CS4246 Project 1
Depression Prediction

Team 01

Antoine Charles Vincent Garcia - A0159072A
Chan Jun Wei - A0112084U
Chen Tze Cheng - A0112092W
Eric Ewe Yow Choong - A0112204E
Han Liang Wee, Eric - A0065517A
Ho Wei Li - A0094679H

Abstract

Depression is a debiliatating mental illness that has good prognosis given early detection and treatment. However, detection is difficult with the various factors that raise the entry barrier and decrease the accuracy of a diagnostic test. In this study we aim to prove that prediction of depression severity on the PHQ-8 scale with voice acoustics is possible through the use of Gaussian Process, potentially lowering entry barriers to diagnostic tests and aiding early detection.

Introduction

Depression has a severe, and at times long-term, negative impact on an individual's quality of life. Major depression is 3rd leading cause of disability worldwide with 65 million life years spent living with the disability or lost due to early death [World Health Organization, 2004]. Depression's annual toll on U.S businesses amounts to about $80 billion in medical expenditures, lost productivity and suicide. Among the costs, close to $10 billion accrued in lost workdays each year and more than $33 billion in other costs accrue from decreased productivity due to symptoms that sap energy, affect work habits, cause problems with concentration, memory, and decision-making. [Greenberg et al., 2015]

Left unchecked, depression increases risk for morbidity, suicide, decreased cognitive and social functioning, self-neglect, and early death [Fiske, Wetherell, and Gatz, 2009]. Death from suicide is one of the top 10 causes of death, above the death rate for chronic liver disease, Alzheimers, homicide, arteriosclerosis or hypertension. [Jiaquan Xu et al., 2016]

Despite the severe consequences, depression is one of the most treatable mental illnesses but it is also one of the most under-diagnosed globally. In general health-care, 48.4% of patients suffering from depression go unrecognized [RC et al., 2003].

Motivation and Hypothesis

The Personal Health Questionaire depression scale (PHQ-8) is one of the many diagnostic tests for depressive disorders and it has been proven to be an effective severity measure for depression in large clinical studies[Kroenke et al., 2008]. Nevertheless, one of the biggest obstacles to successful diagnosis of depression is the unwillingness of patients to admit their predisposition to depression by seeking help.

People often subscribe to the social stigma that being depressed reflects a weakness in their character, a permanent defect in their personality. This stigma manifests itself particularly in a phenomenon known as social distancing whereby people with mental issues are more isolated from others [Smith and Cashwell, 2011]. People suffering from depression hence tend to be ashamed of their condition and are generally convinced that denying and hiding it from others gives them a better shot at integrating with society and living a normal life [Wolpert, 2001]. Even if they do seek help, the accuracy of the PHQ-8 or just questionaires and surveys in general are often adversely affected by the Hawthorne Effect, a type of reactivity in which individuals modify or improve an aspect of their behavior in response to their awareness of being observed. [McCambridge, Witton, and Elbourne, 2014].

In the past decade, there have been research successfully correlating emotion with voice production and speech acoustics [Johnstone, 2001]. Corollary to that, active research of late into the use of voice acoustics as predictors of clinical depression scores has seen some success, proving that it is an effective indicator of depression severity [Hashim et al., 2016].

The aim of the present study is to determine if acoustic measures of voice, characterizing specific spectral and timing proerties, predict clinical ratings of depression severity in a sample of patients using an existing dataset of their voice recordings, on the PHQ-8 scale.

Future Applications

Findings from this experiment will reinforce results from similar experiments and make a stronger case for the

possibility of lowering the entry barriers of depression diagnostic tests and at the same time, increasing the accuracy of depression diagnosis by abstracting the Hawthorne Effect. Diagnostic tests can be designed in a friendly way that does not require the patient to consciously answer questionaires such as the PHQ-8 or even to have complete awareness of the diagnostic process. By taking into account the Hawthorne Effect and the status quo of the social stigma and working around them, diagnoses can be made easier, faster and more accurately and with that, education and treatment can begin early before the mental condition of the patient deteriorates.

