SIGNAL & DATA ANALYTICS IN IoMT: Day 3
Tech-in-Med Summer Camp

PREMANANDA INDIC, PH.D.
DEPARTMENT OF ELECTRICAL ENGINEERING
CALL FOR PAPERS: Computational Modeling of Physiological Systems

Identifying physiological origins of baroreflex dysfunction in salt-sensitive hypertension in the Dahl SS rat

Scott M. Bugenhagen, Allen W. Cowley, Jr., and Daniel A. Beard
Department of Physiology, Medical College of Wisconsin, Milwaukeee, Wisconsin
Submitted 3 February 2010, accepted in final form 23 March 2010

Bugenhagen SM, Cowley AW Jr, Beard DA. Identifying physiological origins of baroreflex dysfunction in salt-sensitive hypertension in the Dahl SS rat. Physiol Genomics 42: 23–41, 2010. First published March 30, 2010. doi:10.1152/physiolgenomics.00027.2010.—Salt-sensitive hypertension is known to be associated with dysfunction of the baroreflex control system in the Dahl salt-sensitive (SS) rat. However, neither the physiological mechanisms nor the genomic regions underlying the baroreflex dysfunction seen in this rat model are definitively known. Here, we have adopted a mathematical modeling approach to investigate the physiological and genetic origins of baroreflex dysfunction in the Dahl SS rat. We have developed a computational model of the overall baroreflex control system based on known physiological mechanisms to analyze telemetry-based blood pressure and heart rate data from two genetic strains of rat, the SS and normotensive SS.13Rn, on low- and high-salt diets. With this approach, physiological parameters are estimated, unmeasured physiological variables related to the baroreflex control system are predicted, and differences in these quantities between the two strains of rat on low- and high-salt diets are detected. Specific findings include: a significant selective impairment in sympathetic gain with high-salt diet in SS rats and a protection from this impairment in SS.13Rn rats, elevated sympathetic and parasympathetic offsets with high-salt diet in both strains, and an elevated sympathetic tone with high-salt diet in SS but not SS.13Rn rats. In conclusion, we have left unidentified because of these interactions. Thus, these types of measurements become diminishingly informative with an increased degree of genetic nonlinearity.

It seems, then, that more detailed phenotypic measurements are required to understand the underlying etiology and to make sense of the genetics associated with this complex disease. Of course, this is not always possible; many measurements of interest are either inaccessible or simply not practical to obtain. In addition, many of these measurements are operating-point dependent and are influenced to a high degree by physiological state. Methods of obtaining these measurements often require invasive techniques that introduce stressors (surgical, pharmacological, etc.) that may themselves alter physiological state and therefore the observed measurements. Thus, differences detected in such experimental measurements may not always indicate differences in underlying physiology but can rather indicate differences in confounding variables related to experimental conditions and/or methods.

Mechanistic mathematical models offer a powerful complement to laboratory measurements (5). By accounting for the
Example 1: Hypertension

Hypothesis: To test the hypothesis that high and low level of salt contents can identify dysfunction in baroreflex mechanisms to indicate hypertension
Example 1: Hypertension

Give two different levels of salt, low level (blue), high level (red) to dysfunction rat (SS; n=9) and compare with healthy rat (SSBN13; n=6)
Example 1: Hypertension

Features:

Mean Blood Pressure (BP)

Standard Deviation of BP
Example 1: Hypertension

Features:

Mean Blood Pressure (BP)
Standard Deviation of BP
Example 1: Hypertension

Features:

Mean Blood Pressure (BP)

Standard Deviation of BP
Example 1: Hypertension

Is there any predictability?

