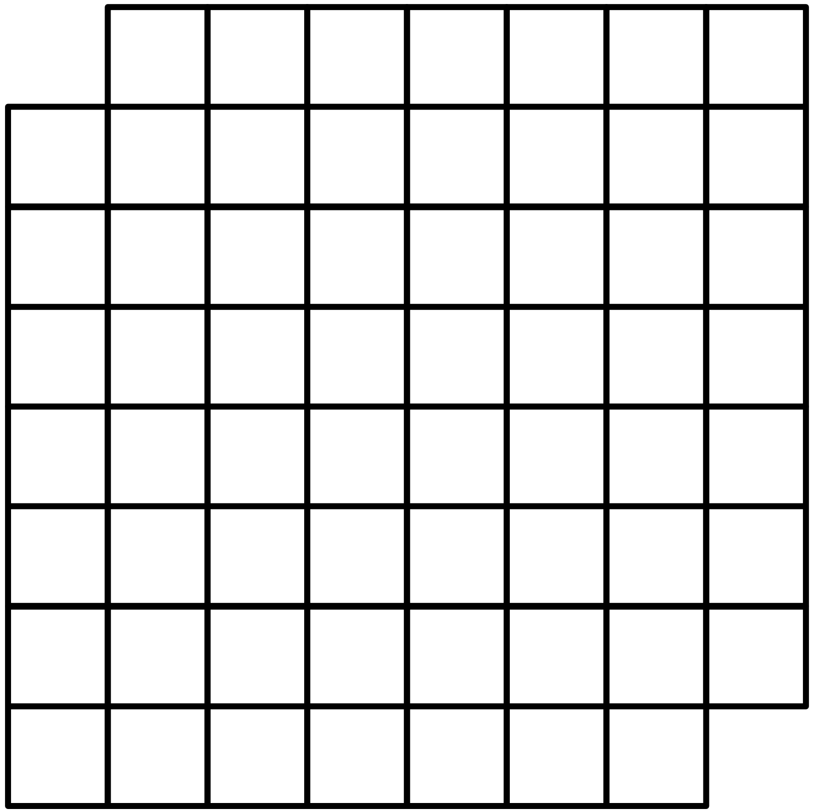
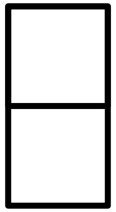
Warm-Up

Can you fill a 8×8 board with the corners missing using dominoes?

Can you tile this?



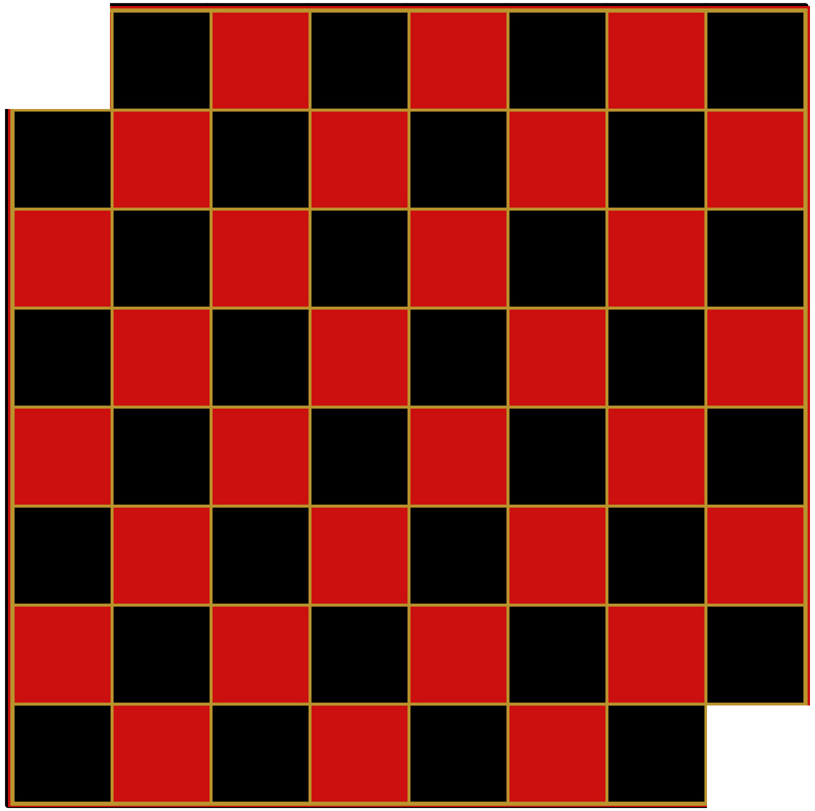
With these?