Gaussian Process Regression Model

As all individuals have varying inherent stress management, the use of the GP model for depression prediction makes use of all samples and feature information to perform the prediction including training data with different or uneven sampling rates. From the mean and variance obtained from previous data, we are able to predict if an individual is depressed.

Qualitative Advantages

GP is exceptionally useful in this application as it enables us to perform prediction on various training data including outliers or those with irregular sampling rates. As our experiment is based on individuals and their personality, the data is therefore subjective for each individual. This helps in our attempt to verify our hypothesis of the correlation between human behaviour and depression.

One key advantage of GP is that it is able to give the confidence interval. Curve-fitting algorithms for interpolation of data do not accomodate well to noise, unlike GP as it can be optimized with the appropriate hyper-parameters and thus, allows a fine and precise trade-off between smoothing and fitting the data.

With GP, we can also obtain a probability distribution such as the predicitve mean and variance. These properties have a significant role to play in estimating the depression levels. For instance, the mean could be use as the baseline for an average working-human mood. The variance here would give us a picture to how depressed is an individual or vice versa. These properties allows us to gauge the depression 'levels' as well as measure the accuracy of the correlation between a subject's emotion and how depress they are feeling.

Requirements

Aside from the audio recordings, our GP model makes use of the following:

1. **Audio Features**
   What are the differences between sounds? To differentiate between sounds, we need to learn about audio features, which will be described below:

   (a) **Energy**
   The Energy feature of a sound refers to the loudness of the sound at various timeframes, hence it is obvious that energy of a sound is directly proportional to the amplitude of the soundwave. This shows that the higher the energy, the louder the sound is going to be.

   (b) **Mel Frequency Cepstral Coefficients (MFCC)**
   Mel Frequency Cepstral (MFC) is a representation of short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The greatest benefit of using MFCC is that the scale approximates the human's auditory system response more closely, hence it allows for a better representation of sound. In order to obtain MFCC, Fourier transform is performed on the sound signals.

   (c) **Magnitude Spectrum**
   Magnitude spectrum can be produced by converting the input signal of an audio into frames. Fast Fourier Transform (FFT) is performed on each frames and this will form the Magnitude Spectrum.

   (d) **Zero-Crossing Rate**
   Zero crossing rate is the rate of sign-changes along a signal. This feature is extremely useful in speech recognition and music information retrieval.

2. **Kernels**
   The similarity measure of all features is usually called kernel. As Gaussian Process uses Kernels to predict the value for an unseen point from training data, understanding different kernels that we are going to use is extremely useful.
   The kernels that would we are intested in is shown below (The kernels are assumed to be defined on two samples $x = (x_1 x_2 x_3 ... x_n)$ and $x' = ({x_1}' {x_2}' {x_3}' ... {x_n}')$, represented as feature vectors in some input space):

   (a) **Radial Basis Function (RBF)**
   The RBF kernel is defined as

   $$K(x, x') = \exp^{-\frac{||x-x'||^2}{2\sigma^2}} \tag{1}$$

   $||x - x'||^2$ represents Squared Euclidean Distance And $\sigma > 0$ can either be a scalar (isotropic variant of the kernel) or a vector with the same number of dimensions as the inputs X (anisotropic variant of the kernel).
   It is also known as the squared exponential kernel. This kernel is infinitely differentiable, which implies that GPs with this kernel as covariance function have mean square derivatives of all orders, and are thus very smooth.

   (b) **Matern**
   The class of Matern kernels is a generalization of

the RBF and the absolute exponential kernel parameterized by an additional parameter nu. The smaller the value of nu, the less smooth the approximated function is. For nu=∞, the kernel becomes equivalent to the RBF kernel and for nu=0.5 to the absolute exponential kernel. Important intermediate values are nu=1.5 (once differentiable functions) and nu=2.5 (twice differentiable functions).

