# Research Proposal Machine Learning to Help Reveal Male and Female Brain Differences

Tram Nghi Pham, UC Berkeley Mathematics
tram.pham18@berkeley.edu

***Abstract*—Machine learning (ML) is becoming increasingly important in a number of fields today. ML consists of statistical/mathematical tools (classified as supervised or unsupervised) used to build a model for predicting, or estimating outputs based on given inputs. Those kind of problems are usually encountered in many fields such as business, medicine, medical imagining, computer vision, etc. In this research, we will be applying ML to the field of functional brain networks. Our aim is to gain a better understanding of between-group differences in large-scale structural and functional brain networks. To be precise, we will be utilizing spectral clustering and several supervised models to analyze the volume and surface area of different regions in the brain. To be precise, we will be utilizing spectral clustering and several supervised models such as Canonical Correlation Analysis (CCA), Linear Discriminant Analysis(LDA), Support Vector Machine(SVM), etc. to analyze the volume and surface area of different regions in brain. Doing so will help us understand how gender can be differentiated based on brain regions and how brain regions are distributed and function based upon gender.**

## I. Introduction

Conventional clustering algorithms like k-means fails to detect the patterns when data is not differentiated by the pairwise distances between points. Therefore, we attempt to use a more advanced method, called spectral clustering - a graph-based algorithm that relies on the eigenstructure of a similarity matrix. The main goal is to see if it is possible to cluster individuals into two groups: male and female. If so, can we explore which brain regions contribute the most to each gender, and if not, can we distinguish any interesting patterns across different clusters? This step is considered as part of an exploratory data analysis. Our next objective is to perform supervised learning methods to extract the brain regions that contribute mainly in differentiating male and female groups. The problem can be rewritten similarly as extracting important features from a classification problem.

## II. Data Description

1) *Volume Data:* Dataset of size $201 \times 258$ where 201 rows represent the individuals and 258 columns represent the surface area of different brain regions. The data is labeled, in which, out of 201 individuals, 55 are female and the rest is male.
2) *Surface Area Data:* Dataset of size $213 \times 68$ where 213 rows represent the individuals and 68 columns represent the volume of different brain regions. The data is labeled, in which, out of 213 individuals, 64 are female and the rest is male.

## III. Research Outline

Research Status: This research is being conducted at Center for Interdisciplinary Brain Sciences Research (CIBSR), Stanford University as part of a supervised independent study course that counts four units toward university credit. This research is also performed in congruence with an honors thesis whose topic is Spectral Clustering and Image Segmentation.

*Goals of Project:* Three major questions will guide our research.

1) Cluster individuals into groups according to their brain regions (surface area, volume, etc) to explore the underlying structure, namely gender. Is there any interesting patterns across different clusters?
2) Which brain regions contribute the most to differentiating male and female groups?
3) Which brain regions are more active in males, and which brain regions are more active in females?

*Work so far and Challenging:*

- Additional features implemented to aid spectral clustering graphical user interface (GUI)an extension of original GUI implemented for the honor thesisin order to see the difference (on an average) between groups in terms of developmental quotient (DQ), etc. See appendix for a demo of GUI.
- CCA has been successfully separating male and female in the projected space. See appendix for data projections.
- Supervised learning algorithms have been conducted and the current accuracy of 72% is obtained.

*Challenging*

- Number of features are much larger than number of observations, causing the algorithm to suffer from over-fitting.
- Principal Component Analysis (PCA) is useless in this dataset(it barely separates male and female groups and eigenvalues are decaying relatively slow)
- Linear Discriminant Analysis (LDA) and CCA are prone to over-fitting, leading to a low accuracy.
- The amount of data is limited.

