DS 6559 Bayesian Machine Learning
Final Project Report
Shaoran Li (sl4bz), Wenxi Zhao (wz8nx), Fang You (fy6vj)

Problem Description

Suicide is a tragic event with strong emotional repercussions for survivors and for families of its victims. It is a major public health problem and a leading cause of death around the globe. Many organizations such as United Nations Development Program, World Bank, World Health Organization are making various effort to prevent suicide from people's last-minute decisions as well as from the root causes. These organizations also collect data to help researchers study potential variables that contribute to people's suicide rates. Our project looks at the suicide data compiled by these organizations and aims to identify and alleviate causes that contribute to suicide rates across the population.

Our project chooses the Suicide Rates Overview from 1985 to 2016 in Kaggle, which is a real world dataset, to apply Bayesian machine learning analysis. This dataset is built to find signals correlated to increasing suicide rates among different groups over the world and across the socio-economic spectrum. This dataset contains 11 features: country, year, sex, age, count of suicides, population, suicide rate, human development index, GDP per year, GDP per capita, and generation. In this dataset, both age and generation have six categories. Our primary goal is to apply different Bayesian machine learning methods to predict the suicide rate, and to evaluate the performance of all Bayesian models.

Before applying Bayesian methods on our data, we start with exploratory data analysis and preprocessing steps. The Fig.1 shows the distribution of number of suicides per year. The graph clearly shows that the suicides number increases from 1988 to 1999, and then decreases a little bit. The suicide number in 2016 is very low, which is due to the data incompleteness.

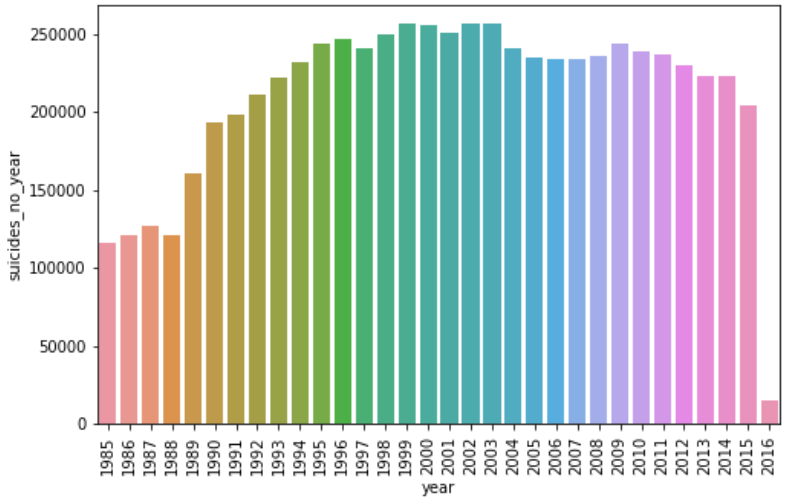

Fig.1 The distribution of no. of suicides per year

The Fig.2 shows the correlation map of all variables. The graph shows that the population and suicide number are highly correlated, since the suicide number is calculated by dividing the suicide number by population. Based on this high correlation, we drop the population variable.

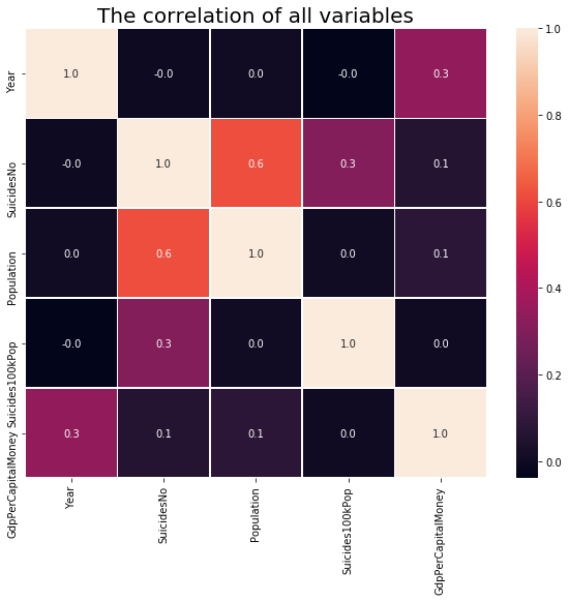

Fig.2 The correlation of all variables

For data preprocessing step, we also drop the human development index since more than half of the data is missing. Then we convert all categorical values to dummy variables by using one-hot encoding method, including year, country, sex, age and generation. We then stratify the response variable (suicide rate) into five categories (zero, low, medium, high, very high), which is based on the "*Archives of Suicide Research*" in 1999 published by Schmidtke, Armin. The zero category only includes zero suicide rate, the low category includes suicide rate ranging from 0 to 10 percent, the medium category includes suicide rate ranging from 10 to 50 percent, the high category includes suicide rate ranging from 50 to 100 percent, and the very high category includes suicide rate exceeding 100 percent. For Bayesian Ridge Regression method, we predict the suicide rate; for other Bayesian methods we predict the categories of suicide rate.

