Capstone Seminar

Fall 2017

- **September 13**
  Time: 1:10–1:25 pm
  Speaker: Wangshu Tu
  Title: Parameters Selection and Comparison in Guassian Kernel SVM *Abstract:* In SVM(Support Vector Machines), it is not so clear that which kernel to choose and how to select proper parameter in kernel function. One of them: Gaussian radial basis function(RBF) is very popular because of only single parameter needs to be determined. In this short talk, it will present different results of applying RBF kernel in binary classification case–Gender Recognition by Voice, with different pairs of (C, r), where C is a regularization parameter to constrain the range of Lagrangian coefficients in dual function F_D, r is reciprocal of single parameter sigma^2 in RBF kernel. For each pair (C, r), compute 10-folder Cross Validation(CV/10) misclassification rate, and it indicates the rate will decrease when C or r increases. The smallest CV/10 misclassification rate among all pairs of (C, r) is also better than LDA and classification tree.

- **September 13**
  Time: 1:25–1:40 pm
  Speaker: Xiang Wang
  Title: Use LDA and QDA To Discriminate Diabetes Data *Abstract:* Diabetes data resulted from a study conducted at the Stanford Clinical Research Center of the relationship between the three clinical classifications and five measurements for 145 instances. It helps the diagnosis and appropriate treatment to the diabetes patients. We draw a scatterplot matrix of all five variables representing the problematic multivari- ate Gaussian distributions, and the assumption of equal covariance matrices is inappropriate, which play the negative roles in LDA(Linear Discriminant Analysis) and QDA(Quadratic Discriminant Analysis). We use LDA and QDA to discriminate the clinical classifications by the five variables, then draw the 2D-scatter-plot of the first two discriminating functions to show that LDA is subject to outliers, but QDA relatively improves the classification by nonlinear discrimination. The misclassification rates in the leave-one-out cross-validation are 11% for LDA and 9.7% for QDA. The fitting for 145 instances indicates LDA and QDA can be quite flexible.

- **September 13**
  Time: 1:40–1:55 pm
  Speaker: Joshua Rovou
  Title: Managing Multinomial Data: Using Aids and Examples from Dr. Ganggang Xu and Julian Faraway's "Generalized Linear Models" *Abstract:* Multinomial data requires careful thought to properly analyze. There are various forms that multinomial data can take that can easily be misclassified, leading to false conclusions. Understanding when, why, and how to recognize and apply these forms and their assumptions allows a data scientist to be mindful of best practices when fitting multinomial models. This paper provides an overview, suggestions, and examples of classifying multinomial data. In addition, this paper provides discussions on applying the appropriate assumptions and models in the R programming language, using the problem and data sets in Faraway's "Generalized Linear Models" Chapter 5 as case studies. These examples seek to illustrate best practices when dealing with multinomial data for the student data scientist.

- **September 13**
  Time: 1:55–2:10 pm

Speaker: Hao Wang
Title: Forest Cover Type prediction *Abstract:* This report is based on the data from UCI machine learning website and the raw data was determined from the US Forest Service (USFS) Region 2 Resource Information System (RIS). There are totally 581,012 observations in this data set and each observation is a 30m x 30m cell containing 54 attributes, which consist of 10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables. The forest cover type is basically a classification problem. By looking from the previous work from Blackard's investigation on this topic (1), we think that we could try using several classifiers such as SVM,K-NN, decision tree and gradient boost model in modeling to increase the accuracy in prediction.

- **September 20**
  Time: 1:10–1:25 pm
  Speaker: Xiaolin Tang
  Title: MODEL SELECTION *Abstract:* TBA.For a dataset, we can come up with plenty of models, however, there should be a best one through all of them. Therefore, we need some criteria to evaluate the model, and a method to find out it, here comes model selection. My presentation will be presented with two parts: first, I will give a quickly review of some basic concepts, including the criteria of a good model and the method to do model selection. Second, I will display several questions of chapter 10 in “Faraway I”(Linear Models with R, second Edition) and talk about the solutions. For Question 10.5, it's about how outliers influence the result of model selection, we use the data with and without the outliers to do model selection separately, and our conclusion is we can obtain the same model, however, the estimated coefficient will be very different for these two models.