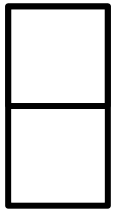
Warm-Up

Can you fill a 8×8 board with the corners missing using dominoes?

Can you tile this?



With these?



CS 3100
Data Structures and Algorithms 2
Lecture 24: Intro to Machine Learning, Part 1

Co-instructors: Robbie Hott and Ray Pettit

Spring 2024

Announcements

- PA5 due next Tuesday; PS10 due tomorrow
- Quizzes 3-4 almost graded!
- Quiz 5 (+ retakes) **next Thursday 5/2 at 7pm**
  - If you have SDAC, please schedule ASAP
  - Review Session: More information coming soon!
- Office hours updates
  - Prof Hott Office Hours:
    - Mondays 11a-12p, Friday 10-11am, 2-3pm

What Is Machine Learning?

- Some definitions:
  - Wikipedia: “statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions”
  - Oxford Languages: “the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data”
- An area that’s part of (or associated with) Artificial Intelligence
- But also Data Science
- Many methods based on statistical techniques
  - So students can find ML taught in CS, DS and Stats departments
  - And applications taught in engineering, science and business courses

Different Ways that Humans Learn

- Rote Learning (memorization)
- Learn by finding patterns
  - Unsupervised Learning
- Learn by example
  - Supervised Learning
- Learn by practice with feedback
  - Reinforcement Learning

Rote Learning

- Memorizing what is learned
  - Difficult to transfer knowledge to different domains or different problems
  - Easy (ish) to recall the information
  - Hard to apply the information

- Computers are GREAT at this
  - Storing information and retrieving it
  - `String knowledge = “America declared independence in 1776”;`
  - The computer is learning (in some sense)
  - …but there are big limitations with what it can do with that knowledge

Learning by Observing Similarities

- What if we are looking at pictures of animals, but aren't given the ground truth of what animal is in which picture?

- We can still group similar photos together
  - E.g., these are all birds, but which are ravens, crows, or starlings?

- This is called **clustering**,
  and is a form of <u>**unsupervised learning**</u>

Image created from: The Caltech-UCSD Birds-200-2011 Dataset, https://authors.library.caltech.edu/records/cvm3y-5hh21

Learning by Example

- If I show you a bunch of examples of a new species of animal, you'll know how to recognize that animal
- I didn't explicitly teach you anything, so how did you do it?

- This is called **<u>Supervised Learning</u>** because a supervisor (teacher, etc.) is telling you the answer to many examples
  - We need "training" on a set of data of known examples

- Then, hopefully you can perform the task independently afterward
  - Our training is used to give answers for new examples

Learning by Doing

- Learn to do something by practicing and retrying until you get better
  - Painting, math, sports, etc.
  - We get feedback about performance that we use to improve

- How can computers do this?
- This is called **Reinforcement Learning** and is a well-studied area

In our CS3100 intro to ML you'll see…

1. This overview, and how many ML algorithms work with data
2. Clustering as an example of unsupervised learning
   – Two algorithms, including one you've seen
3. A simple classification algorithm to show supervised learning
   – This algorithm only uses concepts you've learned already
   – Then a brief overview of other techniques
4. A brief intro to reinforcement learning

