CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Models and Algorithms Part 2

Announcements

- Project Phase 1 released
  - If you did not register as a team yet, you must do so ASAP

Recap from Last Class

- Linear regression
  - Attempt to model the relationship between two (or more) variables
  - $y = \beta_0 + \beta_1 x + \epsilon$
  - β are the coefficients we are solving for, ϵ is our error term, or *noise*
  - The B terms determine the *trend* (or the direction of the line)
  - The noise term determines the amount of *variation* (or how close our observed data is to the line)
  - We can also add more predictor variables, or try non-linear relationships

Clustering with k-means

- Last time we learned Linear Regression, a model that can be useful for ***predicting*** outcomes based on a set of independent variables
- Today we are going to look at an algorithm for ***clustering*** points in a dataset: k-means

Machine Learning Algorithm Classification

- Machine Learning algorithms can be divided into 3 categories:
  - Supervised: We know the "right answer", fit a model to that knowledge
  - Unsupervised: We don't know the answer, want the algorithm to find it
  - Semi-Supervised: We know some, but there is more to learn
- Linear regression is an example of a supervised algorithm. We have a training set of predictors and the observed outcomes, and we fit a model based on that knowledge.

k-means Clustering

- Unsupervised algorithm to find "clusters" in data
  - We don't know or assume anything about the data
- Goal is to "segment" or "cluster" data
  - For example, your data is users, you want to divide them into groups of "similar" users. Why?
    - Serve different ads/provide different experiences
    - Further modeling may differ based on groups

The Algorithm

1. Choose the number of clusters
2. Initialize centroids to some value
    a. Could be via some special algorithm, or could be random
3. Then repeat the following steps…
    a. Reassign all points to the closest centroid
    b. Recalculate the centroids position based on this assignment
4. …until there is no change in centroid values or points stop switching

Interactive Example: Visualizing K-Means Clustering

The Theory Behind the Algorithm

- k-means searches for the minimum *sum of squares* assignment
- In order to converge consider the following two conditions:
  - Re-assigning points reduces the sum of squares
  - Re-computing centroids reduces the sum of squares
- Since both steps reduce the sum of squares, does it converge?
- There are only a finite number of ways to assign the finite number of points to each centroid, so the algorithm must converge
  - It will find a *local* minimum…

Some Issues with k-means

- How do we choose the number of clusters?
  - For 2D data, we may be able to intuit via inspection…but higher dimensional data gets tricky fast
- It will not necessarily find the global minimum
  - Doing so is an NP-Hard problem
- How do you interpret results? What do the clusters represent?
  - Sometimes it makes sense, sometimes it does not

A Small Example

- Consider just the ages of a group of users:

{ 23, 25, 24, 23, 21, 31, 32, 30, 31, 30, 37, 35, 38, 37, 39, 42, 43, 45, 43, 45 }

A Small Example

- Consider just the ages of a group of users:

{ 23, 25, 24, 23, 21, 31, 32, 30, 31, 30, 37, 35, 38, 37, 39, 42, 43, 45, 43, 45 }

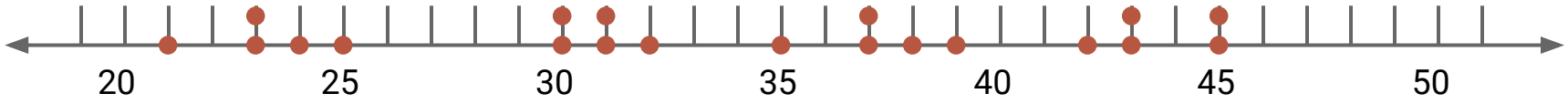
- Let's sort it quick (just for our benefit)

{ 21, 23, 23, 24, 25, 30, 30, 31, 31, 32, 35, 37, 37, 38, 39, 42, 43, 43, 45, 45 }

- Assuming 3 groupings, how might we as humans segment the data?

