ISyE 6740 – Spring 2021
Project Proposal

Team Member Name: Daniel Bae

Project Title: Applications of SVM Beyond Classification Demonstrated via Clinical Data on Dementia

Background:

Dementia is a general term describing a wide group of symptoms related to cognitive impairment and/or cognitive decline. Early detection and planning are crucial as in some cases there exists methods of prevention, treatment, and even reversal of symptoms. There has been success in using analytics for preclinical detection of cognitive change using the popular "clock drawing test", which tracks the movement of a digital pen as participants draw a clock and analyzes the data to determine the risk of cognitive change. This method has made great strides in systematizing a technique for early detection. Since the process is digital, there is potential for it to be administered remotely, it captures minute details imperceptible to the human eye, and data can be collected from participants over repeated trials among other advantages over traditional methods of diagnosis. Advances in the set of analytics tools available in the effort to minimize the impacts of these diseases have significant potential.

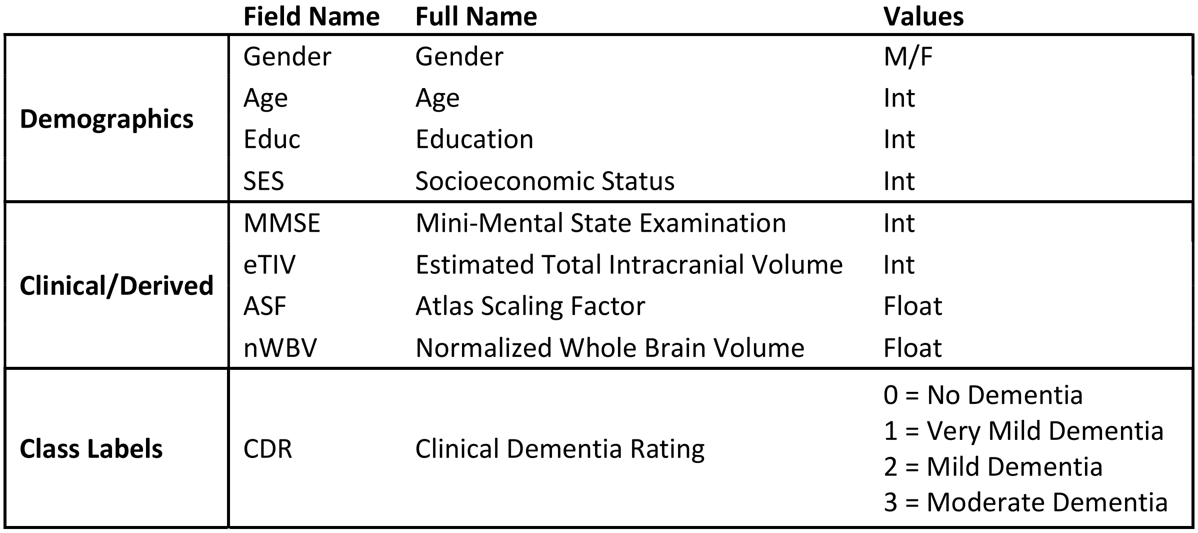
Problem Statement:

This project proposes the application of kernel support-vector machines on data collected from participants that exhibit various levels of dementia in order to discover insights not limited to classification.

Support-vector machines (SVM) are a relatively robust and efficient linear classification tool. With the use of kernels, such as the popular radial basis function (rbf), SVM can also perform non-linear classification. By attempting to maximize the distance between classes, SVM focus on the points nearest or on the boundaries between the classes. These points are known as the support vectors. These support vectors are the "relevant" points that are used to model the boundaries. All other points are, typically, not required by the model and their weights are zero ergo SVM are often described as being memory efficient. A component of this project is to build upon the existing applications of and research in SVM for purposes not limited to classification by using SVM to understand more about the geometry of high dimensional data and observing relationships between features. Specifically, leverage the ratio of support vectors to non-support vectors to determine the relative "roughness" of data and observe change in distance from support vectors given a small change in the value of a feature of a non-support vector to ascertain relative feature importance and relationships between features. By applying these techniques to the data collected from participants demonstrating various levels of dementia, extract information in addition to classification.

