Contest: Child Stuntedness 3

Problem: ChildStuntedness3

Problem Statement

Prizes

- 1st Place - $10,000
- 2nd Place - $5,000
- 3rd Place - $2,500
- 4th Place - $1,500
- 5th Place - $1,000

Stunting affects more than 165 million children, translating to one in four children worldwide. Children with stunted growth have an increased risk of early death, higher burden of disease, compromised physical capacities, and diminished cognitive development. This can reduce the productivity of an entire generation. The roots of stunted growth might start at prior to conception or in the womb, leading to low birth weight infants entering the world already at a deficit. Being able to predict, early in pregnancy, whether a child will have a low birth weight can help initiation of interventions leading to healthy live births. We need, then, to search not only for causes of low birth weight but also for methods for predicting birth-outcomes (weight or length at birth) and pregnancy duration (preterm babies are born before they spend the required 9 months in the womb).

Our goal is to determine a combination of early measures that would be a good predictor for size at birth (weight and length), pregnancy duration, and whether or not a child will require hospitalization at birth. In pursuit of this goal, we have collected time series data from ultrasounds on pregnant mothers. We would like you to use this data to predict a child's birth weight, length, head circumference and birth date (days from pregnancy start).

You may download the learning data set from here. The format for the data in the data set is a csv with details provided below:

Data Description

Column Variable Type Label/Description

- Id int Unique Fetus ID
- Sex int 0 = Male, 1 = Female
- Status STRING 1 of 2 values: CASE or CONTROL
- t.ultsnd float Estimated fetus gestational age from last menstrual recall date
- Columns 5-12 Odv float Dependent variables: Ultrasound observed measurements
- weight float w
- length float l
- head circumference c
- pregnancy duration (or birth date) b
- hosp (Was the baby hospitalized)

For each fetus given sex, status, and multiple ultrasound measurements(columns 5-12) during the pregnancy (time being the variable t.ultsnd). The data from the repeated ultrasounds provides a small time series that can be used for predicting the birth weight and day. More specifically each fetus has 6 ultrasounds done at regular intervals. For almost all IDs, the first ultrasound only one of 8 possible measurements is noted. For each remaining ultrasound each of the remaining 7 measurements are noted almost every time (there are a few cases with missing values). An example of the measurements for a single fetus is shown below.

```
id sx stat t.ultsnd     x1     x2     x3     x4     x5     x6     x7     x8     w      l      c      b      hosp
1  0  CASE 0.2678571429 0.0392 NA     NA     NA     NA     NA     NA     NA     0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.4214285714 NA     0.2782 0.2740 0.3377 0.2514 0.3498 0.3561 0.2708 0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.5392857143 NA     0.4425 0.4274 0.5287 0.4196 0.5278 0.5098 0.4378 0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.6464285714 NA     0.5400 0.5391 0.6561 0.5397 0.6131 0.5943 0.5167 0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.7714285714 NA     0.7092 0.6645 0.7784 0.6882 0.7631 0.7676 0.6840 0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.8964285714 NA     0.8246 0.7539 0.8786 0.7874 0.9151 0.8789 0.8180 0.7678 0.8922 0.9303 1.05   0
1  0  CASE 0.9964285714 NA     0.8909 0.8632 0.9051 0.8878 0.9311 0.9192 0.8534 0.7678 0.8922 0.9303 1.05   0
```

There will be five different flavors of test case, with an equal number of each:

- A single call to predict l, given oid, sex, t.ultsnd, x5 and x6.
- A single call to predict w, given oid, sex, t.ultsnd, x2, x3 and x8.
- A single call to predict c, given oid, sex, t.ultsnd, x4, x6 and x7.
- A single call to predict b, given oid, sex, t.ultsnd, x1, x2, x3, x4, x5, x6, x7 and x8.
- A single call to predict w,l,c,b, hosp using oid, sex, t.ultsnd, x1,x2,x3,x4,x5,x6,x7,x8.

