Computational Optimization and Statistical Methods for Big Data Analytics: Applications in Neuroimaging

W. Art Chaovilawongse
Professor
Department of Industrial & Systems Engineering

Adjunct Professor
Department of Radiology (joint)
Integrated Brain Imaging Center (IBIC), UW Medical Center

Shuai Huang
Assistant Professor
Department of Industrial & Systems Engineering

Outline

• Data Analytics: Overview
• Classification and Regression
• Regularization/Feature Selection
• Health Care and Medical Imaging
  - Cognitive Neuroscience
  - Attention Deficit/Hyperactivity Disorder
  - Connectivity Modeling in Alzheimer’s Disease

Data Analytics

- Reactive analytics focuses on providing statistics for current and historical data to provide insights into what happened and why it happened.
  - statistical modeling, trend reporting, visualization, association and correlation analysis

- Proactive (predictive) analytics focuses on using a known data (called training data), which includes input data features (aka attributes) and response values (aka target patterns), to build a predictive (decision) model to make predictions on unseen data (called test data).
  - supervised learning – support vector machines, linear regression/classification, nonlinear regression (generalized linear model, logistic), decision tree, Bayesian learning, nearest neighbor, and neural networks

Reactive Data Analytics:
Discovering Associations

- Unsupervised learning (hypothesis-driven, clustering, association)
- Investigating “associations” of medical data and outcomes
- Reporting $R^2$ (R-Squared): fitness of regression models
- Reporting $p$-value: separability or fitness of the medical data

Current Landscape: Data Analytics

Machine Learning
  - Feature Extraction
  - Feature Selection
  - Classification/Regression
  - Clustering

Optimization & Statistics
  - Mathematical Programming
  - Convex Optimization
  - Multiple hypothesis testing
  - Bayesian/MCMC
Predictive Data Analytics: It's all about "optimization" and "statistics"

- **Linear Regression**: Min Sum Squared Error
  \[ y = X\beta + e \]

- **Bayesian Classifier**: Max Posterior Probability
  \[ P(Y = y | X) = \frac{P(Y = y)}{P(X)} \]

- **Support Vector Machine**: Max Margin + Min Error
  \[ \arg \min \{ \frac{1}{2} \|w\|^2 + C \sum \xi_i \} \]
  subject to \( \|x_i - w\| - b \geq 1 - \xi_i, \quad \xi_i \geq 0 \)

Support Vector Machines: Concepts and Models

- An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class.
- The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab.
- The best hyperplane for an SVM is the one with the largest margin between the two classes.
- Margin means the maximal width of the slab parallel to the hyperplane.
- **L1-Norm Version**: Minimize \( \|w\|_1 \) subject to \( \sum \xi_i \leq C \)

Classification/Regression Tree

- Start with all input data, and examine all possible binary splits on every feature.
- Select a split with best optimization criterion.
  - Gini's Diversity Index: \( 1 - \sum \hat{p}(i)^2 \)
  - Deviance (deviance): \( -\sum p(i) \log p(i) \)
  - subject to the MinLeaf constraint – min # of observations in the child node
- Impose the split.
- Repeat recursively for the two child nodes.

Support Vector Machines: Solution Approaches

- Sequential Minimal Optimization (SMO) minimizes the one-norm problem by a series of two-point minimizations.
- Iterative Single Data Algorithm (ISDA) solves the one-norm problem using a series on one-point minimizations but does not respect the linear constraint, and does not explicitly include the bias term in the model.
- You can solve the one-norm problem using any quadratic programming solver (e.g., quadprog in Matlab's Optimization Toolbox)

Existing Toolboxes

- Supervised Learning: Regression, support vector machines, patterns and pattern recognition classification, decision trees
  - Linear Regression
  - Nonlinear Regression
  - Generalized Linear Models
  - Support Vector Machines
  - Decision Trees and Ensemble Methods
  - Feature Selection and Feature Engineering
  - Neural Networks

Mathworks © 2015

Classification/Regression Tree

- Start with all input data, and examine all possible binary splits on every feature.
- Select a split with best optimization criterion.
  - Gini's Diversity Index: \( 1 - \sum \hat{p}(i)^2 \)
  - Deviance (deviance): \( -\sum p(i) \log p(i) \)
  - subject to the MinLeaf constraint – min # of observations in the child node
- Impose the split.
- Repeat recursively for the two child nodes.
Naive Bayes Classifier