After preprocessing our data, we apply the linear discriminant analysis to reduce dimensionality. Since we convert all categorical values to dummy variables, our dataset contains more than three thousand columns. For reducing the dimensionality, we choose the linear discriminant analysis instead of principal component analysis because LDA attempts to model the difference between the classes of data, PCA on the other hand does not take into account of any difference in class. For applying Bayesian methods, first we fit the data into these models without applying LDA, then we fit the data into the same models after applying LDA to reduce dimensionality. After both steps, we compare the results and evaluate the performance of all Bayesian models before and after reducing the dimensionality.

The final step is to split train and test set. Our train set contains the first 20000 data points and test set contains the rest of 7820 data points.

Bayesian Methods Used and Mathematical Linkage

Our project applies the following five methods: Bayesian ridge regression, Gaussian naive bayes, Gaussian mixture, hidden Markov model and linear discriminant analysis. We apply these five Bayesian methods to our data by using the following Python libraries: BayesianRidge[1], GaussianNB[2], GaussianMixture[3], hmmlearn[4] and

[1] https://scikit-learn.org/stable/auto_examples/linear_model/plot_bayesian_ridge.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html
[4] https://hmmlearn.readthedocs.io/en/latest/

LinearDiscriminantAnalysis[5]. When applying the five models for the first time, we use the dataset that has only gone through basic preprocessing and one-hot coding steps. When applying the five models for the second time, we use the dataset that has gone through Linear Discriminant Analysis for dimension reduction. Before LDA, the dataset that is used for modeling includes the following 8 predictors: Suicide number, GDP per year, GDP per capital, Gender, Age, CountryYear, Generation, Suicide number per 100k population and 1 label of 5 classes: Zero, low, medium, high and very high suicide rate. After LDA dimension reduction, the dataset that has 9 variables (dimensions) and 5 output classes now have min(5 - 1, 9) == 4 discriminants i.e.: reduced dimensions.

1. Bayesian Ridge Regression
   Rather than generating a point estimate and a confidence interval as in classical regression, Bayesian regression is used to recover whole ranges of inferential solutions for linear models.

   Let $\phi_j(\cdot), j = 1, \ldots, K-1$ be basis function for the variables of x, with basis function and a linear model shown below:

$$f(\boldsymbol{x}_i) = \theta_0 + \sum_{j=1}^{K-1} \phi_j(x_{ij})\theta_j + \eta_i$$

$$f(\boldsymbol{x}) = \theta_0 + \sum_{j=1}^{K-1} \phi_j(\mathbf{x}_j)\theta_j + \boldsymbol{\eta}$$

$$= \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\theta} + \boldsymbol{\eta}$$

Assumptions

$$\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$$

$$|\boldsymbol{\phi}(\boldsymbol{x})| \neq 0$$

   To reduce effect of overfitting, we use Ridge regression which has a cost function.
   The Ridge Regression Bayes optimal solution is shown as below:

$$\hat{\boldsymbol{\theta}}_{Ridge} = \underset{\theta}{\operatorname{argmin}} \ (y - \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\theta})^T (y - \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$

$$\hat{\boldsymbol{\theta}}_{Ridge} = (\lambda \boldsymbol{I} + \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}))^{-1} \boldsymbol{\phi}(\boldsymbol{x})^T y$$

   We decide to use this method since this dataset's response variable--suicidal rate is continuous and linear in nature. The Gaussian prior in Bayesian ridge regression shrinks all effects more heavily than the Laplacian prior of the Bayesian LASSO[6]. In our dataset with numerous features and dimensions, we would like a regression method that shrinks the effect of multiple dimensions more heavily.

   Before LDA, the dataset with 8 predictors is used to fit the model. After LDA, the reduced dataset that has 4 dimensions is used to fit the model. The model aims to find the derived Bayes optimal solution for sum of squared errors for Suicide no. per 100k population.