For the tests to predict a single variable, score will be computed as 1000000 * (1 - SSE / SSE0), where SSE = the sum of the squared errors of the predictions, and SSE0 is the sum of the squared errors based on predicting each value as being the mean of all values presented in the training dataset.

For the test to predict all variables, score will be 2000000 * (1 - SSE / SSE0) + 2000000 * (1 - BR / BR0). Here, SSE = the sum of the squared Mahalanobis distance between the set of predictions and the true values of the variables W, L, C and B. BR = the sum of the Brier scores for each prediction for H. BR0 = the sum of Brier scores based on predicting each value as being the proportion of examples in the training data that required hospitalization. Brier score can be thought of as the squared error of a prediction for the special case that the real value is always exactly 0 or 1. As an example, for a prediction of of 0.3, the Brier score would either be (1.0 - 0.3)^2 = 0.49 or (0.0 - 0.3)^2 = 0.09, depending on whether the event happened or not.

When calculating Mahalanobis score, the inverse covariance matrix is based upon the complete dataset, and is as follows:

```
{  345.19152, -427.6975, -348.6226,  -76.35655 },
{ -427.69754, 1848.5556, -275.5745, -270.12371 },
{ -348.62264, -275.5745, 1661.5829, -148.26433 },
{  -76.35655, -270.1237, -148.2643,  905.19402 }
```

In the String[] trainingData, each String states a record of some fetus, and has 17 tokens, comma-separated, in the same order as described above in the table. The format of testingData is almost same as the trainingData, but with only 12 columns. The datas with same IDs are consecutive. Depending on the type of test case, some of columns 5-12 may be replaced with “NA”. For the calls predicting a single value, the returned value should be a double[] with a single value for each ID (in order by ID). For the call predicting all values, the returned value should be a double[] with five values for each ID, representing the predictions for W, L, C, B, and H respectively.

As an example, if the testing data contains several rows each for IDs 13, 4, and 9, then the return value should have fifteen elements: {W4, L4, C4, B4, H4, W9, L9, C9, B9, H9, W13, L13, C13, B13, H13}.

Notes on Data Set Generation

- All data values are normalized between 0 and 1 as part of data obfuscation requirements.
- The full data set contains approximately 32,000 lines, covering around 5600 ID values.
- The full data set is divided into 20% for example tests, 30% for provisional tests, and 50% for system tests. All data belonging to the same ID is placed in the same data set.
- For each test, approximately 66% of the data (from that segment) is selected for training, and the remainder for testing.
- For provisional tests, all example data is also added to the training set.
- For system tests, all example and provisional test data is also added to the training set.
- In all test cases, training and testing data formats are the same; when applicable, some values may be replaced by "NA", however.

Constraints

- Memory limit = 1GB
- Time limit = 60s for predicting a single value, 300s for predicting all values

Definition

Class: ChildStuntedness3
Method: predictW
Parameters: String[], String[]
Returns: double[]
Method signature:double[] predictW(String[] train, String[] test)

Method: predictL
Parameters: String[], String[]
Returns: double[]
Method signature:double[] predictL(String[] train, String[] test)

Method: predictC
Parameters: String[], String[]
Returns: double[]
Method signature:double[] predictC(String[] train, String[] test)

Method: predictB
Parameters: String[], String[]
Returns: double[]
Method signature:double[] predictB(String[] train, String[] test)

Method: predictAll
Parameters: String[], String[]
Returns: double[]
Method signature:double[] predictAll(String[] train, String[] test)
(be sure your methods are public)

Examples

0)

```
Predict All
Seed = 0
```

1)

```
Predict W
Seed = 0
```

2)

```
Predict L
Seed = 0
```

3)

```
Predict C
Seed = 0
```

4)

```
Predict B
Seed = 0
```

5)

```
Predict All
Seed = 1
```

6)

```
Predict W
Seed = 1
```

7)

```
Predict L
Seed = 1
```

8)

```
Predict C
Seed = 1
```

9)

```
Predict B
Seed = 1
```