*Proposed methods:*

Advisor: Prof. Hadi Hossenei, Stanford Universeity

1) Data preprocessing can often have a significant impact on generalization performance of a supervised ML algorithm. In our case, we decided to normalize the data since all features take on continuous values and we are treating each feature equally.
2) Running Spectral clustering to find interesting patterns across clusters. This step serves as an exploratory data analysis. The challenge in spectral clustering is choosing the kernel parameter $\sigma$, which takes a lot of time and effort to tune.
3) Project data onto k-dimensional subspace using CCA. Given an input state $x \in R^d$ represents for $d$ brain regions/features of an individual $x$, and denote the label as $y \in \{0, 1\}$. Stack them in a matrix. In order to perform CCA, we must first turn our labels y into a one-hot encoding vector $y' \in \{0, 1\}^J$, where each element corresponds to the label. We then need to compute the canonical correlation matrix $\Sigma_{XX}^{\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{\frac{1}{2}}$ and singular value decomposition $U\lambda V^T$. The projection to the canonical variates is defined as
$$x' = U_k^T \Sigma_{XX}^{\frac{1}{2}} x$$
where $k$ is desired dimension of projected subspace, and $\Sigma_{XX}$ is covariance matrix of matrix $X$
4) Project data using LDA dimension reduction technique. It will help us to compare between CCA and LDA's efficacy.
5) Build various classification models to quantify the importance of each brain region. In particular, LDA, Quadratic Discriminant Analysis (QDA), SVM, Ridge Regression, etc will be performed.
6) *Sparse regression* plays an important role in selecting features, preventing over-fitting, reducing computational time, and increasing accuracy. We will focus on Sparse Regression to manage the dataset which has more features than observations.
7) Even though CCA and PCA can reduce dimension, the principal "features" are actually combinations of the features we really collected for our problem, in order to interpret the effect of the original features, we'd need to do a sparse variant of this method, called *Sparse Principal Component Analysis (SPCA)* [3]. A particular disadvantage of ordinary PCA is that the principal components are usually linear combinations of all input variables. Sparse PCA overcomes this disadvantage by finding linear combinations that contain just a few input variables. Running Sparse PCA, Sparse CCA, Sparse Ridge Regression to do feature selection will be proceeded next if the accuracy is still low.
8) Neural network is a promising method though it will be at the risk of memorizing the labels given that we only have a small dataset.
file://./path-to-file

## IV. Appendix

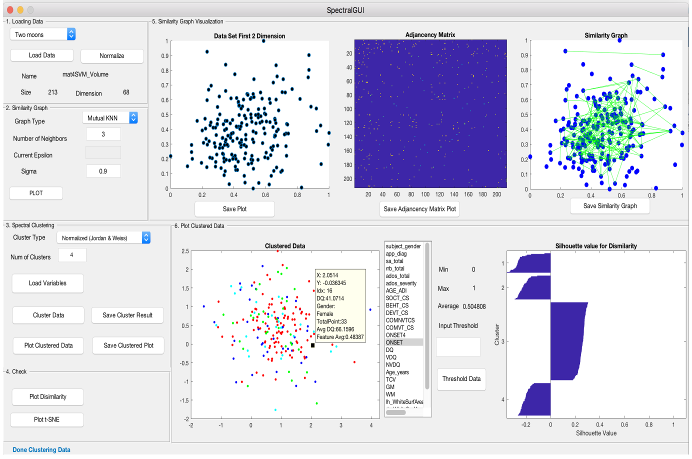
**Spectral Clustering Algorithm** *Normalized Spectral Clustering according to Ng, Jordan, and Weiss (2002)*

**Input:** Given a set of points $S = s_1, ..., s_n$ to cluster into $k$ groups

1) Form the similarity graph with adjacency matrix $W \in R^{n \times n}$ by one of the ways described in section 2.
2) Compute the unnormalized graph Laplacian $L = D - W$
3) Compute the first $k$ eigenvectors $u_1, u_2, ...u_k$ of the graph Laplacian $L_{sym}$
4) Form a matrix $U \in R^{n \times k}$ containing the vectors $u_1, u_2, ...u_k$ as columns.
5) Form the matrix $T \in R^{n \times k}$ from $U$ by normalizing the rows, i.e, set
$$T = (t_{ij})_{i,j=1}^n \; with \; t_{ij} = \frac{w_{ij}}{\left(\sum_k u_{jk}^2\right)^{\frac{1}{2}}}$$
6) For $i = 1, ..., n$, let $y_i \in R^k$ be the vector corresponding to the $i - th$ row of matrix $U$.
7) For $i = 1, ..., n$, let $y_i \in R^k$ be the vector corresponding to the $i$-th row of matrix $T$.
8) Cluster the points $(y_i)_{i=1}^n$ into clusters $C_1, ...C_k$ using the k-means algorithm