2. Gaussian Naive Bayes
   Naive Bayes is a simple technique for constructing classifiers, where models that assign class labels to problem instances are represented as vectors of feature values and the class labels are drawn from some finite set. The mechanism is shown as below:

[5] https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

[6] Pasanen, Leena, Lasse Holmström, and Mikko J. Sillanpää. "Bayesian LASSO, scale space and decision making in association genetics." PloS one 10.4 (2015): e0120017.

For $i \in \{1,\ldots,K\}, j \in \{1,\ldots,p\}, \boldsymbol{x} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$

$$f(\boldsymbol{x}|g_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$
$$= \prod_{j=1}^{p} \mathcal{N}(\mu_{ji}, \sigma_{ji})$$

For $i,k \in \{1,\ldots,K\}$ choose $G = g_i$ if

$$\frac{f(\boldsymbol{x}|g_i)}{f(\boldsymbol{x}|g_k)} > \frac{Pr[G=g_k]}{Pr[G=g_i]}$$
$$\frac{\prod_{j=1}^{p} \mathcal{N}(\mu_{ji}, \sigma_{ji})}{\prod_{j=1}^{p} \mathcal{N}(\mu_{jk}, \sigma_{jk})} > \frac{Pr[G=g_k]}{Pr[G=g_i]}$$

The assumptions of Naive Bayes are as following:

a. Independent identically distributed (i.i.d) random variables
b. The continuous values associated with each class in the predictor values are distributed as a Gaussian distribution

The probability distribution of Naive Bayes is $x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$.

Before LDA, the dataset with 8 predictors is used to fit the model. After LDA, the reduced dataset that has 4 dimensions is used to fit the model. The model aims to maximize the prediction accuracy for class of suicide rate.

3. Gaussian Mixture

The probability density function distribution is $p(x) = \sum_{k=1}^{K} P_k p(x|k;\xi_k).$ where (Pk, ξk) needs to be estimated, (Pk, ξk) is represented by Θ, where Θ is a set of latent variables.
Procedure of Gaussian mixture EM is an iterative algorithm that starts from some initial estimate of Θ (e.g., random), and then proceeds to iteratively update Θ until convergence is detected from the number of mixtures, K, is usually determined by cross-validation.

The assumptions of Gaussian mixtures are listed below:

a. Assumes the underlying data is generated from Mixture of Gaussians.
b. I.i.d random variables

The formula for Gaussian Mixture is shown as following graph:

$$p(\vec{x}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i)$$
$$\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^{\mathrm{T}} {\Sigma_i}^{-1} (\vec{x}-\vec{\mu}_i)\right)$$
$$\sum_{i=1}^{K} \phi_i = 1$$

Since we are representing normally distributed subpopulations within an overall population, we find it to be appropriate to use this method on our dataset, which has many categorical variables. Before applying the LDA dimension reduction method, there were around 3000 features and we would have to use a very sophisticated method to compute the number of clusters, K. For each of the clusters, the mu will be this cluster's mean and covariance matrix will be this cluster's covariance matrix. However, after applying LDA dimension reduction, we get 4 features only and each feature is normalized. In this case the maximum

number of clusters will be K=4, for simplicity reasons we let K= 4. The mu and covariance matrix will simply be the mean and covariance matrix of each converted feature.

4. Hidden Markov Model (HMM)
   The methodology of HMM is a stochastic chain of connected probabilistic states, where each state generates an observation. One can only see the observations, and the goal is to infer the hidden state sequence.
   The assumptions for HMM:
   a. The Markov assumption: the next state is dependent only upon the current state.
   b. The stationarity assumption: state transition probabilities are independent of the actual time at which the transitions takes place.
   c. The output independence assumption: current output (observations) is statistically independent of the previous outputs (observations).

   Before the application of LDA, the states of HMM can be seen as the combination of multiple features, such as year, groupage and GDP. With conversion to dummy variables, there are about 3000 transition states. In the package of HMM in python, the emission probability is 1 over the count of numbers of instances in a specified state; the transition probability is 1 over 3000. After the application of LDA, the states of HMM becomes the combination of normalized components of LDA. The emission probability is 1 over the count of numbers of instances in a specified state; the transition probability is 1 over 4.

