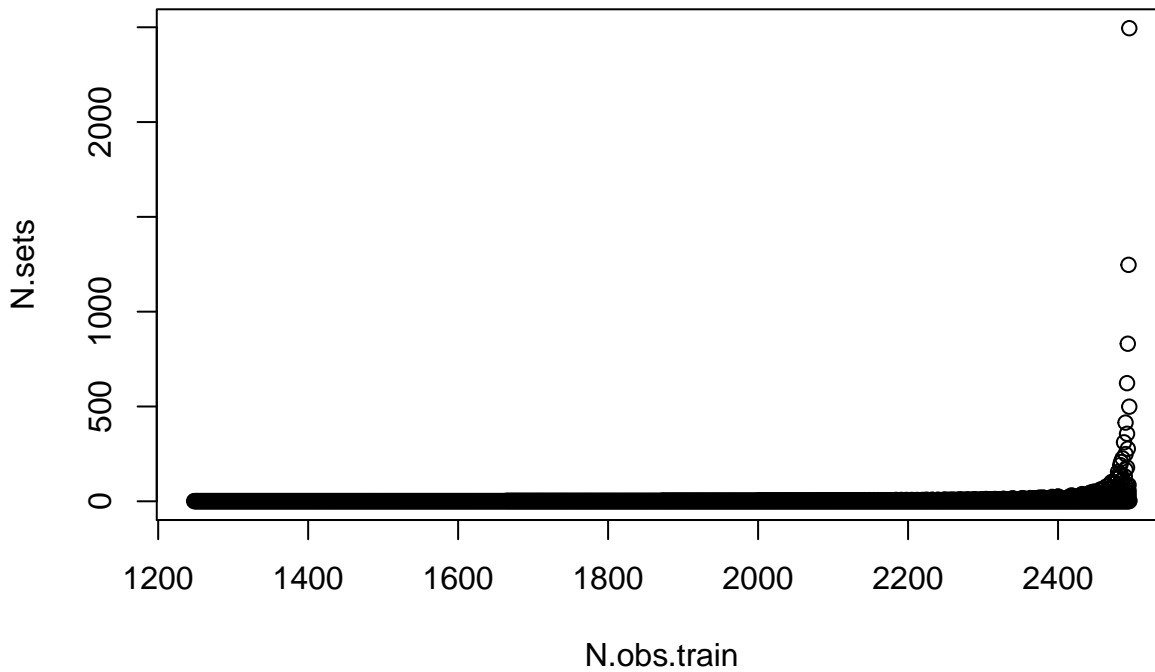
$$S = \Pi_{N.set} \setminus \{\pi_{N.set-1}\}$$

and we define our test data to be:

55

$$\Pi_{N.set}^{Test} = \pi_{N.set-1}$$

```
N.sets=c()  
N=2495  
N.obs=1:2495  
N.sets=floor(N/N.obs)  
N.train.sets=N.sets-1  
N.obs.train=N.sets*N.obs  
plot(N.sets~N.obs.train)
```



---

Code Appendix

59

## Script Dependencies

60

Package Dependencies

61

Working Directory

62

Load Data

63

---

64

## Begin Script

65

## End Script

66