CIS417 Intro. to Business Analytics

Fall 2015, Rajiv Dewan

Basic Course Information

Attached Files: CIS 417 Syllabus Fall 2015.pdf (229.725 KB)

A generic syllabus is available. This web page offers the definitive and up to date details.

The textbook for the course is:
(PF) Foster Provost and Tom Fawcett (2013) *Data Science for Business*, O'Reilly.

Additionally the following books are recommended:
(DW) Dona Wong, (2013), *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*, W.W. Norton & Co.

(BJ) Ben Jones (2014), *Communicating Data with Tableau*, O'Reilly. Note: While this book refers to Tableau 8.1, it is generally applicable to the current version of Tableau.

(JL) Jared Lander (2014) *R for Everyone: Advanced Analytics and Graphics*, Addison Wesley.

Lecture 1: Introduction to Business Analytics

Attached Files:
File 01 Intro.pdf (383.722 KB)
File IBM and Oxford Report on Big Data Nov 2012 MU.pdf (1.708 MB)
File Data Scientist Davenport Patil HBR Oct 2013 MU.pdf (1.656 MB)

Topics:

- Lighthouse examples of Business Analytics:
  - Target
  - Walmart,
  - LinkedIn,
  - Sprint
- What is business analytics?
- Sources of data
- Big Data
- Descriptive analytics
- Predictive analytics
- Intro to data used in this course

Reading: Textbook PF Chapters 1 and 2, the IBM/Oxford Report, and the HBR article.

Lecture 2: Principles of Data Visualization

Attached Files:

File 02 Principles of Data Visualization.pdf (1.198 MB)

Topics:

- Main techniques for descriptive analytics
- Data visualization
- Why visualization matters
  - Expressiveness - Anscombe's quarter
  - Trends
  - Powerful and fast
- Limits of human visual perception
  - acuity
  - limited ability to distinguish between shades of gray and colors
  - limited and fleeting visual memory
  - effect of surroundings and context
- Encoding for rapid perception
- Best practices in data visualization from Wong and Tufte
  - Data sources
  - Editing
  - Plotting
  - Reviewing

Reading: Refer to the supplemental text by Wong (DW) for details on do's and don'ts for data visualization.

Installing Tableau

To prepare for classes coming up, you must start installing some key software now:

Tableau is needed for the next topic. (Windows and Mac versions available):

Landing Page: http://www.tableausoftware.com/tft/activation
Desktop Key: Sent by Email to Registered Students
Instructions: Click on the link above and select Get Started. On the form, enter your university email address for “Business email”; and under "Organization", please input the name of your school.

Lectures 3 and 4: Introduction to Tableau

Attached Files:
File coffee chain data.xlsx (449.101 KB)
File data visualization example.xlsx (8.091 KB)
File 03 Intro to Tableau v2.pdf (1.544 MB)
File 03 Basic Tableau Operations v9-2.pdf (80.1 KB)
File 04 Advanced Tableau Operations.pdf (782.065 KB)
File quarterly revenue.xlsx (10.191 KB)

Topics:

- What is Tableau?
- Connecting to data
- Measures and Dimensions
- Shelves and Marks cards
- Key concept: visual attributes are controlled by data
- Varieties of aggregations
- Displaying how much / how many
- Stacked bar charts
- Heat maps
- Bar-in-Bar chart
- Measuring against a goal

Reading: Optional book by Jones (BJ) is helpful here.

Homework 1: Tableau

Attached Files:
File HW1 Tableau.pdf (140.216 KB)
File coffee chain data.xlsx (449.101 KB)
There are two parts to the homework.

Part I: Skill building exercise that you do not have to turn in. This is available as Build-It-Yourself-Exercises from Tableau. I recommend that you do them all.

Part II: An assignment that you will upload. One per team. It will be graded. The details are in the attached file. You will use the coffee chain data.

Installing R and R Studio

Attached Files:

File Installing R.pdf (678.648 KB)

Lectures 5 and 6: Data Visualization using R

Attached Files:

File 04 intro to R.pdf (638.919 KB)
File rswe.Rda (13.538 KB)
File Coffee Chain data.csv (473.373 KB)
File 04 data vis using R script v2.pdf (80.664 KB)
File 04 Data Vis using R Slides v2.pdf (746.64 KB)
File cc.Rda (71.209 KB)
File roc_sept_wx1.csv (32.889 KB)
File raw_temp.txt (806 B)

Topics:

- Data Visualization Using R
- Rochester weather data set
- ggplot2 package in R
- Visualizing a single measure
- Visualizing one dimension
- Visualizing a measure and a dimension
- Two or more dimensions
- Contingency tables and tests of independence

Readings: From the recommended Book JL
Chapter 5.1 on data frames
Chapter 6.1, 6.2 and 6.5 on Reading data into R
Chapter 7.2 on ggplot
Also, check out the ggplot cheat sheet from RStudio.

