CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 5 Solutions

Note: Your answers to exercises 2 and 3 may differ from mine, depending on which observations you removed. I removed the observations with ID = 317, ID = 142, ID = 61, ID = 127, ID = 214, ID = 321, and ID = 394.

1. One solution is to perform principal components analysis, for which I have provided a detailed description in Lecture 11, and then standarize the resulting uncorrelated explanatory variables.

Another solution is as follows. Let $\mathbf{X}$ be a $j \times 1$ vector containing the standardized explanatory variables. For a generic non-random $j \times j$ matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{AX}] = \mathbf{A}\mathbb{E}[\mathbf{X}] \quad \text{and} \quad \mathbb{V}[\mathbf{AX}] = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^T, \tag{1}$$

where $\mathbb{E}$ and $\mathbb{V}$ are the expectation and variance operators.

Let $\mathbf{R}$ denote the $j \times j$ matrix of correlations among the standardized explanatory variables, and assume that $\mathbf{R}$ is invertible. By Cholesky factorization, we may write $\mathbf{R} = \mathbf{LL}^T$, where $\mathbf{L}$ is a lower diagonal and invertible $j \times j$ matrix. [[Proof of invertibility: a $j \times j$ matrix is invertible if and only if its determinant is nonzero, and the determinant of $\mathbf{R}$ is the square of the determinant of $\mathbf{L}$.]]

Using relation (1) above with $\mathbf{A} = \mathbf{L}^{-1}$, we see that

$$\mathbb{E}[\mathbf{L}^{-1}\mathbf{X}] = \mathbf{L}^{-1}\mathbf{0} = \mathbf{0} \quad \text{and} \quad \mathbb{V}[\mathbf{L}^{-1}\mathbf{X}] = \mathbf{L}^{-1}\mathbf{LL}^T\mathbf{L}^{-1^T} = \mathbf{I}, \tag{2}$$

where $\mathbf{I}$ denotes the $j \times j$ identity matrix. Relation (2) establishes that $\mathbf{L}^{-1}\mathbf{X}$ contains standardized and uncorrelated explanatory variables.

Of course, in practice $\mathbf{R}$ is unknown and must be estimated from the data.

2a. The predicted HeartRate is

$$82.089 - 18.684 \tanh [0.74684 + 0.04186(1 - 2\text{ArrhythmiaAny}) + 0.25687(1 - 2\text{Female})$$

$$-0.00878 S_{AGE} + 0.11241 S_{HEIGHT} - 0.17813 S_{PINT} + 0.01874 S_{PRINT}$$

$$-0.16864 S_{QRSDUR} + 1.01443 S_{QTINT} - 0.14353 S_{TINT} - 0.06350 S_{WEIGHT}],$$

where

$$S_{AGE} = -2.70016 + 0.05764\text{Age}, \quad S_{HEIGHT} = -15.00931 + 0.09176\text{Height},$$

$$S_{PINT} = -3.65312 + 0.04042\text{PInt}, \quad S_{PRINT} = -3.55254 + 0.02286\text{PRInt},$$

$$S_{QRSDUR} = -5.44774 + 0.06063\text{QRSDur}, \quad S_{QTINT} = -11.44879 + 0.03096\text{QTInt},$$

$$S_{TINT} = -4.84898 + 0.02823\text{TInt}, \quad \text{and} \quad S_{WEIGHT} = -4.14434 + 0.06153\text{Weight}.$$

2b. The average squared error on the test data set is 77.33. Noting that $n_{test} = 112$ and that the variance of HeartRate within the test data set is 111.63, we obtain $R^2_{test,NeuralNetwork} = 0.301$. This is much larger than $R^2_{test,LinearRegression} = 0.022$, so the neural network is vastly better at making predictions on the test data set than the linear regression model.

2c. The predicted HeartRate for subject 5 is 69.856. This person's actual HeartRate is missing from the data set. The predicted HeartRate for subject 317 is 64.189. This person's actual HeartRate is reported as 163. [[Whether that is actually correct is unclear.]] The discrepancy between the actual HeartRate and the predicted HeartRate is 98.811.

2d. The standard deviation of HeartRate within the test data set is 10.566, so the discrepancy of 98.811 for subject 317 equals 9.35 standard deviations. Thus, the discrepancy is immense. [[Recall the definition of a standard deviation. In the absence of any statistical modeling, the typical discrepancy between an actual response and a predicted response should be 1 standard deviation.]]

2e. One option is to fit a linear regression model first, before fitting the neural network, and then eliminate from consideration for the neural network whichever explanatory variables were not selected for inclusion in the linear regression model.

Another option is to fit a regression tree first, before fitting the neural network, and then eliminate from consideration for the neural network whichever explanatory variables were not selected for inclusion in the regression tree.

A third option is to fit one neural network using a subset $\mathcal{S}_1$ of the explanatory variables, fit another neural network using another subset $\mathcal{S}_2$, and so forth. Choose whichever subset of explanatory variables yields the neural network with the smallest average squared error on the validation data set.

These options can be used in tandem. For instance, $\mathcal{S}_1$ in the third option can be defined as the subset of explanatory variables selected for inclusion in the linear regression model, $\mathcal{S}_2$ can be defined as the subset of explanatory variables selected for inclusion in the regression tree, and $\mathcal{S}_3$ can be defined as the full collection of explanatory variables.

3a. Average squared error on the validation data set was 101.70 with 8 neighbors, 117.30 with 16 neighbors, and 136.94 with 32 neighbors. As such, we retain 8 neighbors for the subsequent items.

3b. The average squared error on the test data set was 84.48, yielding $R^2_{test,NearestNeighbor} = 0.236$. This is better than $R^2_{test,LinearRegression} = 0.022$ but not as good as $R^2_{test,NeuralNetwork} = 0.301$, so the neural network is best at making predictions on the test data set.

3c. The predicted HeartRate for subject 5 is 64.875. This person's actual HeartRate is missing from the data set. The predicted HeartRate for subject 317 is 68.375. This person's actual HeartRate is reported as 163. The discrepancy between the actual HeartRate and the predicted HeartRate is 94.625.

3d. The standard deviation of HeartRate within the test data set is 10.566, so the discrepancy of 94.625 for subject 317 equals 8.96 standard deviations. Thus, the discrepancy is immense.