Mean Blood Pressure (BP)

Standard Deviation of BP
Project 1: Prediction of House Value
SECTION 1

➢ Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=20640)

- longitude: A measure of how far west a house is; a higher value is farther west
- latitude: A measure of how far north a house is; a higher value is farther north
- housingMedianAge: Median age of a house within a block; a lower number is a newer building
- totalRooms: Total number of rooms within a block
- totalBedrooms: Total number of bedrooms within a block
- population: Total number of people residing within a block
- households: Total number of households, a group of people residing within a home unit, for a block
- medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- medianHouseValue: Median house value for households within a block (measured in US Dollars)
- oceanProximity: Location of the house w.r.t ocean/sea

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
- Linear Regression
- SVM

https://www.kaggle.com/camnugent/california-housing-prices
### SECTION 1

#### Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=5000)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Validation (10 fold) RMSE</th>
<th>R-squared</th>
<th>Test RMSE</th>
<th>Test R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression (using App)</td>
<td>69010</td>
<td>0.64</td>
<td>65501</td>
<td>0.67</td>
</tr>
<tr>
<td>Linear SVM (using App)</td>
<td>70382</td>
<td>0.64</td>
<td>66858</td>
<td>0.66</td>
</tr>
</tbody>
</table>
SECTION 1

➢ Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=5000)
SECTION 2

Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data
2. Identify the features (find correlations between variables)
3. Preprocess the data (missing values, outliers)
4. Train the Model
5. Select the best performance model
SECTION 2

Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data
2. Identify the features (find correlations between variables)
3. Preprocess the data (missing values, outliers)
4. Train the Model
5. Select the best performance model
SECTION 2

➢ Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue \( (N=5000) \)

1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model

Visualize the data, Summarize variables, data cleaning, pre-processing if needed
SECTION 2

➢ Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model

FIND VARIABLE CORRELATIONS TO EACH OTHER
AND THE MEDIAN_HOUSE_VALUE
SECTION 2

➢ Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model

207 Missing values, replace with median values

ocean_proximity: 20636×1 categorical Values:

- <1H OCEAN: 9135
- INLAND: 6550
- ISLAND: 5
- NEAR BAY: 2289
- NEAR OCEAN: 2657

Visualize the data, Summarize variables, data cleaning, pre-processing if needed
SECTION 2

Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data
2. Identify the features (find correlations between variables)
3. Preprocess the data (missing values, outliers)
4. Train the Model
5. Select the best performance model

SPLIT INTO TRAINING AND TEST DATA AND FIT REGRESSION MODELS
SECTION 2

Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model
SECTION 2

Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Validation RMSE</th>
<th>Test RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin regression</td>
<td>70071</td>
<td>65501</td>
</tr>
<tr>
<td>Lin. Regression – fewer variables</td>
<td>69031</td>
<td>65357</td>
</tr>
<tr>
<td>SVM –linear kernel</td>
<td>116370</td>
<td>116130</td>
</tr>
<tr>
<td>SVM –Gaussian Kernel</td>
<td>60099</td>
<td>57708</td>
</tr>
</tbody>
</table>
LASSO REGRESSION

➢ Linear Regression

\[ \hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \ldots + \theta_n x_n^i \]

\[ \hat{Y} = \Theta^T X \]

- Gradient Descent by **Louis Augustin Cauchy** in 1847

Cost Function to Minimize

\[ J = \left( (\hat{y}^i - y^i)^2 \right) = (\hat{Y} - Y)^T (\hat{Y} - Y) \]
LASSO REGRESSION

➢ Linear Regression with Lasso

\[ \hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \ldots \ldots + \theta_n x_n^i \]

\[ \hat{Y} = \Theta^T X \]

Cost Function to Minimize

\[ J = \left( (\hat{y}^i - y^i)^2 \right) = (\hat{Y} - Y)^T (\hat{Y} - Y) + \lambda \sum_{j=1}^{n} |\theta_j| \]
SECTION 3

➢ Home Value Prediction (Lasso Regression): 9 features to predict medianHouseValue (N=5000)

\[ J = \left( (\hat{y}^i - y^i)^2 \right) = (\bar{Y} - Y)^T (\bar{Y} - Y) + \lambda \sum_{j=1}^{n} |\theta_j| \]

Lambda

Lasso removes the ‘total_rooms’ and ‘Ocean Proximity_inland’ variables as least important.