- The naive Bayes classifier is designed for use when predictors are independent of one another within each class.

\[ P(\omega | X) = \frac{P(X | \omega) \cdot P(\omega)}{P(X)} \]

- The key idea is to estimate the distributions.
  - Normal (Gaussian) Distribution
  - Kernel Distribution
  - Multinomial Distribution
  - Multivariate Multinomial Distribution

A good rule-of-thumb

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Predictive Accuracy</th>
<th>Training Speed</th>
<th>Prediction Speed</th>
<th>Memory Usage</th>
<th>Easy to Interpret</th>
<th>Handle Categorical Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>Medium</td>
<td>Fast</td>
<td>Fast</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SVM</td>
<td>High</td>
<td>Medium</td>
<td>*</td>
<td>*</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>Medium</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>**</td>
<td>Fast**</td>
<td>Medium</td>
<td>High</td>
<td>No</td>
<td>**</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>**</td>
<td>Fast</td>
<td>Fast</td>
<td>Low</td>
<td>Yes</td>
<td>Yes**</td>
</tr>
</tbody>
</table>

- Naive Bayes speed and memory usage are good for simple distributions, but poor for kernel distributions and large data sets.
- Nearest Neighbor usually has good predictions in low dimensions, but poor predictions in high dimensions. Nearest Neighbor can have either continuous or categorical predictors, but not both.
- Discriminant Analysis is accurate when the modeling assumptions are satisfied (multivariate normal by class). Otherwise, the predictive accuracy varies.

What’s the Challenge in Healthcare?

Interpretable/generalizable prediction model

- Occam’s razor (law of parsimony)
  - Simplicity is a goal in itself
  - Simplicity leads to greater accuracy

- When the number of features far exceeds the number of samples
  - The number of predictor vars (p) >> The number of observations (n)
  - The number of unknown vars >> The number of linear equations

Ill-posed problem = Overfitting

- This is common in healthcare: never collect enough data samples
- Feature selection is used to construct practical decision models

Feature Selection: Common Approaches

- Feature transformation (e.g., independent component analysis, principal component analysis, factor analysis) is NOT feature selection.

- Using a stepwise (sequential) procedure to select features based on statistical significance in
  - Stepwise regression: Correlation with trained targets (e.g., partial least square, regression weights)
  - Important/Separability index: Separation between two classes (e.g., t-test, mutual information, Fisher’s)

- Sequential feature selection: Similar to stepwise regression but used with classification algorithms

Outline

- Data Analytics: Overview
- Classification and Regression
- Regularization/Feature Selection
- Health Care and Medical Imaging
- Applications in Brain Research
  - Cognitive Neuroscience
  - Attention Deficit/Hyperactivity Disorder
  - Connectivity Modeling in Alzheimer’s Disease

Nearest Neighbor

Key factors:

- # of neighbors
- Distance matrix
  - Euclidean
  - Mahalanobis
  - Cosine
  - Correlation
  - Spearman
  - Hamming
  - Jaccard

Feature Selection:

- Occam's razor: Simplicity is a goal in itself
- Simplicity leads to greater accuracy

- When the number of features far exceeds the number of samples
  - The number of predictor vars (p) >> The number of observations (n)
  - The number of unknown vars >> The number of linear equations

Ill-posed problem = Overfitting

- This is common in healthcare: never collect enough data samples
- Feature Selection is used to construct practical decision models
**Regularization:**

**Objective Function**
- Common objective function of regression/classification: Minimize Prediction/Regression Error
- Regularization - process of introducing a penalty term in the objective function to avoid overfitting
  - Minimize Prediction/Regression Error + Penalty
- Error
  - Regression error: L-1, L-2 norms
  - Classification error: L-0 norm (logistic regression, hinge loss)
- Penalty
  - Continuous: L-1, L-2 norms
  - Discrete: L-0 norms (control the # of features/vars)

**Least Absolute Shrinkage and Selection Operator (Lasso)**
- Lasso (Tibshirani, 1996) is a very popular technique for variable selection for high-dimensional data.
  - A shrinkage and selection method for linear regression that minimizes the sum of squared errors, with a L1-norm penalty
  - \[
  \text{minimize} \quad \frac{1}{2} \| y - X \beta \|^2 + \lambda \| \beta \|_1
  \]
  - Lasso vs. Ridge regression vs. Elastic net
  - If the loss function is replaced by hinge loss, it is L1 norm SVM
  - If the loss function is replaced by logistic function, it’s called logistic regression

