Course Syllabus
Instructor: R J Greene, PhD

email: Rich.Greene@usnh.edu

Text: Data Mining by Witten et al 3rd edition

Link to textbook by WittenLinks to an external site.

Course Administration

0. I assume you are taking this class for vocational reasons - you want to get a higher paying job and more fulfilling than just coding. This means we must develop skills you currently do not have. Mathematics is the language of science and we cannot hide from it or apply it without understanding. For many of you, you will be learning and applying new ideas and skills. This isn't always pleasant and relaxing

1. This course will involve computer programming. If you do not know a computer language well enough to write relatively simple programs, this course is not for you.

2. All work - programming and homework - must be your own - no group work at all. If I catch you cheating, you will be reported to the CS Department for discipline. I have caught several students in the midterm and final cheating using ChatGPT and the result was they got into quite a lot of trouble with the University and lost all hope of getting a letter of recommendation from me. And for what? A ‘good’ grade that now will not matter at all to them!

3. I assume you come to class to hear what I have to say about the lesson at hand. If you must talk or otherwise not listen, the answer is simple: just don't come to class or simply leave the classroom and talk all you wish. Any talking during class is disruptive to those trying to learn and take notes.

4. Please read the assigned work BEFORE coming to class. Class will NOT simply be a rehash of the reading assignment but will assume you have read everything and start from that point.

5. Do NOT submit homework or final late! In the real world, tardiness is simply not tolerated and will destroy your career. This class is intended to force the habit of getting work done early and assuming there will be a lot of unforeseen difficulties>

Grading

Midterm 50%

Final 50%

Primary Text: Witten, Frank, Hall. Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition)

Duration: 11–12 weeks

Format: Weekly 2–3 hour lecture + lab session

Tools: Weka (primary), Python/Scikit-learn (secondary)

Assessment: Midterm 50%, Final 50%

Week 1 – Introduction to Data Mining & Machine Learning

Readings: Ch. 1 (What is Data Mining?)
Topics:
Definitions, scope, and applications of data mining
Data mining process & CRISP-DM
Overview of supervised vs. unsupervised learning
Weka interface walkthrough
Homework:
Install Weka and load the iris and weather datasets.
Perform a simple classification using J48 decision tree.
Write a 1-page reflection on at least 3 real-world applications of data mining with respect to your interests. This is NOT to turn in! Thois is for you to begin looking at DM opportunities for projects or work!

Week 2 – Input: Concepts, Instances, Attributes

Readings: Ch. 2 (Input: Concepts, Instances, Attributes)
Topics:
Data formats & types (nominal, numeric, string, date)
Attribute representation and scaling
Missing values & noise
Data exploration basics in Weka
Homework:
Load contact-lenses dataset, handle missing values, and analyze attribute types.
Compute summary statistics for numeric attributes (Weka & Python).
Short answer: Explain why attribute scaling matters for certain algorithms.

Week 3 – Output: Knowledge Representation

Readings: Ch. 3 (Output: Knowledge Representation)
Topics:
Representations: decision trees, classification rules, linear models
Trade-offs in interpretability vs. accuracy
Visualizing learned models in Weka
Homework:
Train J48 and PART models on weather.nominal dataset.
Compare outputs: Which is more interpretable? Why?
Include screenshots and written interpretation.
Week 4 – Simple Classification & Probabilistic Models

Readings: Ch. 4 (Simple Classifiers)
Topics:
ZeroR, OneR, Naïve Bayes, k-Nearest Neighbor
Evaluating simple baselines
Hands-on: credit-g dataset in Weka
Homework:
Train and evaluate Naïve Bayes and IBk (k-NN) classifiers.
Compare to ZeroR baseline.
Write a 2-page analysis on why simple models can sometimes perform surprisingly well.
Week 5 – Trees & Rules

Readings: Ch. 6 (Trees and Rules)
Topics:
Decision tree induction (ID3, C4.5/J48)
Rule extraction & pruning
Case study: weather & contact-lenses datasets
Homework:
Train J48 and RandomTree models.
Export rules and evaluate accuracy.
Discuss overfitting signs in your models.
Week 6 – Model Evaluation & Overfitting

Readings: Ch. 5 (Evaluation)
Topics:
Cross-validation, hold-out, bootstrap
Confusion matrix, ROC curves, AUC
Bias-variance tradeoff
Homework:
Evaluate credit-g dataset classifiers using 10-fold CV and percentage split.
Plot ROC curves for Naïve Bayes and J48.
Short essay: How do you decide between two models with similar accuracy
Week 7 – Numeric Prediction

Readings: Ch. 7 (Numeric Prediction)
Topics:
Linear regression, model trees (M5)

Evaluation with RMSE and MAE
Applications to real datasets
Homework:
Use CPU.arff dataset to train linear regression and M5 model tree.
Compare errors (RMSE, MAE).
Discuss interpretability differences.

Week 8 – Data Transformations & Attribute Selection

Readings: Ch. 8 (Data Transformations), Ch. 10 (Attribute Selection)
Topics:
Normalization, standardization, discretization
Feature selection methods (filter, wrapper)
PCA in Weka
Homework:
Apply PCA to iris dataset, keeping top 2 components.
Train J48 before and after PCA; compare results.
Short answer: Why might dimensionality reduction help with noisy data?

Week 9 – Clustering & Association Rules

Readings: Ch. 9 (Clustering), Ch. 4 sections on Association Rules
Topics:
k-Means, EM clustering
Apriori algorithm for association rules
Interpreting clusters & rules
Homework:
Cluster iris dataset with k-Means; compare to actual classes.
Run Apriori on a market basket dataset.
Submit top 5 interesting rules (with justification).

Week 10 – Ensemble Methods

Readings: Ch. 16 (Bagging, Boosting, Random Forests)
Topics:
Bagging, AdaBoost, Random Forests
Why ensembles work
Trade-offs: accuracy vs. interpretability
Homework:

Train Random Forest and AdaBoost on credit-g.
Compare to single decision tree performance.
Discuss why ensembles often outperform individual models.
Week 11 – Data Mining in Practice & Final Presentations

Readings: Ch. 17 (Practical Machine Learning), Ch. 19 (Lessons Learned)
Topics:
Data mining workflow from raw data to deployment
Common pitfalls & best practices
Ethical considerations in data mining
Homework:
Submit final project report and presentation: A full data mining case study using a dataset of your choice (Weka or Python).
Include: problem definition, data preprocessing, model selection, evaluation, and interpretation.