RMSE on test data with 7 features = 66443
SECTION 3

Home Value Prediction (Lasso Regression): 9 features to predict medianHouseValue (N=5000)

- 'longitude'  -3.2643  All coefficients multiplied by 10^4
- 'latitude'   -3.2856
- 'housing_median_age' 0.1177
- 'total_rooms'  0
- 'total_bedrooms' 0.0074
- 'population'  -0.0028
- 'households'  0.0014
- 'median_income' 3.8702
- 'op_vbl'}  0
Project 2: Classification of House Value
SECTION 1: Learner App

➢ Home Value Classification: 9 features to classify high vs low medianHouseValue

longitude: A measure of how far west a house is; a higher value is farther west
latitude: A measure of how far north a house is; a higher value is farther north
housingMedianAge: Median age of a house within a block; a lower number is a newer building
totalRooms: Total number of rooms within a block
totalBedrooms: Total number of bedrooms within a block
population: Total number of people residing within a block
households: Total number of households, a group of people residing within a home unit, for a block
medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
medianHouseValue: Median house value for households within a block (measured in US Dollars)
oceanProximity: Location of the house w.r.t ocean/sea

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
Logistic Regression
SVM

https://www.kaggle.com/camnugent/california-housing-prices
SECTION 1: Learner App

 Prediction of House Price Classification Problem

Confusion Matrix

<table>
<thead>
<tr>
<th>True Positive</th>
<th>False Negative</th>
<th>Total Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positive</td>
<td>True Negative</td>
<td>Total Negative</td>
</tr>
</tbody>
</table>

True Positive Rate = True Positive / Total Positive

True Negative Rate = True Negative / Total Negative = 1 – False Positive Rate
SECTION 1: Learner App

DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION
SECTION 1: Learner App

DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION

classificationLearner(Ttrain,'hi_lo_label');

Demo with logistic regression and linear SVM
SECTION 2: Raw Data Analysis

Visualize the data, Summarize variables, data cleaning, pre-processing if needed

207 Missing values, replace with median values

ocean_proximity: 20636×1 categorical
Values:
- <1H OCEAN 9135
- INLAND 6550
- ISLAND 5
- NEAR BAY 2289
- NEAR OCEAN 2657
SECTION 3: Correlation Analysis

FIND VARIABLE CORRELATIONS TO EACH OTHER AND THE MEDIAN HOUSE VALUE

\[ R_{pp} = \text{corr} \left( \text{table2array} \left( T1(:, \text{select \_vars}) \right) \right) \];
**SECTION 4: Logistic Regression**

**SPLIT INTO TRAINING AND TEST DATA AND FIT LOGISTIC REGRESSION MODEL**

Estimated Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-Stat</th>
<th>pValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-154.19</td>
<td>14.421</td>
<td>-10.692</td>
<td>1.10e-26</td>
</tr>
<tr>
<td>longitude</td>
<td>-1.7683</td>
<td>0.17448</td>
<td>-10.135</td>
<td>3.87e-24</td>
</tr>
<tr>
<td>latitude</td>
<td>-1.8133</td>
<td>0.18905</td>
<td>-9.6018</td>
<td>7.85e-22</td>
</tr>
<tr>
<td>housing_median_age</td>
<td>0.044239</td>
<td>0.0051484</td>
<td>8.5928</td>
<td>4.90e-18</td>
</tr>
<tr>
<td>total_rooms</td>
<td>0.0003444</td>
<td>5.7387e-05</td>
<td>5.364</td>
<td>3.00e-0561</td>
</tr>
<tr>
<td>total_bedrooms</td>
<td>0.00080298</td>
<td>0.00084259</td>
<td>0.95299</td>
<td>0.3406</td>
</tr>
<tr>
<td>population</td>
<td>-0.0023529</td>
<td>0.00040995</td>
<td>-11.207</td>
<td>3.77e-29</td>
</tr>
<tr>
<td>households</td>
<td>0.0039573</td>
<td>0.00046559</td>
<td>8.7854</td>
<td>2.85e-05</td>
</tr>
<tr>
<td>median_income</td>
<td>1.0172</td>
<td>0.053904</td>
<td>18.87</td>
<td>2.01e-06</td>
</tr>
<tr>
<td>ocean_proximity_INLAND</td>
<td>-0.053285</td>
<td>0.24937</td>
<td>-0.21368</td>
<td>0.8308</td>
</tr>
<tr>
<td>ocean_proximity_ISLAND</td>
<td>0</td>
<td>0</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ocean_proximity_NEAR_BAY</td>
<td>-0.10616</td>
<td>0.19861</td>
<td>-0.53449</td>
<td>0.593</td>
</tr>
<tr>
<td>ocean_proximity_NEAR_OCEAN</td>
<td>0.11076</td>
<td>0.15948</td>
<td>0.6945</td>
<td>0.49737</td>
</tr>
</tbody>
</table>

