

DS 432 Predictive Modeling for Data Science Lab Assignments (Last Updated: 17/12/2018)

A. CRISP-DM

Reference - *Larose & Larose, Data Mining and Predictive Analytics, Wiley, Second Edition. (Part I - Data Preparation; Chapters 1, 2, 3, and 4.)*

Perform the CRISP-DM analysis as per instructions (Data Preprocessing, Exploratory Data Analysis, and if required, Dimension-Reduction Methods) of the four data sets given below.

1. Bank Loan Default
2. Campaign Offers
3. Retail Forecasting
4. Telco Churn

B. Regression - I

References - *Linear Models with R by Julian J. J. Faraway; Linear Regression Analysis – Theory and Computing by Xin Yan & Xiaogang Su.*

1. Consider three data sets given below. Remove every tenth observation from the each data set for use as a test sample. Use the remaining data as a training sample and build (Diagnostics -> Transformation -> Variable Selection -> Diagnostics) an appropriate linear regression model. Use the models you find to predict the response in the test sample and compare the results.
 - (a) RegD1.txt
 - (b) RegD2.txt
 - (c) RegD3.txt
2. Use the data set RegD14.txt, to fit a model. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant.
 - (a) Check and comment on the constant variance assumption for the errors.
 - (b) Check and comment on the normality assumption.
 - (c) Check and comment on the large leverage points.
 - (d) Check and comment on the outliers.
 - (e) Check and comment on the influential points.
 - (f) Check and comment on the structure of the relationship between the predictors and the response.
 - (g) Compute and comment on the condition numbers.
 - (h) Compute and comment on the correlations between the predictors.

- (i) Compute and comment on the VIF.
- 3. Use the data **RegD4.txt** to fit a model using the following methods.
 - (a) Least squares.
 - (b) Least absolute deviations.
 - (c) Huber method.
 - (d) Least trimmed squares.

Compare the results. Use diagnostic methods to detect any outliers or influential points. Remove these points and then use least squares. Compare the results.
- 4. Use the data **RegD7.txt** to fit a model with Y as the response and only X_3 , X_4 , and X_5 as predictors. Use the Box-Cox method to determine the best transformation on the response.
- 5. Use the data **RegD8.txt** to fit a linear model. Implement the following variable selection methods to determine the "best" model.
 - (a) Backward Elimination.
 - (b) AIC, AICC, BIC.
 - (c) R^2 , R_a^2 .
 - (d) Mallows C_p .
- 6. Consider the data **RegD9.txt**. Remove every tenth observation from the data for use as a test sample. Use the remaining data as a training sample building the following models.
 - (a) Linear regression with all predictors.
 - (b) Linear regression with variables selected using AIC.
 - (c) Principle component regression.
 - (d) Partial least squares.
 - (e) Ridge regression.

Use the models you find to predict the response in the test sample. Make a report on the performance of the models.

C. Regression - II

Reference: *Parallelism in Matrix Computations*, by Efstratios Gallopoulos, Bernard Philippe, and Ahmed H. SamehQ, Springer; *Parallel R*, Ethan McCallum and Stephen Weston, O'Reiley.

```
n1 <- 10000000
x1 <- 1:n1
x2 <- runif(n1,5,95)
x3 <- rbinom(n1,1,.4)
x4 <- rnorm(n1, mean=-30, sd=200)
x5 <- runif(n1,-5000,5000)
b0 <- 17; b1 <- -0.466; b2 <- 0.037; b3 <- -5.2; b4 <- 2; b5 <- 0.00876
```

```

sigma <- 1.4
epsilon <- rnorm(x1,0,sigma)
y <- b0 + b1*x1 + b2*x2 + b3*x3 + b4*x4 + b5*x5 + x1*x2 + epsilon
data14a<-cbind(y,x1,x2,x3,x4,x5)
write.table(data14a, file = "data14a.txt", sep = " ", quote=T)

```

1. Generate, if possible, four moderately large data files (with n_1 equal to 1,000,000, 10,000,000, 50,000,000 and 100,000,000) as per the sample code given above. If required, write data14a.txt into a database file on the hard disk. Suppose we want to fit a linear model

$$y = X\beta + \epsilon \text{ on the data. Compute } \hat{\beta} = (X'X)^{-1}X'y \text{ and } \hat{\sigma} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \text{ using}$$

the R packages given below. Comment and compare (perhaps using GUIProfiler) on your findings, including advantages, errors, space & time complexities, shortcomings (some of the packages below could be obsolete or outdated), etc. You can also search and find other packages and include them in the report.

