# Course Syllabus for DS 740: Data Mining

**NOTE:** This syllabus document contains the basic information of this course. The most current syllabus is available in the full course.

## Course Description

In this course, you'll learn data mining and machine learning methods for supervised and unsupervised learning. Topics include k-nearest neighbors, predictive artificial neural networks, clustering algorithms, and ensemble methods such as random forests and XGBoost. We'll emphasize how to implement these methods in R and interpret the results, but we'll also discuss enough of the theory so you can understand why and how they work.

Prerequisites: DS 705 and DS 710.

## Course Objectives

By the end of this course, you will be able to:

- Compare and decide among methods of data mining.
- Interpret a model, and select and create graphs to support your interpretation.
- Use cross-validation for model selection or assessment.
- Use nested cross-validation for honest assessment of a model selection procedure.
- Implement the following machine learning methods in R:
    - Multiple linear regression
    - Logistic regression
    - K-nearest neighbors for categorical and quantitative response variables
    - Robust regression
    - LDA and QDA
    - Decision Trees
    - Random Forests
    - XGBoost
    - Penalized regression (LASSO, Ridge Regression, and Elastic Net)
    - Support Vector Classifiers and Support Vector Machines
    - Artificial Neural Networks
    - Hierarchical and Non-hierarchical Clustering
    - Principal Components Analysis
    - Association Rules

## Course Components

**Homework:**

To give you a chance to practice the data mining skills you're learning, 11 homework assignments will be due throughout the term. Each homework assignment will involve

programming in R, and some also involve writing short statements interpreting your results.

You are encouraged to communicate about homework problems with your classmates via the discussion board. However, all work that you turn in must be your own and you must fully understand what you write. In particular, you must type your own code and write your own interpretations.

Homework Retake Policy: You can have up to two attempts at each homework assignment. If you submit a homework assignment twice, the second version of each of your short-answer responses is the one that will be graded.

**Lessons and Participation:**

To help you learn the programming skills in this course, we have prepared three types of learning activities:

- Presentations: These are narrated slides and/or demonstrations of how to implement and interpret various data mining techniques. Please watch these and take notes as appropriate. If the presentation includes self-check or review questions, these will be graded on participation only.
- Readings: These have been selected from the textbook and from websites we have found to be particularly helpful for understanding both data mining in general and the specifics of data mining in R. Please read these and take notes as appropriate. If the reading includes exercises, the exercises are not required, although you are welcome to try them if you'd like extra practice.
- WeBWorK Activities: These are questions designed to guide you through understanding and implementing various data mining algorithms. They are designed to complement the readings and presentations, and to prepare you for the homework and the projects. They are implemented in WeBWorK, which gives you automatic, instant feedback about your answers. You must do the sample problems as they count towards your participation grade. The problems will be graded on correctness. However, you are encouraged to try each problem as many times as you need until you get it right! This will help you understand the material better so you can do your best on the homework.

  You are encouraged to work on the sample problems with an R console open next to WeBWorK, so you can go back and forth between testing code and answering questions about it. You are also encouraged to read all of the text in the problem carefully, and take notes on it as you would a reading in the textbook. For some problems, a hint will be made available after you submit an answer to the problem.

  Each set of sample problems can be accessed via a link in this online course.

**Midterm Project:**

The midterm project will provide you with an opportunity to consolidate your understanding of the techniques from the first half of the semester. You'll apply two data mining techniques from the course to a new set of real data, and prepare a non-technical summary of your analysis and its interpretation. You'll also practice properly validating your modeling process, with some structured guidance.

**Final Project:**

The final project is your opportunity to apply what you have learned in this course to answer a question that interests you, by analyzing a real-world data set of your choice and writing a professional summary of your analysis.

## Grading

Your mastery of course content is assessed using a variety of methods:

| **Activity** | **Percentage** |
|---|---|
| Homework Assignments (equally weighted) | 55% |
| Completion of required learning activities (WebWork) | 10% |
| Midterm Project | 10% |
| Final Project | 25% |

Final grades are assigned using the following scale:

| | |
|---|---|
| **90–100%** | A |
| **80–89%** | B |
| **60–79%** | C |
| **0–59%** | F |