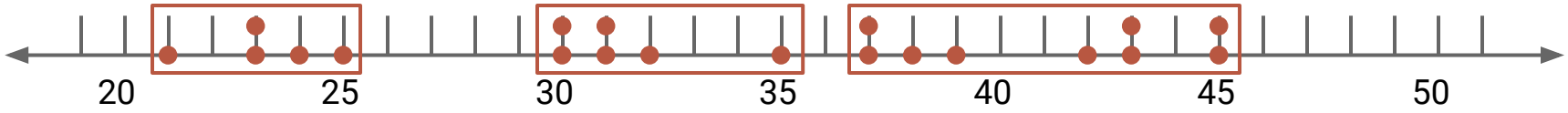
A Small Example

- As humans, it might seem natural to segment by 10s: 20 year olds, 30 year olds, and 40 year olds
- Clustering with k-means gives slightly different results



A Small Example

- As humans, it might seem natural to segment by 10s: 20 year olds, 30 year olds, and 40 year olds
- Clustering with k-means gives slightly different results
  - Is this useful?
  - What can we do with this information?
  - What if we use a different number for k?



Factors to Consider

- What are the basic variables in your problem?
  - What do they represent?
  - What is their scale?
  - How are they related? (based on intuition and observation)
  - What are your predictors?
- What are the underlying processes?
  - Are you attempting to capture them in a model?
- **What do you want to know? (may require a domain expert)**

Demo Examples

- The following examples use the real estate data from DDS Ch 3
- The implementations for linear regression and k-means are from scikit-learn: https://scikit-learn.org/stable/
  - scikit-learn is a Python library machine learning and data analytics

Linear Regression - Example

- In Real Estate, what determines the sale price of a property?

Linear Regression - Example

- In Real Estate, what determines the sale price of a property?
  - Size (of the building and of the land?)
  - Location
  - Date of Sale
  - Age of the property
  - Type of property (commercial, residential, rental, etc)
  - Quality of the build
  - Amenities
  - etc...

Linear Regression - Developing a Model

- What data do we have? (and do we need to get more?)

Linear Regression - Developing a Model

- What data do we have? (and do we need to get more?)
  - Square footage: numeric, presumably bigger properties cost more

Linear Regression - Developing a Model

- What data do we have? (and do we need to get more?)
  - Square footage: numeric, presumably bigger properties cost more
  - Location: neighborhood name…non-numeric
    - Are there underlying factors at the core of this?
      - Crime Rate? Schools? Number of parks? etc…

Linear Regression - Developing a Model

- What data do we have? (and do we need to get more?)
  - Square footage: numeric, presumably bigger properties cost more
  - Location: neighborhood name…non-numeric
    - Are there underlying factors at the core of this?
      - Crime Rate? Schools? Number of parks? etc…
  - Property Type: class of building…non-numeric
    - Does it make sense to include all classes in our model? Maybe we want to model each class separately? Depends on our problem.

Linear Regression - Developing a Model

- What data do we have? (and do we need to get more?)
  - **Square footage: numeric, presumably bigger properties cost more**

To start, let's focus on square footage.

Thought experiment: What if square footage was the only factor in the sale price of a property? What might the model look like?

Demo in JupyterLab

Takeaways?

- The more thorough your cleaning and EDA, the easier the modeling process becomes
- Understand what you are modeling
  - If neighborhood has a large impact on sale price, we need to capture that in our model or use different models per neighborhood
- Sometimes the best answer is more data

k-Means - Example

- How might we cluster properties?

k-Means - Example

- How might we cluster properties?
  - Perhaps we are trying to make policies/decisions based on size and age of a property
    - In a real scenario, may want to include more attributes, but for an example, 2D is easy to visualize
    - Motivation is going to based on problem statement and domain expertise

Demo in JupyterLab

Takeaways?

- Scale is an important part of a k-means model
- Results can be tricky to interpret
- Choice of k is also based on some sense of intuition/domain knowledge

References

- Dataset from Doing Data Science Ch. 3
- pandas tutorial: https://pandas.pydata.org/
- scikit-learn: https://scikit-learn.org/stable/