Lecture 1:
- Survey
- Course Overview
- Basic Concepts and Terms

Homework:
- GitHub account
- Piazza
- Participation Exercise 1
**Course Logistics**

1. **Course webpage:** [www.cs.gonzaga.edu/bowers/courses/cpsc322](http://www.cs.gonzaga.edu/bowers/courses/cpsc322)

2. **Piazza:** for Q&A, announcements (see invite)

3. **GitHub:** for submitting homework

4. **Canvas:** for tracking points

5. **Office Hours:** Tu/Th 3:30–4:30, Wed 1:30–3:30

6. **Grader:** Zach Burge

7. **HW Late Policy:** Up to 1-week late, 20% penalty (email grader)

8. **Grading:** Out of 700 possible points:
   - 8 homework assignments at 35 points each 280 points
   - Final project 70 points
   - 10 quizzes at 10 points each 100 points
   - 2 exams at 100 points each 200 points
   - Attendance (approximately 25 lectures) 50 points

Must score at least 60% on homework (assignments + final project) and 60% on tests (quizzes + exams)

See webpage for syllabus, weekly schedule, homework assignments, etc.
Course Overview: General Ideas & Terms

Basic types of data analytics:

**Descriptive:** understand past and/or current trends
- product sales by month over last two years

**Diagnostic:** understand why past and/or current trends occur
- relationships between product sales and advertising campaigns

**Predictive:** predict future trends
- product sales for next year given anticipated trends
- note: usually based on ability to predict historical data

**Prescriptive:** predict good courses of action / what decisions to make
- place orders for certain amounts of specific products

Main focus of this class is on classic algorithms for predictive analytics

High-level tasks involved in data science:

- Data discovery and selection
- Data engineering (extract, transform, load, catalog)
- Data analysis (exploration, summary stats, feature selection)
- Model development (feature engineering, model training, model evaluation)
- Model engineering (deployment, monitoring)
- Communicate Insights (reports, dashboards, results)
Categorical vs Continuous Values:
  Categorical: values denoting finite categories (e.g., t-shirt sizes, likert values)
  Continuous: generally refers to numeric values (quantities vs categories)
  Note: various approaches for converting from continuous to categorical data

Classification vs Regression Models:
  Classification: predictions to a categorical-valued class “label”
  Regression: predictions to a continuous-valued function (e.g., linear, logistic)

Supervised vs Unsupervised Learning:
  Supervised: learn from already “labeled” data ... i.e., “ground truth” is known
  Unsupervised: find patterns in unlabeled data, sometimes via “cost” function

Third main type of learning is “Reinforcement Learning” (reward functions)

Structured vs Unstructured Data:
  Structured: often implies tables of rows (“instances”) and columns (“features”)
  Unstructured: not naturally tabular data such as text, images, graphs
  Note: Semi-Structured mixes structured and unstructured (e.g., JSON, XML)

Course Goals:
  • Deep(er) dive into classic algorithms (“greatest hits”) for predictive analytics
  • Focus on structured data and classification
  • Focus (mainly) on supervised machine learning
  • Gain more experience programming in Python (implementing algorithms)
Main Course Topics: General overview of the class schedule

1. Python review ... ≈ 2 weeks

2. Data pre-processing ... ≈ 3 weeks
   – Data representation
   – Data preparation (combing and “cleaning” data sets)
   – Data exploration: summarization, visualization

3. Basic predictive analytics ... ≈ 3 weeks
   – Terminology
   – Simple regression (linear) vs classification ($k$-NN)
   – Evaluation approaches (and throughout semester) ... (*)

4. Classic supervised classification algorithms ... ≈ 4 weeks
   – Naive Bayes, decision trees, random forests, ensemble approaches

5. Classic unsupervised classification algorithms ... ≈ 2 weeks
   – $k$-means clustering, association rule mining

6. Basic supervised artificial neural networks (as time allows) ... ≈ 1 week

(*) We’ll also mix in topics related to:
   • explainability and trust
   • data bias and ethical considerations
Course Expectations (also see Syllabus):

• Engage and participate in class (including doing your own work)
• Start assignments early, give yourself enough time to succeed
• Assume you have everything you need (ask when in doubt)
• Study for quizzes, exams, etc
• Come to office hours and frequently check piazza

(Some) collaboration is encouraged ... e.g., discussing assignments

• but avoid plagiarism & other related issues — e.g., no code sharing

Homework will largely be to implement algorithms discussed in class

• discussion of approaches/algorithms will be higher level
• then your job will be to implement (with some hints / design restrictions)

Warning: This course involves a good amount of programming!

Check-In:

• What is meant by “predictive analytics”?
• How are predictive and diagnostic analytics related?
• What is meant by building a “classifier” using “supervised learning”?
• Give an example of data that would be considered “unstructured”.
Participation Exercise #1:

1. Create a GitHub account if you don’t already have one.
2. Log in to piazza for the course and ask one question related to today’s lecture.
3. Download and install miniconda on your laptop (the latest version)
   - Info: https://docs.conda.io/en/latest/miniconda.html
   - To test, from a terminal run conda -h

For Thursday, hand in a hardcopy with:

1. Your name
2. Your github account name
3. Issues you had with piazza (if any)
4. Whether you were able to successfully install miniconda