

**ISE 535 Data Mining - Section 598**

Units: 4

Spring2024 - MW 2:00 p.m.

**Location:** GFS 207

**Instructor: Cesar Acosta**

**Office:** GER 216

**Office Hours:** T 3 p.m. (online)

**Contact Info:** [acostame@usc.edu](mailto:acostame@usc.edu)

**Teaching Assistant:** TBD

**Office:** online

**Office Hours:** TBD

**Contact Info:** TBD

**Teaching Assistant:** TBD

**Office:** online

**Office Hours:** TBD

**Contact Info:** TBD

**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 on Blackboard.

**Technological Proficiency and Hardware/Software Required**

The R programming language and RStudio IDE will be used.

**Required Textbook**

This course does not require a textbook.

**Supplementary Materials** (References)

* James G., *An Introduction to Statistical Learning*, Springer, 2013 (ISLR)

ISBN 978-1-4614-7137-0

* Shmueli, *Data Mining for Business Analytics,* Wiley, 2018, ISBN 9781118879368
* Tan P., Steinbach M., Kumar V., *Introduction to Data Mining,* 2ed.,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

A- 90-94

B+ 87-89

B 83-86

B- 80-82

C+ 77-79

C 73-76

C- 70-72

D+ 67-69

D 63-66

D- 60-62

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.

**535-598 Course Schedule: A Weekly Breakdown**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Date | **Topics/Daily Activities** | **Homework** | **Files** | **R Files** |
| 1 | Jan 8 - **Jan 10** | **Introduction to Data Mining** forDescriptive and Predictive Analytics.  **Introduction to R,** RStudio**,** and rmarkdown. | HW1 R base with rmarkdown | Overview  DMining  Rbase  RStudio | Exercise1  Cars93  auto |
| 2 | Jan 15(recorded) - Jan 17 | **Statistical Analysis with R – Random Variables**  The Multivariate Normal Distribution. Kernel Density Estimator. | HW1 due  HW2 R base plotting | stat2 | qq2  kernels |
| 3 | Jan 22 -Jan 24 | **Statistical Analysis with R – Tests**  Test vs. Confidence Interval.  Hypothesis Testing on *k* populations. | HW2 due | ht2 | simulation  2pop  kpop |
| 4 | Jan 29 -Jan 31 | **Unsupervised Learning.**  Principal Components Analysis (PCA). Dimensionality Reduction, Feature Extraction. | HW3 PCA | pca, examples | Banknote  Stockreturns  universities |
| 5 | Feb 5 -Feb 7 | **Unsupervised Learning.**  Clustering Methods.  K-Means clustering.  Hierarchical clustering | HW3 due | unsupervised  kmeans  hierarchical | simulated  kmeans  hierarchical  husarrests |
| 6 | Feb 12 -Feb 14 | **Tidyverse** R library (readr, tidyr, dplyr, stringr)  **Data Visualization** libraryggplot2 | HW4 clustering | dplyr4  ggplot | StudyArea  ggts2, mpg |
| 7 | Feb 19(recorded)-- Feb 21 | **Unsupervised Learning.** Clustering Methods. Density-based Spatial Clustering (DBSCAN)  Model-based Clustering. Mixtures. | HW4 due | dbscan  modelbased | simulation2  geyser, dbscan  contourpoints  diabetes |
| 8 | Feb 26 -Feb 28 | **Midterm Exam** |  |  |  |
| 9 | Mar 4 – Mar 6 | **Unsupervised Learning.** Mining marketing data. Association Rules. Performance measures |  | rules | fplates  groceries |
| 10 | Mar 11-Mar 13 | **Spring Break** |  |  |  |
| 11 | Mar 18 -Mar 20 | **Classification – Part 1**. Entropy vs Gini index.  KNN, Naïve Bayes. | HW5 Clustering | classification  knn1, nbayes2 | nbayes, tan,  nb2.csv  bostonknn4 |
| 12 | Mar 25 -Mar 27 | **Classification – Part 2**  Discriminant Analysis (linear, quadratic) | HW5 due | slides | admission  wine |
| 13 | Apr 1 - 3 | **Classification – Part 3.** Rule learners. |  | slides |  |
| 14 | Apr 8 -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 15 -Apr 17 | **Data Visualization.** R library ggmap.  Spatial and geographical visualization. | HW6 due | slides2 | map22, map33  choropleths |
| 16 | Apr 22 -Apr 24 | **Review** |  |  |  |
|  | TBD | **Final Exam 2 p.m.** |  |  |  |

**Statement on Academic Conduct and Support Systems**

**Academic Conduct:**

Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, “Behavior Violating University Standards” [policy.usc.edu/scampus-part-b](https://policy.usc.edu/scampus-part-b/). Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, [policy.usc.edu/scientific-misconduct](http://policy.usc.edu/scientific-misconduct).

**Support Systems:**

*Counseling and Mental Health - (213) 740-9355 – 24/7 on call*

[studenthealth.usc.edu/counseling](https://studenthealth.usc.edu/counseling/)

Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

*National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call*

[suicidepreventionlifeline.org](http://www.suicidepreventionlifeline.org/)

Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

*Relationship and Sexual Violence Prevention and Services (RSVP) - (213) 740-9355(WELL), press “0” after hours – 24/7 on call*

[studenthealth.usc.edu/sexual-assault](https://studenthealth.usc.edu/sexual-assault/)

Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

*Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298*

[equity.usc.edu](https://equity.usc.edu/), [titleix.usc.edu](http://titleix.usc.edu)

Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following *protected characteristics*: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations. The university also prohibits sexual assault, non-consensual sexual contact, sexual misconduct, intimate partner violence, stalking, malicious dissuasion, retaliation, and violation of interim measures.

*Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298*

[usc-advocate.symplicity.com/care\_report](https://usc-advocate.symplicity.com/care_report/)

Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

*The Office of Disability Services and Programs - (213) 740-0776*

[dsp.usc.edu](http://dsp.usc.edu/)

Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

*USC Support and Advocacy - (213) 821-4710*

[uscsa.usc.edu](https://uscsa.usc.edu/)

Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

*Diversity at USC - (213) 740-2101*

[diversity.usc.edu](https://diversity.usc.edu/)

Information on events, programs and training, the Provost’s Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

*USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call*

[dps.usc.edu](http://dps.usc.edu/), [emergency.usc.edu](http://emergency.usc.edu/)

Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

*USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call*

[dps.usc.edu](http://dps.usc.edu/)

Non-emergency assistance or information.