- (a) lm
- (b) biglm
- (c) bigmemory/biganalytics
- (d) ff/biganalytics
- (e) RNetCDF
- (f) DBI/RSQLite/sqlldf/RPostgreSQL/RODBC
- (g) snow/snowfall
- (h) HadoopStreaming
- (i) Rhipe
- (j) RHadoop

Also, check and give a comparative report on how the above computation works in Python (Numpy, SciPy, Pandas, Scikit-learn, etc.)

D. Time Series

References: *Introductory Time Series with R* by Cowpertwait and Metcalfe, Springer, and *An Introduction to Analysis of Financial Data with R* by Ruey S. Tsay, Wiley.

- (a) Identify the ARIMA(p, d, q) model and the white noise variance estimate for the given data sets. (Use `ts.plot`, `acf`, `pacf`, `eacf`, `arima`, etc. Avoid using `auto.arima` except for verifying your answer.)
 - i. TSD1
 - ii. TSD2
 - iii. TSD3
 - iv. TSD4
 - v. TSD5

- vi. TSD6
 - vii. TSD7
 - viii. TSD8
 - ix. TSD9
 - x. TSD10
- (b) Simulate a series of $n = 500$ Gaussian white noise observations as in and compute the sample ACF, $\hat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. Now repeat the same by using only $n = 50$. How does changing n affect the results?
 - (c) Consider the `so2` data set, which is part of `astsa` package. Fit an $\text{ARIMA}(p, d, q)$ model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast the data into the future four time periods ahead (about one month) and calculate 95% prediction intervals for each of the four forecasts. Comment.
 - (d) Perform a time series model specification, estimation, model diagnostics, and forecasting for a financial stock of your choice.

E. Classification

Reference: *An Introduction to Statistical Learning by James, Witten, Hastie, and Tibshirani, Springer; Applied Predictive Modeling (Chapter 8), Kuhn, Springer.*

- (a) Consider the `Weekly` data set, which is part of `ISLR` package. It contains the weekly stock market returns for 21 years.
 - i. Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any pattern?
 - ii. Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appears to be statistically significant? If so, which ones?
 - iii. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
 - iv. Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010.)
 - v. Repeat (d) using linear discriminant analysis (LDA).
 - vi. Repeat (d) using quadratic discriminant analysis (QDA).
 - vii. Repeat (d) using K-nearest neighbors (KNN) with $K = 1$.
 - viii. Which of these methods appears to provide the best results on this data?
- (b) This problem involves predicting `Salary` from the `Hitters` data set which is part of the `ISLR` package.

- i. Remove the observations for whom the salary information is unknown, and log-transform the salaries.
 - ii. Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.
 - iii. Perform boosting on the training set with 1000 trees for a range of values of the shrinkage parameter λ . Produce a plot with different shrinkage values on the x -axis and the corresponding training set MSE on the y -axis.
 - iv. Produce a plot with different shrinkage values on the x -axis and the corresponding test set MSE on the y -axis.
 - v. Which variable appear to be the most important predictors in the boosted model?
 - vi. Apply bagging to the training set. What is the test set MSE for this approach?
 - vii. Apply random forests to the training set. What is the test set MSE for this approach?
- (c) Generate a simulated two-class data set with 100 observations and two features in which there is a visible but non-linear separation between the classes. Show that in this setting, a support vector machine with a polynomial kernel (with degree greater than 1) or a radial kernel will outperform a support vector classifier on the training data. Which technique performs best on the test data? Make plots and report training and test error rates in order to back up your assertions.

F. Unsupervised Learning (Principal Component Analysis and Clustering)

Reference: *An Introduction to Statistical Learning (Chapter 10)* by James, Witten, Hastie, and Tibshirani, Springer.

1. In this problem, we will generate simulated data, and then perform PCA and K-means clustering on the data.
 - (a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables. (Be sure to add a mean shift to the observations in each class so that there are three distinct classes.)
 - (b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.
 - (c) Perform K-means clustering of the observations with $K = 3$. How well do the clusters that you obtained in K-means clustering compare to the true class labels?
 - (d) Perform K-means clustering with $K = 2$. Describe your results.
 - (e) Now perform K-means clustering with $K = 4$, and describe your results.
 - (f) Now perform K-means clustering with $K = 3$ on the first two principal component score vectors, rather than on the raw data. That is, perform K-means clustering on the 60 by 2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

- (g) Using the `scale()` function, perform K-means clustering with $K = 3$ on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (b)? Explain.