Data Source:

The dataset leveraged to test the utility of SVM, in addition to classification, is the "OASIS-1: Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults" dataset from the "Open Access Series of Imaging Studies" (OASIS) project. Demographic, clinical, and derived data was collected from 436 participants. There are 11 features in total, however, not all of the features will be included in this experiment. For example, all of the participants were right-handed therefore the feature on dominant handedness will be excluded. The dataset also exhibits missing data for roughly half of the participants. Participants with missing data will not be included in this experiment. The total final population will likely be closer to 216 participants each with 8 features. Below is a figure representing the 8 features and the class labels.



	Field Name	**Full Name**	**Values**
Demographics	Gender	Gender	M/F
	Age	Age	Int
	Educ	Education	Int
	SES	Socioeconomic Status	Int
Clinical/Derived	MMSE	Mini-Mental State Examination	Int
	eTIV	Estimated Total Intracranial Volume	Int
	ASF	Atlas Scaling Factor	Float
	nWBV	Normalized Whole Brain Volume	Float
Class Labels	CDR	Clinical Dementia Rating	0 = No Dementia 1 = Very Mild Dementia 2 = Mild Dementia 3 = Moderate Dementia

Methodology:

First and foremost is the preprocessing of the data after only relevant features and participants without missing data have been retained. Using one-class SVM, check for the existence of outliers. Outliers could have large influences on the smaller dataset being used for these experiments. Once outliers have been taken care of, if they do in fact exist, each of the features will be standardized by subtracting the mean and dividing by the standard deviation. The data will then be split into a training set comprising 70% of the data with the remaining 30% being reserved for testing. Since the dataset is relatively small, proper randomization during the shuffling process will be critical.

After the data has been preprocessed, a SVM model using the rbf kernel can be trained on the training data using cross-validation to tune the model's parameters. For the model's gamma parameter, the heuristic "median trick" will be employed to choose a suitable value. The model with the best average performance, based on accuracy, during cross-validation will then be tested on the reserved testing dataset in order to confirm whether, given a set of features, it can reasonably accurately classify a participant into one of four classes.

Given the trained model, several observations will be made based on the relation between support vectors and non-support vectors. Data that will be collected includes, the ratio of support vectors to non-support vectors and the amount of “contact” between support vectors of different classes. Non-support vectors will also be randomly selected and given a small change in one of their features, their distance to the nearest support vector from each class will be measured. The distance to the nearest support vector from each class will be determined using a k-nearest neighbors algorithm to identify the nearest vectors and L2 norm to measure the distance. This will be repeated using combinations of features to observe whether there are any relationships.

Evaluation and Final Results:

Since a rough object is likely to have a greater surface area to volume ratio, a greater support vectors to non-support vectors ratio should also give some sense of the geometry of the data even in higher dimensions. By observing relative ratios compared to the plot of their 2D representations achieved using PCA and ISOMAP, it may be possible to learn about the irregularity, or absence of, in the data without having to reduce the number of dimensions to a number that can be visualized. While PCA and ISOMAP are popular methods for linear and non-linear dimensionality reduction respectively, some information may be lost in the process.
The magnitude of change in Euclidean distance between non-support vectors and support vectors given changes in their features will be compared to the results obtained using the mutual information statistic. Given that mutual information is able to describe any type of dependency, if the results follow similar patterns, it may be possible to learn a lot about the features given even just the SVM model itself.

Conclusion:

There is often no single best method to achieve a goal in every analytics scenario. Having multiple methods to tackle a problem can improve the amount learned overall. SVM as classification tools can bolster the effort to detect cognitive decline early and accurately. Furthermore, by observing the relationship between the support vectors and the unweighted datapoints it may be possible to learn even more about the underlying data.

Works Cited:

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

OASIS brains. (n.d.).
Retrieved from https://www.oasis-brains.org/#data

Hyunseokc. (2018, March 06). DETECTING EARLY ALZHEIMER'S.
Retrieved from https://www.kaggle.com/hyunseokc/detecting-early-alzheimer-s

Informs. (n.d.). Detecting Preclinical Cognitive Change.
Retrieved from https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Detecting-Preclinical-Cognitive-Change

Bron, E. E., Smits, M., Niessen, W. J., & Klein, S. (2015). Feature Selection Based on the SVM Weight Vector for Classification of Dementia. *IEEE journal of biomedical and health informatics*, *19*(5), 1617–1626. https://doi.org/10.1109/JBHI.2015.2432832