**Incorporating Domain Knowledge:**

**System-level understanding**
- Regularization usually has no control over what features to be selected - making it difficult to interpret the final model
- Make use of a system-level understanding of the data (e.g., prior knowledge or hypotheses)
- Test if incorporation of such prior knowledge will help the model generalize better (i.e., more accurate)

**Generalized Lasso:**

**Incorporating Domain Knowledge or Information Theory**
- **Lasso with fixed features**: fix a set of k desired features in the model - reduction to the standard Lasso problem
  - Objective: Fix a set of k desired features in regression while using Lasso
  - Method: WLOG, consider those features as \(X_{p-k+1}, X_{p-k+2}, \ldots, X_p\), fixing those features is equivalent to setting them as “penalty free” in regression, i.e., in the matrix, we have \(d_{p-k+1} = d_{p-k+2} = \ldots = d_p = 0\). Now \(D\) is rank-deficient matrix. So we make it full rank by remove the last k rows from the original matrix \(D\), and adding \(k \times p\) matrix \(A\). The original matrix can be expressed as \(\bar{D} = [D \ A]^T\), where \(D\) is the upper part \((p - k) \times p\) matrix

**Making prediction models more interpretable/generalizable**
- Motivation:
  - Screen good/informative features
  - Incorporating prior knowledge
  - Consider the generalized Lasso model

**Regularization with domain knowledge:**

**Lasso with fixed features**
Proof:
Reduction to a standard Lasso problem

\[ \hat{\beta} = [D \ A]^T, \] where \( D \) is the upper part \((p-k) \times p\) matrix, and \( A \) is \( k \times p \) matrix.

Since \( A \)'s rows are orthogonal to those in \( D \), following the procedures in Tibshirani's paper (10) and (11), we get a standard lasso problem regarding \( \beta_{\theta} \) which is related to the coefficient vector of the first \( p-k \) features that are NOT in the fixed set.

\[
\begin{align*}
\min_{\beta} & \quad \frac{1}{2} \|Y - X \beta\|^2 + \lambda \sum_{j=1}^{p-k} |\beta_j| \\
\text{subject to} & \quad \|\beta\|_0 = k
\end{align*}
\]

where \( P = \begin{bmatrix} X_k^T X_{\theta}^T \end{bmatrix}^T, X = \begin{bmatrix} X_1 \ A_1 \ A_2 \ \cdots \ A_{p-k} \end{bmatrix}, X_{\theta} = \begin{bmatrix} X_{\theta 1} \ \cdots \ X_{\theta k} \end{bmatrix}, P = \begin{bmatrix} X_{\theta 1} \ \cdots \ X_{\theta k} \end{bmatrix} \).

The solution from lasso is \( \hat{\beta}_{\theta} \) and \( \hat{\beta}_{\theta} = (X_k^T X_{\theta}^T)^{-1} X_k^T (Y - X \hat{\beta}) \).

\[
\hat{\beta} = [D \ A]^T \hat{\beta} = \hat{\beta}_{\theta}
\]

---

Generalized Lasso:
Incorporating Domain Knowledge or Information Theory

- **Logical Lasso**: If feature \( i \) selected, then feature \( j \) must be selected

Let \( D_{ij} = -\alpha \Rightarrow D_{ij} = 0 \), \( D_{ij} = 1 \).

\[
\begin{align*}
\theta_0 &= D_{ij} \\
\theta_1 &= D_{ij} \beta - (\beta_1 + \beta_2 + \cdots + \beta_{ij} + \cdots + \beta_{ij} \cdots + \beta_{ij}) \\
\theta_2 &= D_{ij} \theta_1 - (\beta_1 + \beta_2 + \cdots + \beta_{ij} + \cdots + \beta_{ij} \cdots + \beta_{ij}) \\
\text{set} \ i \& \ j \ \text{desired, solve lasso}
\end{align*}
\]

If \( \theta_0 = 0 \), then \( \beta_0 = a \beta_j \), and \( j \) being selected together.