mdl = fitglm([Ttrain(:,1:9) table(y)],'Distribution','binomial');

Remove Insignificant features
mdl1 = fitglm([Ttrain(:,[1:4 6:8]) table(y,'variablenames',{'Hi_lo_label'})],'Distribution','binomial');

plotDiagnostics(mdl1,'leverage')
SECTION 6: Classification (Clean Data)

TEST MODEL FOR TWO CLASS CLASSIFICATION (Logistic Regression)

2 class Confusion Matrix

Test Data N = 1500
(30% of 5000)

Missing Values
Insignificant Features
Outliers
SECTION 7: SVM Classification

REGULARIZATION OF VARIABLES DONE AUTOMATICALLY, NO NEED TO CHOOSE FEATURES SEPARATELY AS WAS DONE EARLIER FOR LOGISTIC REGRESSION

Test Data N = 1500  
(30% of 5000)

Linear SVM

```
SVMModel = fitcsvm(Ttrain(:,1:9),y,'standardize',true);
```
SECTION 8: SVM Classification

LINEAR vs RADIAL BASIS FUNCTION (RBF) KERNEL

\[ \text{fitcsvm}([x1 \ x2], y1); \]

\[ \text{fitcsvm}([x1 \ x2], y1, 'KernelFunction', 'rbf'); \]

\[ x1: \text{Age of House} \]
\[ X2: \text{Median Income} \]
SECTION 9: Multiclassification (SVM)

ONE CLASS vs REST

Also perform one to one class

Mdl = fitcecoc(Ttrain(:,1:8),y,'Learners',t,'Coding',coding,'ResponseName',responseName,... 'PredictorNames',predictorNames,'ClassNames',classNames);
Mdlp = fitcecoc(Ttrain(:,1:8),y,'Learners',t,'FitPosterior',true,...
'ClassNames',{'low','mod','high'},...
'Verbose',2);
Project 3: Oxygen desaturation
Differentiating Smokers vs Non-Smokers

Pattern Analysis of Oxygen Saturation Variability in Healthy Individuals: Entropy of Pulse Oximetry Signals Carries Information about Mean Oxygen Saturation

Amar S. Bhogal and Ali R. Mani*
LG, Division of Medicine, University College London, London, United Kingdom

Pulse oximetry is routinely used for monitoring patients’ oxygen saturation levels with little regard to the variability of this physiological variable. There are few published studies on oxygen saturation variability (OSV), with none describing the variability and its pattern in a healthy adult population. The aim of this study was to characterize the pattern of OSV using several parameters; the regularity (sample entropy analysis), the self-similarity (detrended fluctuation analysis (DFA)) and the complexity (multiscale entropy (MSE) analysis). Secondly, to determine if there were any changes that occur with age. The study population consisted of 26 individuals. The “young” population consisted of 20 individuals [Mean (±1 SD) age = 21.0 (±1.36 years)] and the “old” population consisted of 16 individuals [Mean (±1 SD) age = 50.0 (±10.4 years)]. Through DFA analysis, OSV was shown to exhibit fractal-like patterns. The sample entropy revealed...