(c) **Automatic Relevance Detection (ARD) Kernels**
If we chose the anisotropic variant of the RBF and Matern kernel, then these kernels would be also called as ARD kernels. So, what it means for the model is, (let's assume the numbe of dimensions of the input is N) for each σ_i, where i = 1,2,3,...,N, might be different, depending the importance of the corresponding input features.

There are a few advantages for our project if we used ARD: [Cawley and Talbot, 2014]

i. Generalisation performance is potentially improved.
ii. The data would be better explained.
iii. Increase the efficiency of feature extraction for audio.

(d) **Dot Product**
The Dot Product kernel is non-stationary and can be obtained from linear regression by putting N(0, 1) priors on the coefficients of x_d (d = 1, . . . , D) and a prior of N(0, σ_0^2) on the bias. The Dot Product kernel is invariant to a rotation of the coordinates about the origin, but not translations. It is parameterized by a parameter σ_0^2. For $\sigma_0^2 = 0$, the kernel is called the homogeneous linear kernel, otherwise it is inhomogeneous. The kernel is given by

$$K(x, x') = \sigma_0^2 + x \cdot x' \tag{2}$$

While Matern is a very generalized kernel, both isotropic and ARD variant of the Matern kernel is selected for our model. On the other hand, as dot product kernel is a very efficient and simple kernel, we decide to give it a try too.
As we favor simplicity, we actually think that dot product kernel would give us the best result.

2.3 Important Machine Learning Methods

Machine learning is about using models to learn from existing data for some improvement or predictions. The following is some methods used in machine learning:

1. **k-nearest neighbors algorithm**
The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. K is the user defined constant which classify the class of a vector data. The training examples are vectors in a multidimensional feature space, each with a class label.

2. **Support Vector Machines (SVM)**
Support Vector Machines (SVM), with the help of the libsvm, a developed library for SVM. SVM is a type of linear classifier that aims at finding a unique solution in the form of an optimal hyperplane. This optimal hyperplane is defined as the one that maximizes the margins between the two classes. It will be positioned at equal distance between the closest points of each class and will create the largest possible corridor with no points from either classes inside. These closest points on the border of the corridor are called the support vectors.

3. **Random Forest Regressor**
Random Forest is an ensemble of decision trees, whereas a decision tree is trying to separate the data into different leaf nodes where the data points in each node have certain similarity. Each decision tree will predict a value and all of the values will be averaged. The average value would be the prediction result.

4. **AdaBoost**
AdaBoost is the short form of Adaptive Boost. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost first determines the classifier with the least errors, then focuses on those errors later by adding a weight to the outliers, subsequently it searches for another classifier with least errors again. The process is repeated for all classifiers which will lastly be combined to form a final classifier which will specify the class of data. With this process, AdaBoost has the capacity to be sensitive to noisy data and outliers.

5. **Naive Bayes**
Naive Bayes is a prediction algorithm that uses Bayes rule where the value of a particular feature is independent of the value of any other feature given the class variable. Naive Bayes takes the product of all the probabilities of a particular event, then gives a prediction of the probabilities of that event. All the data are extracted from the dataset to represent the probability which are the number of times the particular event occurs in the dataset out of the total number of events. Naive Bayes can produce great results if the features in the dataset is as close to being independent as possible.

Feature Selection

Feature selection is the process of selecting a subset of relevant features including variables or predictors to be used in a model construction for machine learning. The usage of feature selection is to reduce the complexity of a model in machine learning to be interpreted easier. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and most cost-effective predictors, and providing a better understanding of the underlying process that generated the data. [Guyon and Elisseeff, 2003]

In our audio dataset, we extracted many features including MFCC, MS, Zero-Crossing rate and Energy level, and thus the dimension of our features are huge (about 100 columns). While the number of datas we have are not large enough, having a large feature dimensions will lead to overfitting. As shown in the Figure 1, when the model is overfitting, it will produce a worse result.

Besides, having less features would significantly reduce the time amount of training the model.

Based on these rationale, we decided to do feature selection.