If \( \theta_0 = -1 \), then \( \beta_0 = a \beta_j \), \( \beta_0 = a \beta_j \), and \( j \) being selected together.

---

Regularization with domain knowledge

- **Lasso with fixed features**: Fix a set of \( k \) desired features in the model – reduction to the standard Lasso problem

- **Lasso with an adjustable penalty term**: Penalize individual features differently based on prior knowledge

It is equivalent to set different \( \lambda \) for each dimension of \( \beta \), where \( \lambda_i = \alpha_i \lambda \). This can achieve the goal of setting different \( \lambda \) for each feature.

\[
\min_{\beta} \frac{1}{2} \|Y - X \beta\|^2 + \sum_{i=1}^{k} \lambda_i |\beta_i|
\]

- **Logical Lasso**: If feature \( i \) selected, then feature \( j \) must be selected

Let \( D_{ij} = -\alpha \Rightarrow D_{ij} = 0 \), \( D_{ij} = 1 \).

\[
\begin{align*}
\theta_0 &= D_{ij} \\
\theta_1 &= D_{ij} \beta - (\beta_1 + \beta_2 + \cdots + \beta_{ij} + \cdots + \beta_{ij} \cdots + \beta_{ij}) \\
\theta_2 &= D_{ij} \theta_1 - (\beta_1 + \beta_2 + \cdots + \beta_{ij} + \cdots + \beta_{ij} \cdots + \beta_{ij}) \\
\text{set} \ i \& \ j \ \text{desired, solve lasso}
\end{align*}
\]

---

Conditional Correlation!

**Ground Truth V.S. Correlations**

- Difficult to interpret, many correlations are byproducts of the intrinsic correlations
- Inefficient to capture changes, e.g., treatment effect evaluation
Another Example: A Spring System

![Image of a spring system with nodes and arrows]

<table>
<thead>
<tr>
<th>Observed correlations</th>
<th>Conditional correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.83 0.67 0.50 0.33 0.17</td>
<td>2 -1 0 0 0</td>
</tr>
<tr>
<td>0.67 1.33 1.00 0.67 0.33</td>
<td>-1 2 -1 0 0</td>
</tr>
<tr>
<td>0.50 1.00 1.50 1.00 0.50</td>
<td>0 0 -1 2 -1</td>
</tr>
<tr>
<td>0.33 0.67 1.00 1.33 0.67</td>
<td>0 0 0 -1 2</td>
</tr>
<tr>
<td>0.17 0.33 0.50 0.67 0.83</td>
<td></td>
</tr>
</tbody>
</table>

Generally, how to Learn the Intrinsic Interactions using Data?

How does ML Estimation Work?

![Diagram of observed covariance matrix and implied covariance matrix]

Trade-off
- Larger $\lambda$ → sparser $\Theta$ (more zero)
- Model fit (likelihood)
- $\ell_1$-norm: Model complexity

Formulation of Sparse Inverse Covariance Estimation (SICE)

<table>
<thead>
<tr>
<th>Our developments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure identification: developed an efficient computational algorithm (SICE) to identify the zero elements in $\Theta$</td>
</tr>
<tr>
<td>Strength estimation: proved a monotone property of SICE which provides a method for estimating the strength of the connections between brain regions</td>
</tr>
<tr>
<td>Applied to the ADNI data set</td>
</tr>
</tbody>
</table>

Illustration of SICE

![Diagram of SICE illustration]

Penalty parameter from small tolerance

Monotone Property of SICE

A quasi measure for connectivity strength
- Definition: two brain regions have connectivity if there is a path (can be one or more arcs) between them
- Monotone property: when $X_i$ and $X_j$ become disconnected at $\lambda_i$, they will never connect again at any $\lambda_j > \lambda_i$ (then $\lambda_i$ can be a strength measurement)

![Diagram of two islands in the network and proof for monotone property]

Basic idea
- Recall that two brain regions have a connection if a path between them
- No path means the two regions belong to two isolated “islands” in the network
- Mathematically two isolated “islands” in the network are equivalent to two blocks in the corresponding covariance matrix.
Proof for Monotone Property

We prove that, for any \( \lambda_1 > \lambda_2 \), \( \Sigma^e \) has the same block structure as \( \Sigma^p \).