- **September 20**
  Time: 1:25–1:40 pm
  Speaker: Shaofei Zhao
  Title: Use Dataset aatemp to Predict Temperature in the Future *Abstract:* Dataset aatemp is from the U.S. Historical Climatology Network. They are the annual mean temperatures in Ann Arbor, Michigan going back about 150 years. The data contains 115 observations on two variables, year is the year from 1854 to 2000, temp is the annual mean temperature. Our motivation is that by analyzing the data, we may give a reasonable prediction of the mean temperature in 2020. To analyze, we need firstly check the assumptions of error terms, such as constant variance, normality and outliers. Then we proceed linear regression considering response transformation and predictor transformation, in the long talk we may also consider some methods of time series to achieve a better model. At the moment, we use a model with 3-degree polynomials, which can increase our R^2 from 0.05 to 0.13, and base on this model, we may predict the temperature of 2020 is about 46 degrees.

- **September 20**
  Time: 1:40–1:55 pm
  Speaker: McInroy,Alexander
  Title: Binomial Regression and its Applications in Medical Diagnosis *Abstract:* Regression analysis remains one of the most directly applicable tools in the field of statistics and machine learning today. In particular, binomial regression is especially useful for diagnosing medical conditions where a test leads to a positive or negative result. This talk will use examples from Extending the Linear Model with R, Faraway to illustrate this. Technical topics covered will be data analysis, data interpretation, model selection, and model prediction. Furthermore, selecting appropriate cutoff points for a binomial model’s response will be discussed, since mitigating type I and type II error is a subjective balancing act depending on the situation..

- **September 20**
  Time: 1:55–2:10 pm
  Speaker: Schepis,Michael

Title: Classification Methodology *Abstract:* One of the primary motivations of analyzing data is our ability to accurately categorize it. This is the purpose of classification; Different statistical techniques allow us to divide categorical data into relevant groups, however this ability is only as useful as our ability to interpret these groups in a meaningful way. Most data of interest cannot be perfectly separated without error, so it is also crucial to analyze mis-classification rates with each of these statistical tools. In this talk, we will provide relevant examples from Alan J. Izenman's Multivariate Statistical Techniques and discuss the tools of Linear Discriminant analysis, Bayes Rule, and Quadratic Discriminant Analysis and when each of those techniques would be most appropriate and how to create a confusion matrix to examine the accuracy of our findings..

- **September 27**
  Time: 1:10–1:25 pm
  Speaker: Yanwei Jiang
  Title: Lineal Model diagnostics in error terms *Abstract:* When a linear model has been constructed, we need to check whether the model is appropriate and do some adjustment if necessary, this process is called model diagnostics. There are several aspects that we need to consider, in this short talk, we will focus on the error term of linear model. The model assumptions of the error term are constant variance, normality and independent. We will introduce some common and useful tools such as residual plot, qq-plot and other tools, both in theory and method. Some examples will also be demonstrated to show how these tools work.

- **September 27**
  Time: 1:25–1:40 pm
  Speaker: Yifei Zeng
  Title: Application of Multidimensional Scaling *Abstract:* Multidimensional Scaling (mds) is a method to reduce high-dimensional-data to low-dimensional-data, most likely to 2-dimensions or 3-dimensions which is the dimension that human can visualize. Also mds can also be used to cluster different groups of data. In this talk, we will discuss different types of mds and apply them to a data set from Izenman. The data set contains the distance between 48 cities in UK which forms a dissimilarity matrix. We will compare different types of mds with respect to visualization, computation, and fits of data etc. Also within a distinct type of mds, we will talk about how to choose the dimension so when we are reducing the dimension, we will still get enough information from data. We will finally present the map of the 48 cities and compare with the mds' map to see the fit of mds.

