CSC 390
Topics in Artificial Intelligence

“Unsupervised Machine Learning”

Fall 2016
Prof. Sara Mathieson
Smith College
Outline: 9/8

• Introductions
• Syllabus and course overview
• What can we do with unsupervised learning?
• Classical AI example
• Crash course on supervised learning
Introductions
To discuss with a partner:

1) Do you think we as humans learn in a “supervised” or “unsupervised” way? (thinking about these words in a non-scientific sense)
To discuss with a partner:

1) Do you think we as humans learn in a “supervised” or “unsupervised” way? (thinking about these words in a non-scientific sense)

2) How would you identify a leaf?
To discuss with a partner:

1) Do you think we as humans learn in a “supervised” or “unsupervised” way? (thinking about these words in a non-scientific sense)

2) How would you identify a leaf?

3) Also discuss what you hope to get out of this course.
Identification options:

- Go through a nature guide until you find a match
  - Issues?
Identification options:

• Go through a nature guide until you find a match
  o Issues?

• Use “features” (coniferous vs. deciduous, type of lobes or tips, waxiness, color, etc)
  o Issues?

Image: keys.lucidcentral.org
Identification options:

- Go through a nature guide until you find a match
  - Issues?

- Use “features” (coniferous vs. deciduous, type of lobes or tips, waxiness, color, etc)
  - Issues?

- Collect tons and tons of leaves and cluster them somehow
  - Issues?
Identification options:

• Go through a nature guide until you find a match
  o Issues?

• Use “features” (coniferous vs. deciduous, type of lobes or tips, waxiness, color, etc)
  o Issues?

• Collect tons and tons of leaves and cluster them somehow
  o Issues?
Identification options:

- Go through a nature guide until you find a match
  - Issues?

- Use “features” (coniferous vs. deciduous, type of lobes or tips, waxiness, color, etc)
  - Issues?

- Collect tons and tons of leaves and cluster them somehow
  - Issues?
Identification options:

- Go through a nature guide until you find a match
  - Issues?

- Use "features" (coniferous vs. deciduous, type of lobes or tips, waxiness, color, etc)
  - Issues?

- Collect tons and tons of leaves and cluster them somehow
  - Issues?
Syllabus
Senior Seminar

- Capstone experience that ties together what you have learned in CS (and other courses) so far

- Focus on effective scientific communication
  - Writing
  - Discussions
  - Oral presentations

- Individual research projects

- Learning to read scientific literature

- Due to the course style, enrollment is limited
Prerequisites

• CSC 111, Introduction to Computer Science
• MTH 111, Calculus 1
• MTH 220 or another intro statistics course
• A 200-level computer science course
• Linear algebra helpful but not required
Class Meetings

- Interactive lecture (slides + board)
- Small in-class labs (not usually turned in, but often homeworks will build on labs)
- Paper discussions or presentations
Assignments

• Homeworks: programming (Python), pencil-and-paper, mid-semester presentation (15-20min)
  ○ 40%
Assignments

• Homeworks: programming (Python), pencil-and-paper, mid-semester presentation (15-20min)
  o 40%

• Midterm assignment (usually a take-home exam)
  o 20%
Assignments

• Homeworks: programming (Python), pencil-and-paper, mid-semester presentation (15-20min)
  o 40%

• Midterm assignment (usually a take-home exam)
  o 20%

• Final project presentation and writeup
  o 30%
Assignments

- Homeworks: programming (Python), pencil-and-paper, mid-semester presentation (15-20min)
  - 40%

- Midterm assignment (usually a take-home exam)
  - 20%

- Final project presentation and writeup
  - 30%

- Participation (in-class discussion, labs, Piazza)
  - 10%
Resources

- Textbook (free online!)
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction

- Piazza for online discussion, announcements, etc
  - https://piazza.com/smith/fall2016/csc390/home
Resources

• Spinelli Center for Quantitative Learning
  
  https://www.smith.edu/qlc/

• Disability Services
  
  https://www.smith.edu/ods/
Software (Python)

Packages:

• numpy
• scipy
• matplotlib
• sklearn

Enthought Canopy:

https://store.enthought.com/downloads/#default
Tentative Topics

• Overview of AI
• Supervised vs. unsupervised learning
• Key methods in supervised learning
• Clustering (k-means, hierarchical, UPGMA)
• Principal components analysis (PCA)
• Non-negative matrix factorization
• Autoencoders
• Graphical models and latent variables
• Topic modeling
• Natural Language Processing (NLP) applications
Tentative Topics

- Expectation-maximization (EM)
- Hidden Markov models (HMM)
- Combining unsupervised and supervised learning
- Neural networks and deep learning
- Deep learning application: image identification
Course Policies

1) Email: use Piazza for all questions that might be relevant to others in class
Course Policies

1) **Email**: use Piazza for all questions that might be relevant to others in class

2) **Sending code**: do not send long blocks of code
Course Policies

1) **Email**: use Piazza for all questions that might be relevant to others in class

2) **Sending code**: do not send long blocks of code

3) **Late work**: one 3-day extension, no other late work
   Exceptions: accommodations letters, notice from Dean or Heath Services
Course Policies

1) **Email**: use Piazza for all questions that might be relevant to others in class

2) **Sending code**: do not send long blocks of code

3) **Late work**: one 3-day extension, no other late work
   Exceptions: accommodations letters, notice from Dean or Heath Services

4) **Electronic devices**: fine in class as long as directed toward class material
Course Policies

1) **Email**: use Piazza for all questions that might be relevant to others in class

2) **Sending code**: do not send long blocks of code

3) **Late work**: one 3-day extension, no other late work
   Exceptions: accommodations letters, notice from Dean or Heath Services

4) **Electronic devices**: fine in class as long as directed toward class material

5) **Attendance**: two missed classes without effect
“Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations. All submitted work of any kind must be the original work of the student who must cite all the sources used in its preparation.”
Examples of Unsupervised Learning
Unsupervised learning: HMM

Modern humans, Neanderthal, Denisova

The complete genome sequence of a Neanderthal from the Altai Mountains, Prüfer et al (2014)
Unsupervised learning: PCA

Genes mirror geography within Europe (2008)
Example from Classical AI
Decision Trees

• We could make a decision tree for our leaf example, or a diagnostic example
Decision Trees

• We could make a decision tree for our leaf example, or a diagnostic example

• What are the advantages/disadvantages?
Decision Trees

• We could make a decision tree for our leaf example, or a diagnostic example

• What are the advantages/disadvantages?

• Modern machine learning makes use of theory and statistics to make principled inference
Supervised Learning: makes use of examples where we know the underlying "truth" (sometimes called a label)

Unsupervised Learning: Learn underlying structure or features without labeled "training" data
### Supervised Learning:
makes use of examples where we know the underlying "truth" (sometimes called a label)

### Unsupervised Learning:
Learn underlying structure or features without labeled "training" data
Crash Course on Supervised Learning
Supervised Learning

- Labels/outputs are quantitative (regression)
- Labels/outputs are qualitative (classification)
Example data with 3 classes

Question: how to classify a new data point?
Nearest Neighbor

Kind of like a guidebook. Disadvantages?

Image: CS231n Machine Learning at Stanford
5-Nearest Neighbor

Often more robust. Disadvantages?
FIGURE 2.2. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.
Unsupervised Nearest Neighbor?
Please turn in notecards!