5. Linear Discriminant Analysis
   The linear discriminant analysis has strong assumptions such as:
   a. Multivariate normality of the explanatory variables with equal covariance matrices during populations
   b. I.i.d with low multicollinearity

   Note that the LDA method is more suitable for smaller data sets. It generates higher bias and lower variance. Probability function $f_i(x)$ is the class—conditional density of X in class G=i, π_i is the prior probability of class i, with $\sum_{i=1}^{I}\pi_i = 1$. The probability function for LDA is as the following picture:

$$\Pr(G = i \mid X = x) = \frac{f_i(x)\ \pi_i}{\sum_{i=1}^{I}\pi_i f_i}$$

   Suppose:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} EXP^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$$

   The LinearDiscriminantAnalysis library we use come with built-in regularization and dimension reduction steps. With this built-in reduction functionality, we use our pre-LDA dimension reduction dataset for the LDA prediction model, and the result is the prediction result based on the reduced dataset. Note that within this package of LDA, the class prior probability is simply 1 over the count of observations in that specific class. Practically speaking, the actual prior data is very hard to be obtained for numerous groups of different frequencies. With LDA prediction, we are also looking for maximal prediction accuracy for class of suicide rate.

Results and Conclusions

We begin with the evaluation of the results of each model before applying linear discriminant analysis. In the table below, both Gaussian Naive Bayes and Gaussian Mixture do not perform very well and the accuracy are both around 30%. For Hidden Markov Model, it takes too long to calculate the prediction accuracy since after using one-hot encoding, we have more than three thousand features, which makes the calculation become very complex. In order to find a baseline comparison for the result of Bayesian Ridge Regression, we also fit our data into the linear regression. It clearly shows that the result of those two methods are very similar, and Bayesian Ridge Regression does not improve the mean square error.

Methods	Gaussian Naive Bayes	Gaussian Mixture	Hidden Markov Model
Prediction Accuracy	33.36%	27.11%	N/A

Methods	Linear Regression	Bayesian Ridge Regression
Mean Square Error	262.39	265.87

We then evaluate the result of each model after applying linear discriminant analysis. This time the Gaussian Naive Bayes improves significantly, reaching 81.56% prediction accuracy. The outstanding performance of Gaussian Naive Bayes validates the assumption which requires normal distribution of each feature is met after applying LDA. Gaussian Mixture also improves a little bit, but it is still below 40%.

The Hidden Markov Model produces the lowest prediction accuracy, due to its violation of stationarity assumption and Markov assumption. The combination of multiple features such as “year”,“agegroup”, etc. could be used for specifying states, but in this case, the suicidal rate of next state is not based on the combination of current year, agegroup etc.. The states of suicide rate is not sequential in nature. If we want to use all the features to generate states that are dependent on previous state, the Hidden Markov model becomes utterly complicated. We would therefore not go further with this method either because its low prediction accuracy and its violation of two important assumptions.

Next, instead of using LDA to reduce dimensionality, we also use it to make predictions. The Linear Discriminant Analysis reaches second highest prediction accuracy, which is 65.15%. Finally, we still use linear regression as a comparison of Bayesian Ridge Regression. This time the mean square errors for both methods decrease a lot, but the Bayesian Ridge Regression is still slightly worse than linear regression. After applying LDA, the number of features reduces a lot, and this could make the penalizing component of Bayesian Ridge Regression have less effect.

In conclusion, the Gaussian Naive Bayes could be used for future modeling and prediction in preventive suicide studies.

Methods	Gaussian Naive Bayes	Gaussian Mixture	Hidden Markov Model	Linear Discriminant Analysis
Prediction Accuracy	81.56%	37.28%	25.90%	65.15%

Methods	Linear Regression	Bayesian Ridge Regression
Mean Square Error	183.817	183.824

Reference

1. Schmidtke, Armin, et al. "Suicide rates in the world: update." Archives of Suicide Research 5.1 (1999): 81-89.
2. Krebs, W. G., et al. "Tools and databases to analyze protein flexibility; approaches to mapping implied features onto sequences." Methods in enzymology. Vol. 374. Academic Press, 2003. 544-584.
3. Warakagoda, Narada. “Assumptions in the Theory of HMMs.” Assumptions in the Theory of HMMs, 10 May 1996, jedlik.phy.bme.hu/~gerjanos/HMM/node5.html.
4. Probabilistic Learning: Theory and Algorithms, CS 274A. “The EM Algorithm for Gaussian Mixtures” https://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf
5. Barber, David. Bayesian reasoning and machine learning. Cambridge University Press, 2012.
6. Bayesian Least Squares Regression, DS 6559 slides, Donald E. Brown, University of Virginia, 2019
7. NaiveBayesHO, DS 6559 slides, Donald E. Brown, University of Virginia, 2019
8. Pasanen, Leena, Lasse Holmström, and Mikko J. Sillanpää. "Bayesian LASSO, scale space and decision making in association genetics." PloS one 10.4 (2015): e0120017.