\[
\Sigma^p_{ij} = \begin{bmatrix}
\Sigma^p_{11} & \cdots & \Sigma^p_{1k} \\
\cdots & \ddots & \cdots \\
\Sigma^p_{k1} & \cdots & \Sigma^p_{kk}
\end{bmatrix}
\]

\[
\Sigma^e_{ij} = \begin{bmatrix}
\Sigma^e_{11} & \cdots & \Sigma^e_{1k} \\
\cdots & \ddots & \cdots \\
\Sigma^e_{k1} & \cdots & \Sigma^e_{kk}
\end{bmatrix}
\]

\[
\Sigma^p_{ij} = \begin{bmatrix}
\Sigma^p_{11} & \cdots & \Sigma^p_{1k} \\
\cdots & \ddots & \cdots \\
\Sigma^p_{k1} & \cdots & \Sigma^p_{kk}
\end{bmatrix}
\]

\[
\Sigma^e_{ij} = \begin{bmatrix}
\Sigma^e_{11} & \cdots & \Sigma^e_{1k} \\
\cdots & \ddots & \cdots \\
\Sigma^e_{k1} & \cdots & \Sigma^e_{kk}
\end{bmatrix}
\]

Outlining

- Data Analytics: Overview
- Classification and Regression
- Regularization/Feature Selection
- Health Care and Medical Imaging
- Applications in Brain Research
  - Cognitive Neuroscience
  - Attention Deficit/Hyperactivity Disorder
  - Connectivity Modeling in Alzheimer's Disease

Medical Imaging 101: Alphabet Soup

Medical Imaging: Brain Imaging Example

- Computed tomography (CT) or Computed Axial Tomography (CAT) scanning uses a series of x-rays of the head taken from many different directions – typically used for quickly viewing brain injuries (hard tissues).

- Magnetic resonance imaging (MRI) uses magnetic fields and radio waves to produce high quality three-dimensional images of brain structures (soft tissues).

- Position emission tomography (PET) measures emissions from radioactively labeled metabolically active chemicals (tracers) that have been injected into the bloodstream. It produces images of the distribution of the chemicals in the brain.
MRI/Functional MRI
- MRI/fMRI are largely used due to its low invasiveness, lack of radiation exposure, and relatively wide availability.

Outline
- Data Analytics: Overview
  - Classification and Regression
  - Regularization/Feature Selection
- Health Care and Medical Imaging
  - Applications in Brain Research
    - Attention Deficit/Hyperactivity Disorder
    - Cognitive Neuroscience
    - Connectivity Modeling in Alzheimer's Disease

ADHD Diagnosis: Background and Motivation
- No single standard test for ADHD diagnosis. The gold standard diagnosis process requires experiences and extended involvement from physicians
- ADHD is thought to be caused by impairments in executive functioning modulated by cortex and associated brain networks
  - Cortical thickness of region of interests (ROIs) in the cortex can be measured by structural MRI
- Data: A total of 35 right handed participants (23 ADHD, 12 Control)
  - children and adolescents between 9 and 15 years of age matched on gender, socioeconomic status, and ethnicity
  - 45 cortical thickness features (regions of interest).

ADHD Diagnosis: Using Structural MRI Analysis
- Blascik JC, Xiao C, Mehta SH, Chaovalitwon W, and Grabowski TG
  - Psychiatry and Behavioral Medicine
  - Seattle Children's Hospital
  - Radiology, Industrial & Systems Engineering, Medicine
  - Integrated Brain Imaging Center
  - University of Washington, Seattle, WA, USA

ADHD Diagnosis: Background and Motivation
- No single standard test for ADHD diagnosis. The gold standard diagnosis process requires experiences and extended involvement from physicians
- ADHD is thought to be caused by impairments in executive functioning modulated by cortex and associated brain networks
  - Cortical thickness of region of interests (ROIs) in the cortex can be measured by structural MRI
- Data: A total of 35 right handed participants (23 ADHD, 12 Control)
  - children and adolescents between 9 and 15 years of age matched on gender, socioeconomic status, and ethnicity
  - 45 cortical thickness features (regions of interest).