TAXONOMY OF MACHINE LEARNING METHODOLOGIES

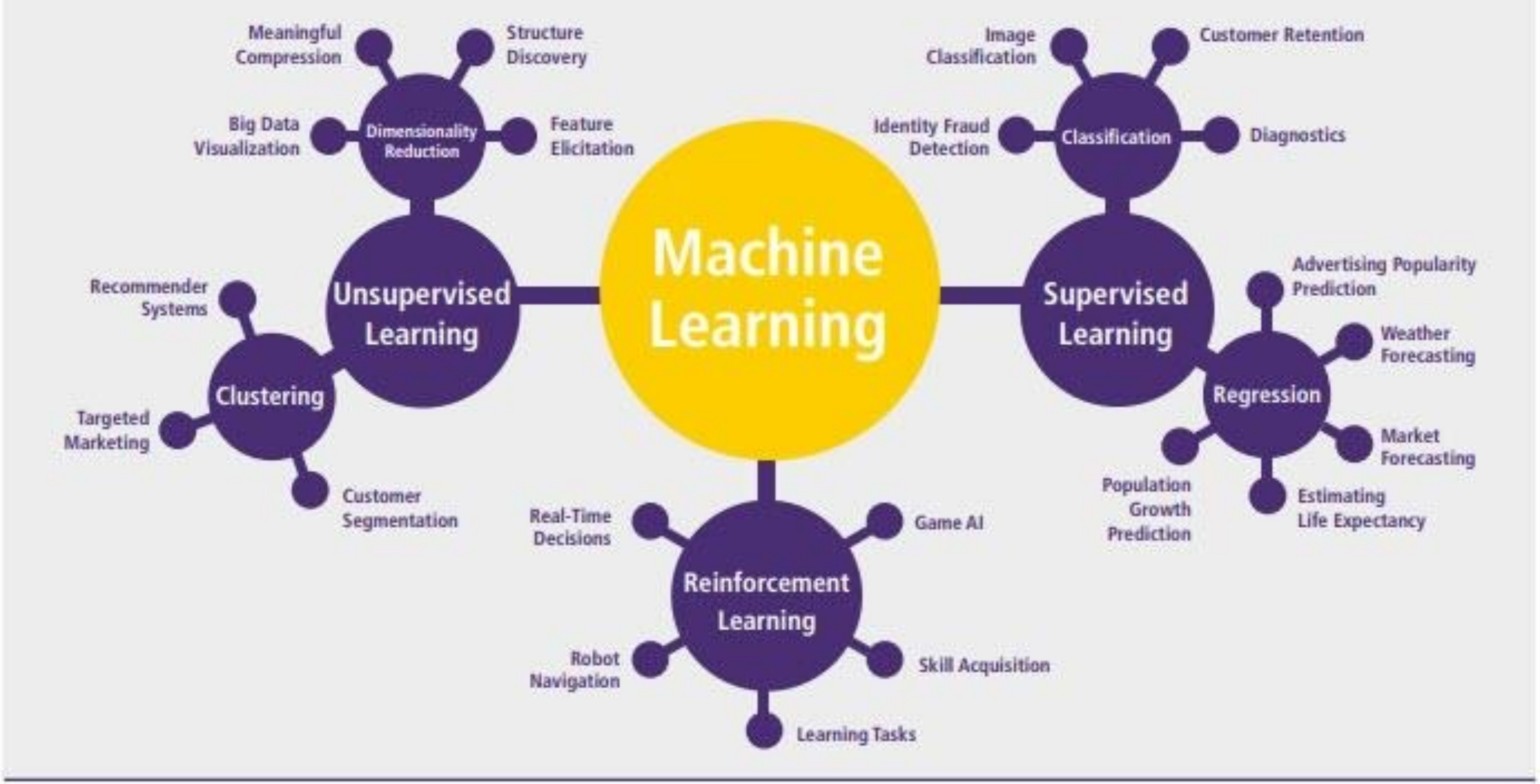


Figure 10: An overview of machine learning techniques; **Source:** Jha, V.

Image from:: https://www.forbes.com/sites/louiscolumbus/2018/06/11/10-ways-machine-learning-is-revolutionizing-supply-chain-management/?sh=27f405dc3e37

Where We Are Now

1. This overview, and **<u>how many ML algorithms work with data</u>**
2. Clustering as an example of unsupervised learning
   – Two algorithms, including one you've seen
3. A simple classification algorithm to show supervised learning
   – This algorithm only uses concepts you've learned already
   – Then a brief overview of other techniques
4. A brief intro to reinforcement learning

Data for Machine Learning

- ML techniques use data for individual observations (items, examples) to build a model of that set of data
  - Explain the data in some way
  - Use the data to make predictions
- For each observation, there are a set of **features** or attribute values
  - Measurements or observations for that example