- **September 27**
  Time: 1:40–1:55 pm
  Speaker: Gang Cheng
  Title: Problems with the error term *Abstract:* In ordinary least square regression, we assume the error term ϵ is independent and identically distributed. Furthermore, in order to carry out the usual statistical inference, we also assume the error term are normally distributed. However, in many cases, this assumption always violated and we have to consider alternatives. (i) When the errors are dependent, like time series, we use \emph{Generalized Least Squares}(GLS); (ii) When the errors are independent, but not identically distributed, we can use \emph{Weighted Least Squares}. (iii)When the errors are not normally distributed, we can use \emph{Robust Regression}. In this talk, I will mainly focus on the theory part of these regressions and one or two example(s) of these.

- **September 27**
  Time: 1:55–2:10 pm
  Speaker: Chenxi Wang
  Title: Linear Model with categorical predictors *Abstract:* My topic for seminar is focused on Chapter 14 of the book Linear Models of R by Faraway. This chapter mainly talks about categorical predictors for linear regression models.

I will give brief talk about basic concepts related to categorical predictors. Also, together with exercises at the end of this chapter, I will give specific examples on how to deal with categorical factors in regression analysis in practical world.

- **October 4**
  Time: 1:10–1:40 pm
  Speaker: Hao Wang
  Title: Forest Cover Type prediction *Abstract:* This report is based on the data from UCI machine learning website and the raw data was determined from the US Forest Service (USFS) Region 2 Resource Information System (RIS). There are totally 581,012 observations in this data set and each observation is a 30m x 30m cell containing 54 attributes, which consist of 10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables. The forest cover type is basically a classification problem. By looking from the previous work from Blackard's investigation on this topic (1), we think that we could try using several classifiers such as SVM,K-NN, decision tree and gradient boost model in modeling to increase the accuracy in prediction.

- **October 4**
  Time: 1:40–2:10 pm
  Speaker: Chenxi Wang
  Title: Linear Models with Categorical Predictors *Abstract:* My topic for seminar is focused on Chapter 14 of the book Linear Models of R by Faraway. This chapter mainly talks about categorical predictors for linear regression models. I will give brief talk about basic concepts related to categorical predictors. Also, together with exercises at the end of this chapter, I will give specific examples on how to deal with categorical factors in regression analysis in practical world.

- **October 11**
  Time: 1:10–1:40 pm
  Speaker: Shaofei Zhao
  Title: Use Different Model to Predict the Temperature of Binghamton *Abstract:* Weather forecasting has always been interesting and challenging, since we have already evaluated dataset aatemp during our short talk, using linear model may not get a good consequence for a variety of reasons, so it's necessary for us to consider making some changes both at the dataset and the methods, after these changes, we are supposed to get a better model and more precise predictions. This time we go to the same website U.S. Historical Climatology Network and download average daily temperature in Binghamton, and perform several different methods to fit the data, including linear model, time series model, etc. We will also test the goodness of fit and other model assumptions. Finally, we use different models to predict the future temperature, and make comments..

- **October 11**
  Time: 1:40–2:10 pm
  Speaker: Yifei Zeng
  Title: Application of Multidimensional Scaling *Abstract:* Multidimensional Scaling (mds) is a method to reduce high-dimensional-data to low-dimensional-data, most likely to 2-dimensions or 3-dimensions which is the dimension that human can visualize. Also mds can also be used to cluster different groups of data. In this talk, we will discuss different types of mds and apply them to a data set from Izenman. The data set contains the distance between 48 cities in UK which forms a dissimilarity matrix. We will compare different types of mds with respect to visualization, computation, and fits of data etc. Also within a distinct type of mds, we will talk about how to choose the dimension so when we are reducing the dimension, we will still get enough information from data. We will finally present the map of the 48 cities and compare with the mds' map to see the fit of mds.