ADHD Diagnosis: Using Structural MRI Analysis
- Blascik JC, Xiao C, Mehta SH, Chaovalitwon W, and Grabowski TG
  - Psychiatry and Behavioral Medicine
  - Seattle Children's Hospital
  - Radiology, Industrial & Systems Engineering, Medicine
  - Integrated Brain Imaging Center
  - University of Washington, Seattle, WA, USA

Our Algorithm: Fixing informative ROIs
- Fixing informative ROIs
- Prediction Results: Comparison with Other Feature Selection Methods
  - Algorithm
    - Testing Accuracy
    - Training Accuracy
    - # of Selected Features
  - Comparison with Other Feature Selection Methods
    - MI + Lasso with fixed features: 0.81, 0.87, 4
    - Forward selection: 0.77, 0.82, 4
    - SVM: 0.74, 0.89, 7
    - Lasso: 0.70, 0.90, 18

Prediction Results: Comparison with Other Feature Selection Methods
- Algorithm
  - Testing Accuracy
  - Training Accuracy
  - # of Selected Features
  - Comparison with Other Feature Selection Methods
    - MI + Lasso with fixed features: 0.81, 0.87, 4
    - Forward selection: 0.77, 0.82, 4
    - SVM: 0.74, 0.89, 7
    - Lasso: 0.70, 0.90, 18

(1) right rostral anterior cingulate
(2) left inferior parietal
(3) left pars orbitalis
(4) right frontal pole
Confirmation: An independent study

Right rostral anterior cingulate cortex (Right ACC): significant differences between ADHD and control groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>ADHD (n = 53)</th>
<th>Control (n = 53)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total brain volume (cm³)</td>
<td>135.2±14.37</td>
<td>129.9±13.23</td>
<td>0.023</td>
</tr>
<tr>
<td>Right ACC</td>
<td>3.13±1.25</td>
<td>3.43±1.73</td>
<td>0.113</td>
</tr>
<tr>
<td>Left ACC</td>
<td>2.98±2.98</td>
<td>3.65±2.98</td>
<td>0.036</td>
</tr>
<tr>
<td>Right ACC</td>
<td>3.31±3.15</td>
<td>3.40±2.87</td>
<td>0.023</td>
</tr>
<tr>
<td>Control ACC</td>
<td>3.03±2.98</td>
<td>2.98±2.88</td>
<td>0.608</td>
</tr>
</tbody>
</table>

Haxby’s Data: Experiments
- There are 10 runs (blocks), each producing 121 fMRI data points.
- Each block displayed image exemplars from all 8 conceptual categories: 1) face, 2) house, 3) cat, 4) bottle, 5) scissor, 6) shoe, 7) chair, and 8) ‘scrambled picture’.

Classification Results: Subject 1 – VTC (577 voxels)

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Validation Accuracy</th>
<th>Testing Accuracy</th>
<th>Avg # voxels selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stepwise Selection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI ranking + SVM</td>
<td>0.88</td>
<td>0.88</td>
<td>284.4</td>
</tr>
<tr>
<td>MI ranking + LR</td>
<td>0.94</td>
<td>0.90</td>
<td>359.7</td>
</tr>
<tr>
<td>MI ranking + GNB</td>
<td>0.75</td>
<td>0.75</td>
<td>265.9</td>
</tr>
<tr>
<td>MI ranking + SVM + Lasso</td>
<td>0.87</td>
<td>0.93</td>
<td>260.0</td>
</tr>
<tr>
<td>MI ranking + LR + Lasso</td>
<td>0.93</td>
<td>0.88</td>
<td>399.3</td>
</tr>
<tr>
<td>MI ranking + GNB</td>
<td>0.78</td>
<td>0.70</td>
<td>295.4</td>
</tr>
<tr>
<td>MI ranking + SVM + Ridge</td>
<td>0.86</td>
<td>0.85</td>
<td>8.25</td>
</tr>
<tr>
<td>MI ranking + LR + Ridge</td>
<td>0.82</td>
<td>0.78</td>
<td>14.95</td>
</tr>
<tr>
<td>MI ranking + Lasso</td>
<td>0.83</td>
<td>0.92</td>
<td>577 (over fitting)</td>
</tr>
</tbody>
</table>

A classic classification problem... but what is the problem?
- The ROI VTC contains a total of 577 voxels.
- The whole brain space covers a total of 43,193 voxels.