Feature Examples

- Imagine a data set for patients at risk for a particular type of cancer
- Possible features
  - Weight, body-mass index, cholesterol levels, …
  - Gender, ethnicity, socioeconomic category, zipcode, level of alcohol consumption, physical activity, …
  - Treated previously with Drug *X*
  - Patient has had this cancer
- Note the features can be
  - numeric values, or categorical values, or binary (yes/no) values

A Model to Answer Questions

We want to create a **model** from this data that we could use to answer questions like:

- Which features or combination of features correlate mostly highly with having this kind of cancer?
- Are there patterns in the set of patient data that help us understand the data better?
  - Are there "trends", outliers, etc.? Features that matter a lot, or very little?

- For new patients, can we predict the likelihood of them developing this cancer in the future?
  - Note the feature "had this cancer" divides the set of data into **<u>classes</u>** or **<u>categories</u>**
  - We call a feature used this way a **<u>label</u>**
  - This question is the <u>classification</u> problem

Multidimensional Data

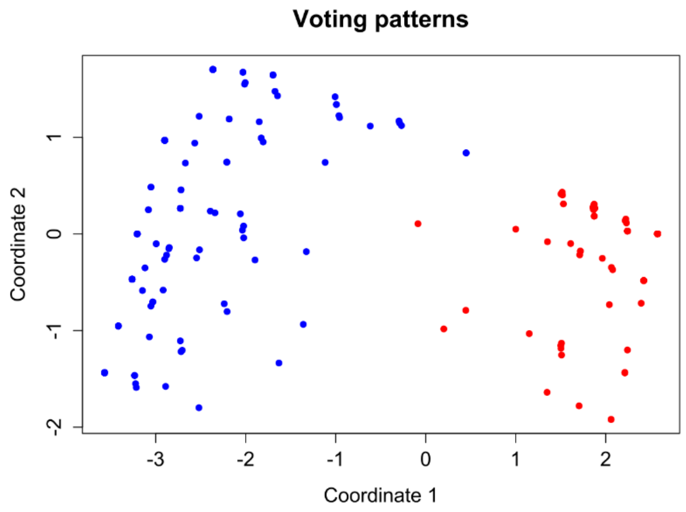
If the data is numeric, we can model the set as points in a multidimensional space

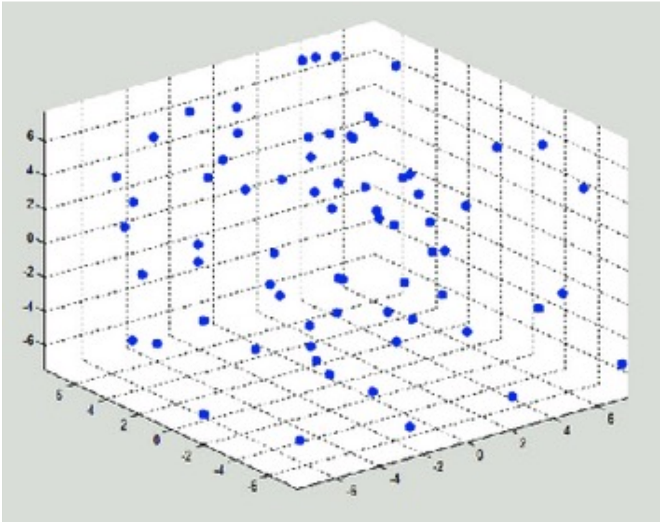
- So *n* features → *n*-dimensional space

Similarity of two samples is based on a distance metric

- Euclidean distance
  (or sometimes another measure)

In the field of Statistics, there are many multivariate statistical methods that model data this way





Final Comments on Features

Many issues to consider (that we won't say more about here in CS3100)

- In a multidimensional space, do we scale numerical values for features that have different means or ranges?
  (E.g. age vs. cholesterol level)
- Does our algorithm work well with a combination of numeric and non-numeric features? (ex: yes/no features)
- It can be harder to get useful information with a large number of features, so can we extract a smaller subset or combination of features that works better? (**Feature Extraction**)

Curse of dimensionality

Where We Are Now

1. This overview, and how many ML algorithms work with data
2. **Clustering as an example of unsupervised learning**
   – Two algorithms, including one you've seen
3. A simple classification algorithm to show supervised learning
   – This algorithm only uses concepts you've learned already
   – Then a brief overview of other techniques
4. A brief intro to reinforcement learning