- **October 18**
  Time: 1:10–1:40 pm
  Speaker: Gang Cheng
  Title: Problem with the error *Abstract:* The Robust Regression method down weighting the extreme cases, but sometimes, when the large errors are sufficient numerous and extreme in value, it still failing. We need methods which fit the data well even in the presence of such problems. Least Trimmed Regression is a good way for dealing with data with many bad entries. Unlike Robust Regression, it gives no weights on such bed entries. In terms of determining the significance of variables, the R does not provide us with the standard error or p.value. Instead, we can solve it by bootstrap. In this talk, I will focus on the properties of LTS estimator and give examples of its working in real situation.

- **October 18**
  Time: 1:40–2:10 pm
  Speaker: Schepis,Michael
  Title: Classification Methodology Using Examples from Izenmana's Modern Multivariate Statistical Techniques *Abstract:* One of the primary motivations of analyzing data is our ability to accurately categorize it. This is the purpose of classification; Different statistical techniques allow us to divide categorical data into relevant groups, however this ability is only as useful as our ability to interpret these groups in a meaningful way. Most data of interest cannot be perfectly separated without error, so it is also crucial to analyze misclassification rates with each of these statistical tools. In this talk, we will provide relevant examples from Alan J. Izenmanâ's \emph{Multivariate Statistical Techniques} and discuss the tools of Linear Discriminant analysis, Bayes Rule, and Quadratic Discriminant Analysis and when each of those techniques would be most appropriate and how to create a confusion matrix to examine the accuracy of our findings. For each technique we will discuss how the classifyier functions and give examples in which it might be applied.

- **October 25**
  Time: 1:10–1:40 pm
  Speaker: McInroy,Alexander
  Title: Binomial Regression and its Applications in Medical Diagnosis *Abstract:* Regression analysis remains one of the most directly applicable tools in the field of statistics and machine learning today. In particular, binomial regression is especially useful for diagnosing medical conditions where a test leads to a positive or negative result. This talk will use examples from Extending the Linear Model with R, Faraway to illustrate this. Technical topics covered will be data analysis, data interpretation, model selection, and model prediction. Furthermore, selecting appropriate cutoff points for a binomial model’s response will be discussed, since mitigating type I and type II error is a subjective balancing act depending on the situation. Special attention will be given to selecting an appropriate cutoff point.

- **October 25**
  Time: 1:40–2:10 pm
  Speaker: Rovou,Joshua
  Title: Managing Multinomial Data: With examples from Dr. Ganggang Xu and Faraway's Extending the Linear Model *Abstract:* Multinomial data requires careful thought to properly analyze. There are various forms that multinomial data can take that can easily be misclassified, leading to false conclusions. Understanding when, why, and how to recognize and apply these forms and their assumptions allows a data scientist to be mindful of best practices when fitting multinomial models. This paper provides an overview, suggestions, and examples of classifying multinomial data. In addition, this paper provides discussions on applying the appropriate assumptions and models in the R programming language, using the problem and data sets in Faraway’s “Generalized Linear

Models" Chapter 5 as case studies. These examples seek to illustrate best practices when dealing with multinomial data for the student data scientist.

- **November 1**
  Time: 1:10–1:40 pm
  Speaker: Xiaolin Tang
  Title: MODEL SELECTION *Abstract:* For a dataset, we can come up with plenty of models, however, there should be a best one through all of them. Therefore, we need some criteria to evaluate the model, and a method to find out it, here comes model selection. My presentation will be presented with two parts: first, I will give a quick review of some basic concepts, including the criteria of a good model and the method to do model selection. Second, I will talk about several questions of chapter 10 in "Faraway I"(Linear Models with R, second Edition). To be specific, 10.1,10.4,10.5 and 10.6 will be covered. 10.1 is basic a review of model selection method. 10.4 talks about how different the reduced model can be with different selection direction, by direction I mean forward, backward etc. And the answer for this problem is the model will be quite different. 10.5 is about how outliers influence the result of model selection, we use the data with and without the outliers to do model selection separately, and we obtain the same model but with different estimated coefficient. And lastly, we will talk about how model selection infects prediction interval, it will be a little bit different from full model.

