Introduction and Overview of k-means clustering

Ryan Miller

Grinnell College
Statistics

Defining “Machine Learning”

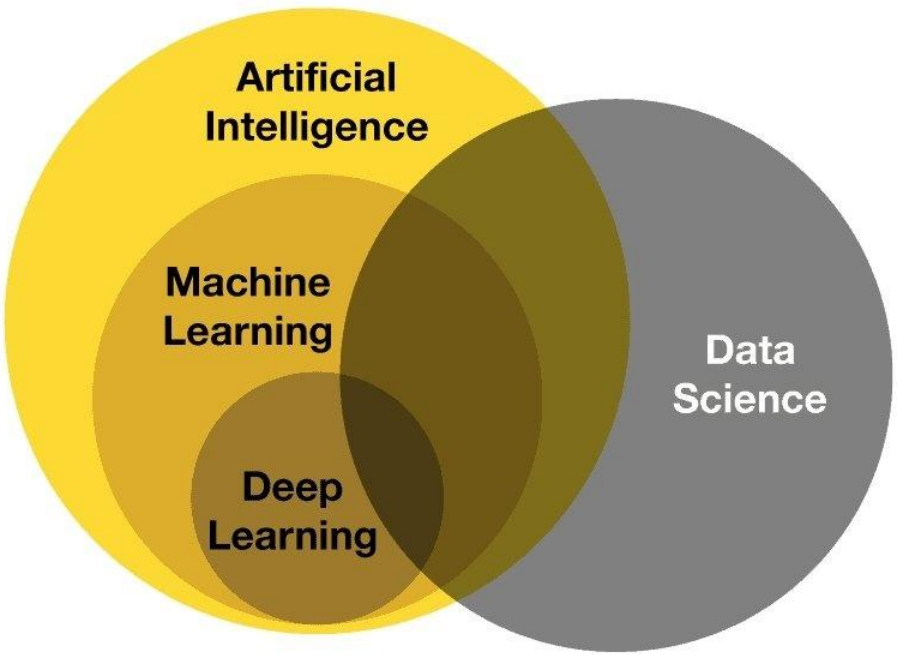
- How would you define “machine learning”?

Defining "Machine Learning"

- How would you define "machine learning"?
  - IBM: a branch of artificial intelligence (AI) focusing on the use of data and algorithms to imitate the way humans learn (gradually improving with experience)

Defining "Machine Learning"

- How would you define "machine learning"?
  - IBM: a branch of artificial intelligence (AI) focusing on the use of data and algorithms to imitate the way humans learn (gradually improving with experience)



What does it mean to learn?

Consider an arbitrary data set containing n samples of p features, what are some things we could "learn"?

What does it mean to learn?

Consider an arbitrary data set containing n samples of p features, what are some things we could "learn"?

- **Supervised learning**:
  - **Classification** - how to classify each sample into a discrete category
  - **Regression** - how to predict a numeric quantity corresponding to each sample
- **Unsupervised learning**:
  - **Clustering** - how to categorize the *samples* into meaningful groups (that aren't predefined)
  - **Dimension reduction** - how to meaningfully represent each sample using $p^* < p$ features by finding patterns among the original features
  - **Anomaly detection** - how to identify samples that deviate significantly from "normal data"

First steps

- We'll devote most of our attention this semester to supervised learning, but we'll start with a short unit on unsupervised learning for a few reasons:
  - It's manageable to learn Python, unsupervised learning algorithms, and notation simultaneously
  - Unsupervised methods can be used as *pre-processing steps* to prepare data for supervised learning

Notation

To begin, let's consider **tabular data** that is organized in a matrix, $\mathbf{X}$, consisting of n samples (rows) with p features (columns)

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

In this class I'll try to stick to the following conventions:

1. Bold capital letters to denote a matrix
2. Bold lower-case letters to denote a vector
3. non-bold capital letters to denote a random variable
4. non-bold lower-case letters to denote a scalar value

Clustering

- The goal of clustering is categorize our n samples into groups
  - A simple approach is to find k non-overlapping groups, where k is pre-specified
  - As such, the first approach we'll discuss is **k-means clustering**

k-means clustering (set up)

- Consider a set of **prototypes** defined by the following coordinates in the p-dimensional space of our data:

$$\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$$

- Let $A()$ denote an *assignment function* that assigns a sample to the group associated with a prototype
  - For example, $A(\mathbf{x}_i) = j$ indicates the i^{th} sample is assigned to the j^{th} prototype

k-means clustering (objective function)

- Generally speaking, machine learning methods occurs by optimizing an **objective function** (in supervise learning we'll call this a *cost function*)
  - In k-means clustering, we seek to minimize the following:

$$\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{z}_{A(\mathbf{x}_i)}||^2 = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{z}_{A(\mathbf{x}_i)})^T (\mathbf{x}_i - \mathbf{z}_{A(\mathbf{x}_i)})$$

- The data (the $\mathbf{x}_i$'s) are fixed, and we can simplify things by assuming an assignment function, such as the nearest prototype by euclidean distance: $d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{k} (x_{1,j} - x_{2,j})^2}$
  - k-means clustering amounts to solving for the prototype coordinates that minimize this objective function

k-means (algorithm)

The k-means algorithm alternates between two steps:

1. Given the current prototype coordinates, assign each sample to the nearest prototype using euclidean distance (ie: apply $A()$ to each sample).
2. Update the prototype coordinates to now be the mean coordinates of their respective assigned samples.

