

Data Mining

Instructor: Cesar Acosta-Mejia

Course Description

This course is about data analytics tools, methods, and applications. It focuses on data mining, Data Visualization, and Unsupervised Learning Methods. The course shows how to do feature engineering and how to reduce data complexity. Data visualization techniques are reviewed to find useful information from spatial data, now available from different online providers.

Unsupervised Learning Methods are used for clustering analysis, anomaly detection, and dimension reduction. The course reviews many unsupervised learning methods and shows how to apply them by means of case studies for model construction and evaluation.

The main computational tool is the R language. Libraries for statistical analysis, data visualization, and statistical learning are reviewed. RStudio is the interface of choice.

Prerequisite(s): None.

Recommended Preparation Expected to have knowledge of Engineering Statistics at the level of ISE 225 and working knowledge of a programming language.

Learning Objectives and Outcomes

In this course students learn to

- Preprocess dataframes (missing, duplicates, and data types)
- Understand the importance of Dimensionality Reduction.
- Apply Principal Components for Data Reduction.
- Apply clustering methods for unsupervised learning.
- Learn and apply Discriminant Analysis for classification.
- Use statistical learning Classification methods (Naïve Bayes, discriminant analysis)
- Apply Classification methods for unbalanced data.
- Apply data visualization tools on spatial data.
- Use association rules for mining market basket data.

Course Notes

The course material is available online.

Technological Proficiency and Hardware/Software Required

The R programming language and the RStudio IDE will be used.

Required Textbook

- James G., *An Introduction to Statistical Learning*, Springer, 2013 (ISLR)
ISBN 978-1-4614-7137-0

Supplementary Materials (References)

- Shmueli, *Data Mining for Business Analytics*, Wiley, 2018, ISBN 9781118879368
- Tan P., Steinbach M., Kumar V., *Introduction to Data Mining*, Second Edition, Pearson, 2018, ISBN 978-0133128901

Description and Assessment of Assignments

- Midterm** will be in-class based on the schedule and 2 hours length.
- Final Examination** a two-hour comprehensive exam scheduled by USC.
- Homework** are assigned every other week. Homework is based on the material of the previous and current week. Must be submitted by the due date, during the class session. No late homework to be accepted.

Grading Policy

Assignment	Points	% of Grade
Homework	100 each (6 homework assignments)	30
Midterm	100	30
Final	100	40
TOTAL		100

Grading Scale (Course final grades will be determined using the following scale)

A	95-100	B-	80-82	D+	67-69
A-	90-94	C+	77-79	D	63-66
B+	87-89	C	73-76	D-	60-62
B	83-86	C-	70-72	F	59 and below

Assignment Submission Policy

Assignments should be typewritten and clean. They should be submitted in class by the due date. Email submissions and late submissions are not allowed. No make-up exams are considered.

Timeline and Rules for submission

Assignments are to be returned the week after submission. Solutions will be released soon after the homework submission date.

Course Schedule: A Weekly Breakdown

	Date	Topics/Daily Activities	Homework	slides	R Files
1	Jan 9	Introduction to Data Mining for Descriptive and Predictive Analytics. Introduction to R , RStudio, and rmarkdown.	HW1 R base with rmarkdown	Overview DMining Rbase RStudio	intro Cars93 auto
2	Jan 16	Statistical Analysis with R – Random Variables The Multivariate Normal Distribution. Kernel Density Estimator.	HW1 due HW2 R base plotting	stat1	qq kernels
3	Jan 23	Statistical Analysis with R – Tests Test vs. Confidence Interval. Hypothesis Testing on k populations.	HW2 due	ht2	simulation 2pop kpop
4	Jan 30	Unsupervised Learning. Principal Components Analysis (PCA). Dimensionality Reduction, Feature Extraction.	HW3 PCA	slides2 10 largest.r cancer.csv	banknote
5	Feb 6	Unsupervised Learning. Clustering Methods. K-Means clustering. Hierarchical clustering	HW3 due	unsupervised kmeans hierarchical	simulated kmeans hierarchical usarrests correlationbased
6	Feb 13	Tidyverse R library (readr, tidyr, dplyr, stringr) Introduction to ggplot2	HW4 clustering	dplyr4 hierarchical	husarrests
7	Feb 20	Data Visualization. R libraries ggplot2, and extensions	HW4 due	ggplot	StudyArea ggs2, mpg
8	Feb 27	Midterm Exam			
9	Mar 6	Unsupervised Learning. Clustering Methods. Density-based Spatial Clustering (DBSCAN) Model-based Clustering. Mixtures.		db modelbased	simulation2 geyser contourpoints dbscan diabetes
10	Mar 13	Spring Break			
11	Mar 20	Classification – Part 1 KNN, Naïve Bayes.	HW5 Clustering	classification knn, nb	nbayes, tan bostonknn4
12	Mar 27	Classification – Part 2 Discriminant Analysis (linear, quadratic)	HW5 due	slides	admission wine
13	Apr 3	Classification – Part 3 Multinomial regression		slides	
14	Apr 10	Classifying Unbalanced Data. New metrics, Sensitivity, Specificity, False Positive rate (FPR), Recall, Precision. ROC Curve, and the AUC.	HW6	slides	roc5
15	Apr 17	Data Visualization. R library ggmap. Spatial and geographical visualization.	HW6 due	slides2	map22, map33 choropleths
16	Apr 24	Association Rules		rules	fplates groceries
16	TBD	Final Exam 2 p.m.			