Classification Results: Subject 1 – VTC (577 voxels)

- Using an untrained Lasso parameter

Results:
- **Validation Accuracy**: Overall Accuracy on the training data.
- **Testing Accuracy**: Overall Accuracy on the test data.
- **Avg # voxels selected**: Average number of voxels selected for each classification algorithm.
### Classification Results:

**Subject 1 – Whole Brain (43,193 voxels)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Accuracy</th>
<th>Validation Accuracy</th>
<th>Testing Accuracy</th>
<th>Avg # voxels selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR + Lasso</td>
<td>0.99</td>
<td>0.73</td>
<td>0.76</td>
<td>13.25</td>
</tr>
<tr>
<td>LR + ridge</td>
<td>0.98</td>
<td>0.66</td>
<td>0.61</td>
<td>14.4167</td>
</tr>
<tr>
<td>LR + top M</td>
<td>0.99</td>
<td>0.44</td>
<td>0.40</td>
<td>43,193 (over fitting)</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>0.95</td>
<td>0.48</td>
<td></td>
<td>n/a</td>
</tr>
</tbody>
</table>

### Generalizing to Other Studies:

**Test Results**

<table>
<thead>
<tr>
<th>Task</th>
<th>reg. type</th>
<th>% train acc.</th>
<th>% valid acc.</th>
<th>% test acc.</th>
<th>average # of selected voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iowa</td>
<td>Lasso</td>
<td>99.52</td>
<td>85.19</td>
<td>83.85</td>
<td>68.00</td>
</tr>
<tr>
<td></td>
<td>elastic</td>
<td>100.00</td>
<td>89.45</td>
<td>84.04</td>
<td>258.00</td>
</tr>
<tr>
<td></td>
<td>ridge</td>
<td>100.00</td>
<td>85.32</td>
<td>85.28</td>
<td>2,045.00</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>100.00</td>
<td>58.59</td>
<td>59.31</td>
<td>2,045.00</td>
</tr>
<tr>
<td>CMU</td>
<td>Lasso</td>
<td>78.84</td>
<td>81.77</td>
<td>82.00</td>
<td>482.50</td>
</tr>
<tr>
<td></td>
<td>elastic</td>
<td>100.00</td>
<td>82.84</td>
<td>81.77</td>
<td>12,474.50</td>
</tr>
<tr>
<td></td>
<td>ridge</td>
<td>100.00</td>
<td>70.37</td>
<td>71.36</td>
<td>12,474.50</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>100.00</td>
<td>52.73</td>
<td>48.33</td>
<td>12,454.50</td>
</tr>
</tbody>
</table>

### Understanding the Alzheimer's Disease

- PET images
- Output of SI CE
- Better visualization

### Application in AD

- PET images:
  - Subjects: patient (AD, n=49); normal aging (NC, n=67); mild cognitive impairment (MCI, n=116)
- Imaging preprocessing:
  - AAL (Automated Anatomical Labeling)

### In Gaussian Systems: Conditional independence = Inverse Covariance

**Assume a multivariate Gaussian distribution for the functional data of all the brain regions (nodes)**

- Structure identification
- Inverse covariance matrix of the distribution $\theta$ is a symmetric matrix

### Result – Functional Connectivity Networks

- Connectivity network structure:
  - AD: significant loss of connectivity in temporal lobe
  - AD: significant increase of connectivity in frontal lobe compared with NC
- Connectivity strength:
  - AD: between-lobe connectivity is weaker than those within-lobe
  - AD: same regions in left and right hemisphere are connected much weaker than NC

The significances of those differences are tested by bootstrap hypothesis testing.
Result – Hippocampus Network

Hippocampus network for AD patient

Hippocampus network for normal aging

Significant loss of connectivity between hippocampus and temporal lobe

Result – Strength of Connections in AD

Visualization of the strength of the connections - a dendrogram presentation

Temporal_Sup_L

Cingulum_Post_R & L

All in temporal lobe

Result – Strength of Connections in MCI

Result – Strength of Connections in NC

Credits

- "Decision Model for Patient-Specific Motion Management in Radiation Therapy Planning"
- "Network Optimization of Functional Connectivity in Neuroimaging for Differential Diagnoses of Brain Diseases"
- "Computational Framework of Robust Intelligent System for Mental State Identification and Human Performance Prediction with Biofeedback"
- "IBIC: Integrated Brain Imaging Center for the University of Washington"
- "Continuous Assessment of Cognitive Load in Information Seeking"

Thank you