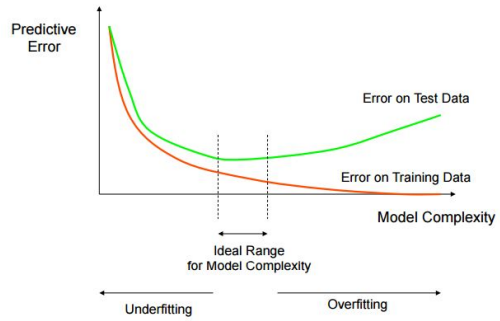

Figure 1: How Overfitting affects Prediction

Hypothesis: Doing feature selection before applying to the ARD Kernel Gaussian Process model will improve the result. Without doing feature selections, as the feature dimensions would be huge, we would also have large number of kernel parameters. As a result, it is likely to lead to overfitting and generating poor result. [Cawley and Talbot, 2014] Thus, to increase the effectiveness of the Gaussian Process Model, we would apply feature selection to the data first.

Hypothesis: MFCC is the most important feature. Depression is actually tied to emotions : it can either mean sadness or people who deny feeling sad. [Stratou et al., 2015] On the other hand, as the baseline accuracy of using only MFCC to predict emotions is quite high, [El Ayadi, Kamel, and Karray, 2011], we actually guess that MFCC is the most important feature in depression prediction. In other words, since we have limited datas and large feature dimension, using MFCC feature only for the input vector of the model might actually improve the result.

Evaluation

In order to test our Gaussian Process model, we conducted tests on data obtained from Audio/Visual Emotion Challenge and Workshop(AVEC 2016) [Valstar et al., 2016]. The goal of AEVC is to weigh-in on the various approaches(visual, audio) used to recognize emotions under unambiguous conditions. AVEC 2016 provided 2 pieces of data as input: visual and auditory data. However, we would be reducing the scope of the experiment, limiting the experiment to only the auditory data. Two Sub-Challenges are lised in AVEC 2016. We are only interested in the Depression Classification Sub-Challenge, which requires participants to classify inputs by the PHQ-8 score.

Data

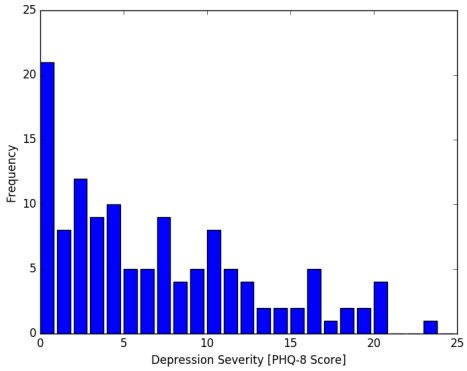

Figure 2: PHQ-8 scores' histogram of both training and development set

The depression data used in AVEC 2016 was obtained from the benchmarking database, the Distress Analysis Interview Corpus - Wizard of Oz(DAIC-WOZ). Data collected from DAIC-WOZ include audio and video recordings and the corresponsing PHQ-8 score[CITE:27](0-24), which is a frequently used self-report scheme to access severity of depression[CITE]. Henceforth, we would need to pre-process the auditory data before we use it in our Gaussian Process Model. The data is pre-processed as described in the Section [REF]. The distribution of the depression severity scores in both training and development set is given in Figure 2. The data provided are split into 2 sets: training and development. A summary of the data is given in Table 1.

	Training	Development	All
n	95	31	126
μ	6.326	7.548	6.626
σ	5.597	6.690	5.909

Table 1: Summary of Datasets provided

Measure of Accuracy

AVEC 2016 provided a baseline classifier that consistently predicts the PHQ-8 score with RMSE $= 6.7418$[CITE]. In order to provide a meaningful and consistent comparison to the baseline provided, we would be only using Root Mean Square Deviation Error(RMSE) to measure the error rate on both Training and Development datasets. RMSE(Equation 3) is a commonly used in machine learning communities to measure the differences between the values predicted by a

model and the values actually observed.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \quad (3)$$

[CITEDBLP:journals/corr/ValstarGSRLTSSC16]

Experimental Setup

We compared our Gaussian Model against commonly used machine learning algorithms. The list of algorithms and their hyperparameters are given in Table 2. The hyper-parameters are either determined by the defaults used in the popular machine learning library, Scikit Learn[CITE] or some reasonable values were used. Each machine learning algorithm is trained against the training set and thereafter tested against the development set using RMSE as the error metric. The process used is shown in Figure 3.