**Output:** Clusters $A_1, ..., A_k$ with $A_i = \{j | y_j \in C_i\}$

### Spectral Clustering Graphical User Interface



SpectralGUI

1. Loading Data
Two moons
Name mat4SVM_Volume
Size 213 Dimension 68

2. Similarity Graph
Graph Type: Mutual KNN
Number of Neighbors: 3
Current Epsilon:
Sigma: 0.9

3. Spectral Clustering
Cluster Type: Normalized (Jordan & Weiss)
Num of Clusters: 4

4. Check

X: 2.0514
Y: -0.036345
Idx: 16
DQ:41.0714
Gender:
Female
TotalPoint:33
Avg DQ:66.1596
Feature Avg:0.48387

Min 0
Max 1
Average 0.504808
Input Threshold

Done Clustering Data

### CCA Projections

Figure 1 shows the data after being projected on 2 dimensional subspace.

## References

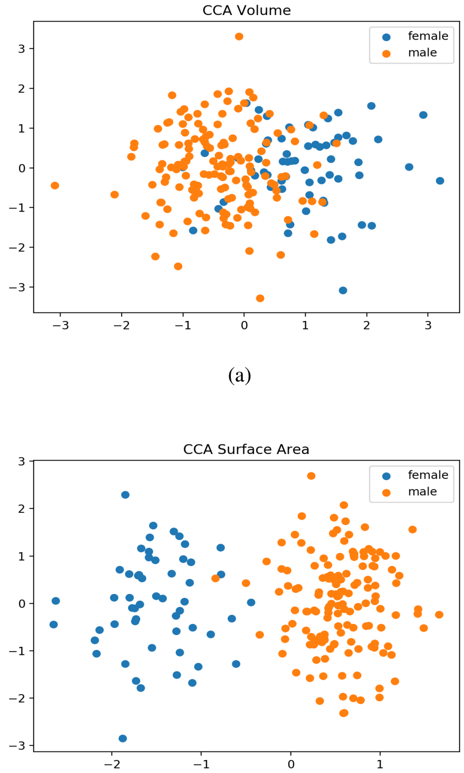
[1] U. von luxburg, *A tutorial on spectral clustering, 2007.* https://www.cs.cmu.edu/~aarti/Class/10701/readings/Luxburg06_TR.pdf.

[2] Lihi Zelnik-Manor, Pietro Perona, *Self-Tunning Spectral Clustering* http://www.vision.caltech.edu/lihi/Publications/SelfTuningClustering.pdf,AccessedSept18,2017

[3] Hui ZOU, Trevor HASTIE, and Robert TIBSHIRANI, *Sparse Principal Component Analysis. American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America Journal of Computational and Graphical Statistics, Volume 15, Number 2, Pages 265286, 2006.*

[4] Olcay Kursun, Ethem Alpaydin, and Oleg V Favorov, *Canonical correlation analysis using within-class coupling, Pattern Recognition Letters, vol. 32, no. 2, pp. 134144, 2011.*

[5] M. A.Ng and Y.Weiss, On spectral clustering: Analysis and an algorithm., 2011. http://ai.stanford.edu/~ang/papers/nips01-spectral.pdf, Accessed 27 June 2017.



(a)

(b)

Fig. 1: Projection on 2-dimensional subspace using CCA on (a) Volume Data and (b) Surface Area Data. Male and female are well-separated in the projected space.

[6] Heysem Kaya, Florian Eyben, Albert Ali Salah, Bjorn Schuller, *CCA BASED FEATURE SELECTION WITH APPLICATION TO CONTINUOUS DEPRESSION RECOGNITION FROM ACOUSTIC SPEECH FEATURES* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.432.4086&rep=rep1&type=pdf, Accessed 19 Oct 2017