Homework 2: Heatmap using R

Attached Files:

File Coffee Chain data.csv (473.373 KB)
File HW2 Heat Map using R.pdf (118.954 KB)
File Skill Building Exercise for R.pdf (25.934 KB)

There are two parts to the homework.

Part I: Skill building exercise that you do not have to turn in. This is available in the attached file. I recommend that you do them individually.

Part II: An assignment that you will upload. One per team. It will be graded. The details are in the attached file. You will use the coffee chain data.The homework is to be done in teams, and upload a single submission for the team. Each of you should attempt the homework individually as well before meeting as a team. This will help you further build your skill.

Lecture 7: Introduction to Classification

Attached Files:
File 05 Intro to Classifiers.pdf (347.613 KB)
File bm.t.Rda (381 B)
File gain from splits.xlsx (11.961 KB)
File intro_rpart.R (849 B)
File intro_rpart.R.pdf (36.276 KB)
Textbook PF Chapter 3

Topics:

- What is classification? How is it used in business?
- The process of building, testing, validating, and using models
- Varieties of classifiers
- A simple tree classifier for bank marketing example
- Attribute choice for splitting
- Entropy, deviance, Gini and misclassification rate criteria
- Gain from splitting
- ID3 Classifiers and other Recursive Partition algorithms
- RPART package in R

Lectures 8 and 9: Evaluating Classifiers

Attached Files:
File eval.R.pdf (95.404 KB)
File bank-data.csv (33.753 KB)
File bm.raw.Rda (8.544 KB)
File bm.Rda (2.489 KB)
File 06 Evaluating Classifiers v2.pdf (277.37 KB)
Textbook PF Chapter 7

Topics:

- Dangers in interpreting re-substitution errors

- Laplace correction
- Introducing the full Bank IRA data set, clean up and discretization
- Splitting datasets into training and testing datasets (hold-out dataset)
- Building the tree model on the training dataset
- Creating the confusion matrix for training and test datasets
- K-fold cross-validation
- Performance metrics from information science
- Economic significance of recall, specificity, and precision
- Problems with unbalanced classes

Required Preparation for the Midterm

Attached Files:

File Blackboard – Uploading multiple files.pdf (204.688 KB)
File cc_test.Rda (71.209 KB)
File cc_test.xlsx (449.101 KB)
Every student must do this individually by 10/28 to make sure that their set up is working in preparation for the exam.

1. Download the two files cc_test.xlsx and cc_test.Rda.
2. Read cc_test.xlsx into Tableau. Make a simple visualization and save the workbook file with the following file name: first last.twb where first and last are your first, and last names, respectively.
3. Read the cc_test.Rda files into RStudio. It has the dataframe cc. Write two lines of code, one to examine the structure of cc and another to provide a summary of cc. Save the R SCRIPT file with a similar approach to naming, first last.R (the extension is .R)
4. Read the attached file: "Blackboard - Uploading Multiple Files" so you know how to upload multiple files. Essentially, do not submit unitl you have selected each file to upload by using the "Browse my computer" button more than once.
5. Log into Blackboard, and upload the two files (TWB and R script), save and submit the assignment.

Homework 3: Classification

Attached Files:

File churn.CSV (1.312 MB)
File HW3 Churn part I.pdf (75.768 KB)

Midterm Examinaton

Attached Files:

File cc_test.Rda (71.209 KB)
File cc_test.xlsx (449.101 KB)
File Rajiv.Profitability.Rda (618 B)
File Midterm Day Fall 2015 v2.pdf (630.541 KB)
File midterm exam day fall 2015 template v2.R (2.135 KB)

Lecture 9: Tableau - advanced visualization

Attached Files:

File Tableau Dashboard step-by-step.pdf (4.376 MB)

Recommended book by Jones (BJ) is useful for this topic.

Homework 4: Churn Part II

Attached Files:

File HW4 Churn part II.pdf (449.715 KB)

Due 11/23 extended to 11/25

Lecture 10: Avoiding Overfitting

Attached Files:

File 07 trees-pruning v2.R (5.015 KB)

File 07 trees-pruning v2.R.pdf (136.266 KB)

File 07 Avoiding Overfitting v2.pdf (361.332 KB)

Textbook (PF) Chapter 5

Topics:

- Effect of model flexiblility on training and test data set errors
- Components of MSE on test sets
- Variance and Square of Bias tradeoff
- Manual snipping
- Pre-pruning
- Post Pruning using complexity parameter

Lecture 11: Using Classifiers for Business Decision

Attached Files:

File 09 Using Classifiers for Business Decisions v2.pdf (237.489 KB)

File 09 roc curves v2.R (4.057 KB)

File 09 roc curves v2.R.pdf (106.715 KB)

Textbook PF Chapters 8 and 11

Topics:

Receiver Operating Characteristic (ROC) Curves

Constructing the ROC curve for a tree classifier

Determining cost of errors in a business setting

Finding the optimal threshold to minimize expected cost

Making predictions using the optimal threshold

ROCR package in R

Lecture 12: Bayesian Classifiers

Attached Files:

File ad.datax.Rda (dataset)   (127.789 KB)
File 10 Bayesian Classifiers v2.pdf   (248.299 KB)
File 10 naivebayes v2.pdf   (14.602 KB)
File 10 naivebayes v2.R   (3.371 KB)
File ad.datax.csv   (9.803 MB)

Topics:

- Bayesian classfiers
- How they work
- The Naive Bayes approach
- R package e1071
- Real world case: predicting ads on web pages
- Dealing with datasets with a large number of features
- Data disretization using rpart tree classifier
- Visualizing model output
- Advantages and disadvantages of Naive Bayes classifiers

Reading: Textbook PF Chapter 9

Lecture 13: Descriptive Analytics with Text

Attached Files:

File 11 textMining basic v2.pdf   (16.363 KB)
File 11 Text Descriptive Analytics v2.pdf   (372.857 KB)
File Mission.csv   (1.385 KB)
File 11 textMining basic v2.R   (4.941 KB)

Topics:

- Prevalence of textual data
- Structure in text data
- Natual Language Processing basics
- R packages tm and SnowballC
- Stopwords
- Cleaning text
- Bag-of-words approach
- Stemming
- Document Term Matrix
- Term frequency
- Wordcloud

Reading: Textbook Chapter 10

Hw 5: Tweet Challenge

Attached Files:

File HW5 Classifying Text.pdf (91.837 KB)

File Tweets.zip (16.675 KB)

Your grade will depend on how well you do on the set of mysterious evaluation tweets for which the class is not provided! See details in the attached files.

Lectures 14: Predictive Analytics with Text Data

Attached Files:

File Text Data for Class: Reviews.zip (3.739 MB)

File 12 textMiningNB v4.R.pdf (24.118 KB)

File 12 Text Predictive Analytics v2.pdf (274.179 KB)

File 12 textMiningNB v4.R (11.53 KB)

Textbook (PF) Chapter 10

Topics

- R packages tm and SnowballC
- Reading text files from sorted directories
- Associating classes with files
- Cleaning up text data: stopwords and stemming
- Effect of sparsity
- Naive Bayes model for detecting sentiment polarity
- TF and TFIDF
- Trading off Variance and square of Bias

Lectures 15 and 16: Dimension Reduction using Unsupervised Learning

Attached Files:

File student survey.csv (10.747 KB)

File 13 dim_reduction3.R.pdf (17.929 KB)

File 13 Dimension Reduction v3.pdf (335.371 KB)

File 13 dim_reduction3.R (6.915 KB)

File 13 Tableau Story.pdf (1,005.577 KB)

File clus.coord.Rda (1.454 KB)

File cor.ss.Rda (10.751 KB)

File means.ss.Rda (1.316 KB)

File pca.pve.Rda (538 B)

Textbook (PF) Chapter 6

Topics:

- Data: Student Satisfaction Survey
- Visual exploration of the data
- Feature extraction using Principal Components Analysis
- R function prcomp() for PCA
- Cleaning up data, dealing with missing values
- Scree plot
- Loadings and meaning of scores
- Clustering
- Use of clustering in marketing
- Kmeans algorithm
- R function cluster() and plotcluster()
- Clustering student satisfaction data
- Cluster membership

Preparation for the Final

Attached Files:
File b test.Rda (44.493 KB)
Download b test.Rda and make yourself familiar with it.

This is a data set is from a bank that is marketing Certificates of Deposits (CD). The data set has 11 features and a class variable for a total of 12 columns that are described in the table below. It has 4,521 observations.

Feature	Description
age	Age of customer in years (integer)
job	Job category, one of 12 possible values, (string)
marital	Marital status, one of 3 possible values, (string)
education	Level of education, one of 4 possible values, (string)
default	Whether the customer has previously defaulted on a loan (yes or no)
balance	Average daily balance (numeric)
housing	Whether the customer has a mortgage with the bank (yes or no)
loan	Whether the customer has a revolving loan credit with the bank (yes or no)
date	Date of last contact
contact	Method of last contact, one of three values (string)
duration	Duration of last contact, in minutes, (integer)
cd	Whether the customer purchased a CD or not (yes, or no), the class variable

Final Exam