Algorithm	Hyper-parameters
K-Nearest Neighbors	x
Linear SVM	x
RBF SVM	x
Decision Tree	x
Random Forest	x
AdaBoost	x
Naive Bayes	x
Decision Tree	x

Table 2: List of Machine Learning Algorithms with their corresponding hyper-parameters

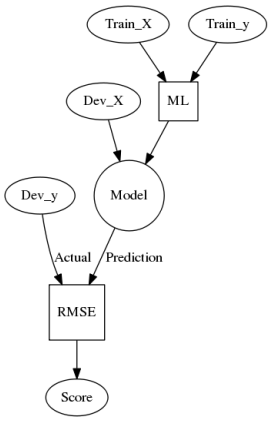

Figure 3: Experimental process

Results

The results of the experiment is shown in t

	RMSE	
Algorithm	Training	Development
K-Nearest Neighbors	x	x
Linear SVM	x	x
RBF SVM	x	x
Decision Tree	x	x
Random Forest	x	x
AdaBoost	x	x
Naive Bayes	x	x
Decision Tree	x	x
Gaussian Process	x	x

Table 3: RMSE results of the different machine learning algorithms

Conclusion

Our work has succesfully modelled the various levels of depression with GP by taking audio recordings as our input.

Contributions

- **Antoine Charles Vincent Garcia**: Scripting the program, setting up machine learning libraries and running tests.
- **Chan Jun Wei**: Project technicalities such as problem formulation and modelling, mathematics and experiment planning.
- **Chen Tze Cheng**: Project technicalities such as problem formulation and modelling, mathematics and experiment planning.
- **Eric Ewe Yow Choong**: Documentation especially writing of the motivation, recording research findings and keeping track of requirements.
- **Han Liang Wee, Eric**: Scripting the program, setting up machine learning libraries and running tests.
- **Ho Wei Li**: Documentation especially writing up the motivation, recording research findings and keeping track of requirements.

References

Cawley, G. C., and Talbot, N. L. C. 2014. Kernel learning at the first level of inference. *Neural Networks* 53:69–80.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44(3):572–587.

Fiske, A.; Wetherell, J. L.; and Gatz, M. 2009. Depression in older adults. *Annual Review of Clinical Psychology* 5:363–389.

Greenberg, P. E.; Fournier, A.-A.; Sisitsky, T.; Pike, C. T.; and Kessler, R. C. 2015. The economic burden of adults with major depressive disorder in the united states. *The Journal of Clinical Psychiatry* 76(2):155–162.