- **November 1**
  Time: 1:40–2:10 pm
  Speaker: Xiang Wang
  Title: Linear Discriminant Analysis for Diabetes Data

Abstract: Diabetes data resulted from a study conducted at the Stanford Clinical Research Center of the relationship between the three clinical classifications and five measurements for 145 instances. It helps the diagnosis and appropriate treatment to the diabetes patients. Based on all five variables representing the problematic multivariate Gaussian distributions and the inappropriate assumption of equal covariance matrices, misclassification rates in the leave-one-out cross-validation are 11% for LDA and 9.7% for QDA. Thus we will talk about the LDA via Multiple Regression and Logistic Discrimination, which is a semi-parametric model and asymptotically less efficient than is Gaussian LDA. In the view of Logistic Discrimination, we can use data transformation, variable selection techniques, polynomial regression, and other powerful methods to improve the result of LDA and QDA. Then we verify that when the Gaussian distributional assumptions or the common covariance matrix assumption are not satisfied, Logistic discrimination performs much better, and is more robust to non-normality than Gaussian LDA.

- **November 8**
  Time: 1:10–1:40 pm
  Speaker: Wangshu Tu
  Title: Parameters Selection in Guassian Kernel SVM *Abstract:* In Support Vector Machines(SVM), to get small misclassification rate(MCR), we need to adjust C, which is regularization parameter to constrain the range of Lagrangian coeffcients α in dual function FD, and parameters τ in kernel function. But often it is not so clear that how to select them. In this long talk, Gaussian radial basis function(RBF) will be used as kernel function, because only one parameter(C,τ) needs to be determined. In order to get desired parameter, one method is to find minimal missclassification error by using 10-folder Cross Validation(CV/10) with different pairs of (C, τ). Based on that, other optimization methods, and An Automatic Method for Selecting the Kernel Parameter τ will be discussed.

- **November 8**
  Time: 1:40–2:10 pm
  Speaker: Yangwei Jiang
  Title: Lineal Model diagnostics *Abstract:* When a linear model has been constructed, we need to check whether the model is appropriate and do some adjustment if necessary, this process is called model diagnostics. There are several aspects that we need to consider, during the short talk, we have discussed about the error term. In this long talk, we will discuss about the outliers, leverage and influential points, and we will also introduce partial regression plot and partial residual plot to detect outliers and non-linearity of the regression model.

Spring 2017

- **February 23**
  Time: 1:15–2:15 pm
  Speaker: Yu Hu
  Title: Vehicle's Fuel Economy Data Analysis *Abstract:* This project I did aimed at providing reliable estimates for comparing vehicles in 2016. The purpose is to help car buyers choose the most fuel-efficient vehicle that meets their needs. Using the linear regression knowledge to analysis the correlation between each factor(vehicles's weight, cylinder, engine size, etc.) and response(fuel economy).
- **March 2**
  Time: 1:15–2:15 pm
  Speaker: Hao Wang
  Title: On testing independence and goodness of fit in linear models *Abstract:* We consider a linear regression model and propose an omnibus test to simultaneously check the assumption of independence between the error and the predictor variables and the goodness of fit of the linear regression model.
- **March 16**
  Time: 1:15–2:15 pm
  Speaker: Liping Gu
  Title: Analysis of the Dataset “Wine” *Abstract:*In the talk, I discuss the statistical analysis on the dataset “wine”. First I converted the numerical predictors into factorial ones and used the method which is similar to the factorial design to find the impact of different factors.Then making use of the algorithms of LDA and QDA I analyzed the data. The result of the computation shows a clear tendency on the means of the predictors that stand out as significant in the result of the factorial model. The data analysis suggests that 4 predictors are significant, whereas the other 7 are not.

- **May 2**
  Time: 12–1 pm
  Speaker: Hao Wang
  Title: On testing independence and goodness of fit in linear models *Abstract*: We consider a linear regression model and propose an omnibus test to simultaneously check the assumption of independence between the error and the predictor variables and the goodness of fit of the linear regression model.