Clustering

- An example of **unsupervised learning**
- **Goal:** divide samples in the data set into some number of *k* groups where items in a group are highly similar
  - *k* is sometimes specified, but some algorithms find a value of *k* that divides the clusters “best”
- **Usages**
  - Data exploration/understanding
  - Sometimes as a first step for supervised learning (coming soon)

Cluster Usage Examples

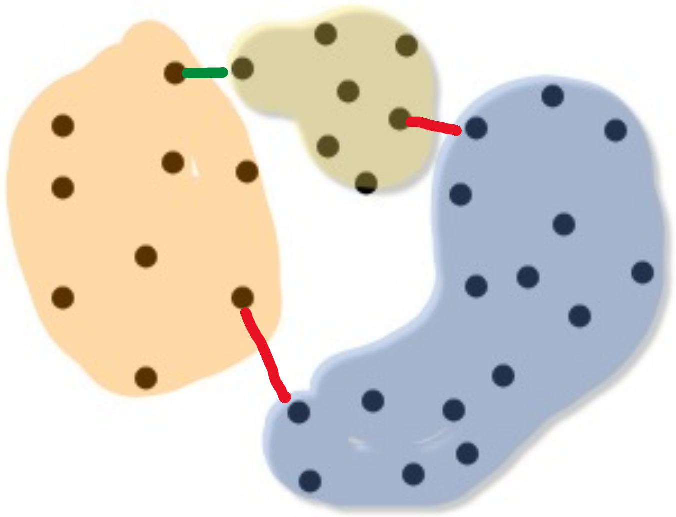
- Market Segmentation
  – Figure out different types of customers your company has.
- Organize Data Centers based on characteristics of the network traffic you are getting.
- Social Network Analysis
  – Automatically compute different types of users in a social network
- Recommender Tools
  – “People who viewed this also liked….”

Remember PA3?

- Wait, that was a problem about **graphs**, wasn't it?
  - And not the kind of multidimensional data we just talked about… 😕
- Remember for PA3:
  - We thought of our input as a complete, undirected, weighted graph
  - Our input was like a distance matrix (weights between all possible pairs)
- For multidimensional data and this algorithm, we note:
  - Each observation is like a graph node
  - The similarity between pairs of observations is like the edge weights in the complete graph

Clustering Strategy in PA3

- PA3's strategy is called **single-link clustering**
- Reminders:
  - – The distance $D(C_i, C_j)$ between the pair of clusters C_i and C_j is the smallest distance between any two observations in the pair of clusters
    - In the example, the red and green lines
  - – For <u>any given division</u> into clusters, the clustering-score is the smallest of the $D(C_i, C_j)$ values, i.e. $min_{i,j}\ D(C_i, C_j)$
    - In the example shown, it's the length of the green line
  - – Find <u>the best division</u> into clusters, i.e. the one that maximizes that value (i.e. the smallest of the between-pair distances)



Solution Using MST Algorithms

- To find the best clustering, remove the highest-weight k-1 edges from the MST
  - For Kruskal's, stop after adding n-k edges
  - For Prim's, build complete MST, sort MST edges, remove k-1 largest
- Time complexity for both of these solutions?
  - Depends on use of indirect heaps for Prim's or Union/Find improvements for Kruskal's
  - They're reasonable!

To Learn More….

https://en.wikipedia.org/wiki/Single-linkage_clustering

- This algorithm may produce long thin clusters in which nearby elements of the same cluster have small distances, but…
- Elements at opposite ends of a cluster may be much farther from each other than two elements in another cluster
  - Wikipedia notes that this characteristic makes sense for astronomers grouping clusters of galaxies
- Wikipedia describes other algorithms that don't use MSTs

What's Next?

1. This overview, and how many ML algorithms work with data
2. Clustering as an example of unsupervised learning
   – Single-link clustering (like PA3)
   – **Another algorithm: *k* means clustering**
3. A simple classification algorithm to show supervised learning
   – This algorithm only uses concepts you've learned already
   – Then a brief overview of other techniques
4. A brief intro to reinforcement learning