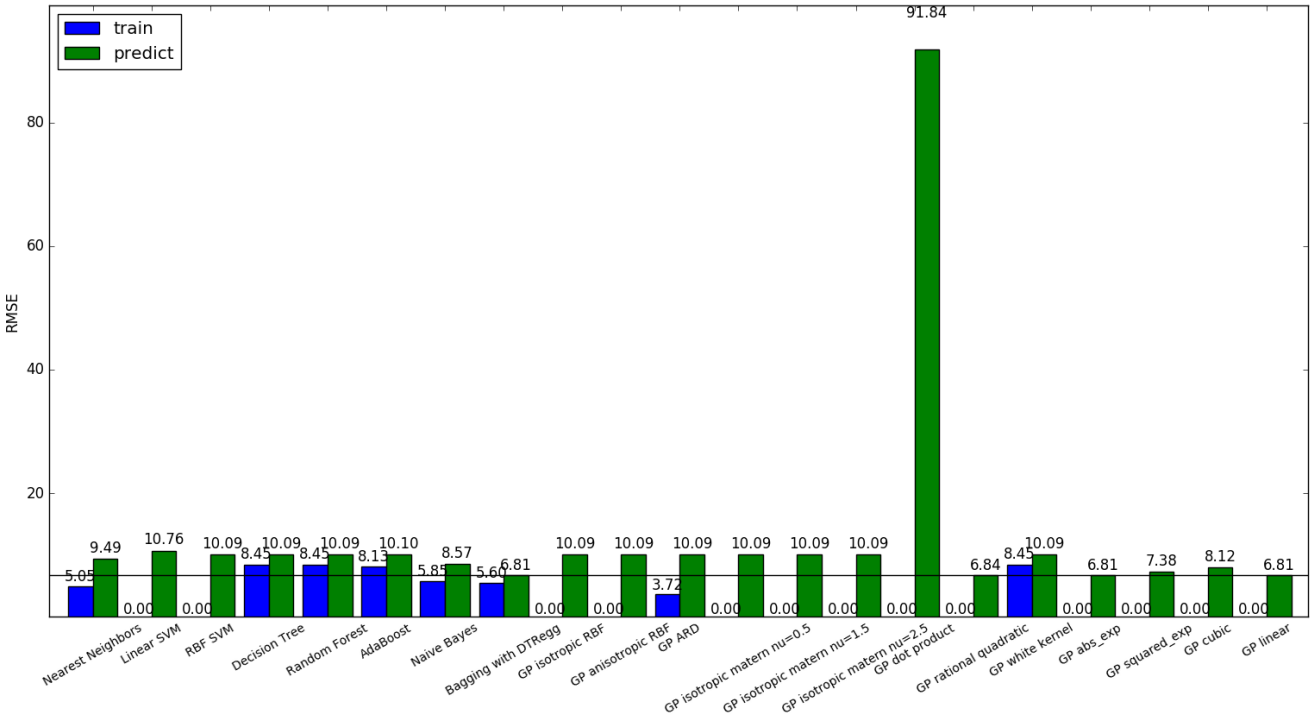

Figure 4: Chart showing RMSE(Training and Development) for the different classifiers

Guyon, I., and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)* 3(3):1157–1182.

Hashim, N. W.; Wilkes, M.; Salomon, R.; Meggs, J.; and France, D. J. 2016. Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice* 0(0).

Jiaquan Xu, M.; Sheery L. Murphy, B.; Kenneth D. Kochanek, M.; and Brigham A. Bastian, B. 2016. Deaths: Final data for 2013. *National Vital Statistics Reports* Vol 64, No. 2.

Johnstone, T. 2001. *The effect of emotion on voice production and speech acoustics*. Ph.D. Dissertation, University of Western Australia.

Kroenke, K.; Strine, T.; Spitzer, R. L.; Williams, J. B.; Berry, J. T.; and Mokdad, A. 2008. The phq-8 as a measure of current depression in the general popoulation. *Journal of Affective Disorders* 114(1-3):163–73.

McCambridge, J.; Witton, J.; and Elbourne, D. R. 2014. Systematic review of the hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67(3):267–277.

RC, K.; P, B.; O, D.; and et al. 2003. The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). *JAMA* 289(23):3095–3105.

Smith, A. L., and Cashwell, C. S. 2011. Social distance and mental illness: Attitudes among mental health and nonmental health professionals and trainees. *The Professional Counselor: Research and Practice* 1(1):13–20.

Stratou, G.; Scherer, S.; Gratch, J.; and Morency, L. P. 2015. Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender. *Journal on Multimodal User Interfaces* 9(1):17–29.

Valstar, M. F.; Gratch, J.; Schuller, B. W.; Ringeval, F.; Lalanne, D.; Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; and Pantic, M. 2016. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR* abs/1605.01600.

Wolpert, L. 2001. Stigma of depression - a personal view. *British Medical Bulletin* 57(1):221–224.

World Health Organization. 2004. The global burden of disease: 2004 update. Technical report, World Health Organization.