These steps repeat until additional iterations no longer improve the objective function.

k-means (algorithm demo)

Here is animation similar to what we walked through on the board:

- Link to k-means GIF animation

k-means (initialization)

Unfortunately, k-means is sensitive to the initial set of prototype coordinates. The `sklearn` implementation allows us two options:

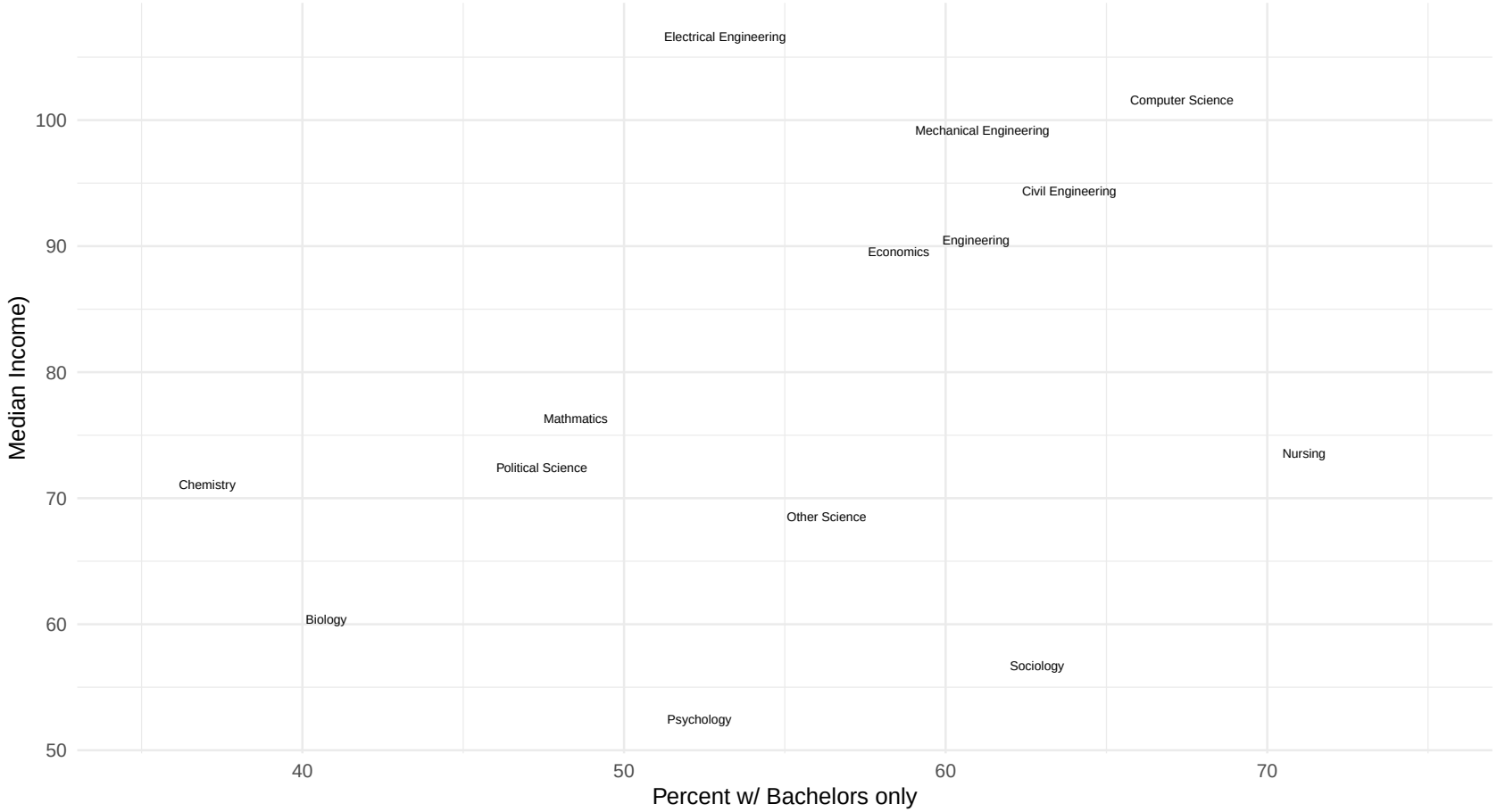
1. `random` (also known as Forgy) - randomly select k of the observed samples and use their coordinates as the initial prototypes
2. `k-means++` - choose the first prototype uniformly at random from the observed samples, then use probability distributions find additional prototypes that are far from existing prototypes (click here for details)

k-means (the role of k)

- The k-means algorithm requires a pre-specified value of a k, making it an example of what's known as a **hyperparameter** or **tuning parameter**
- A simple way to intelligently decide upon the value of hyperparameter is a method known as **grid search**, which is an exhaustive search over an enumerated set of candidate values
  - In the context of k-means, we might look at the value of the objective function over a sequence of k values
    - We can then visually display this relationship and select k at an inflection point (where additional prototypes no longer produce the same improvement in the object function)

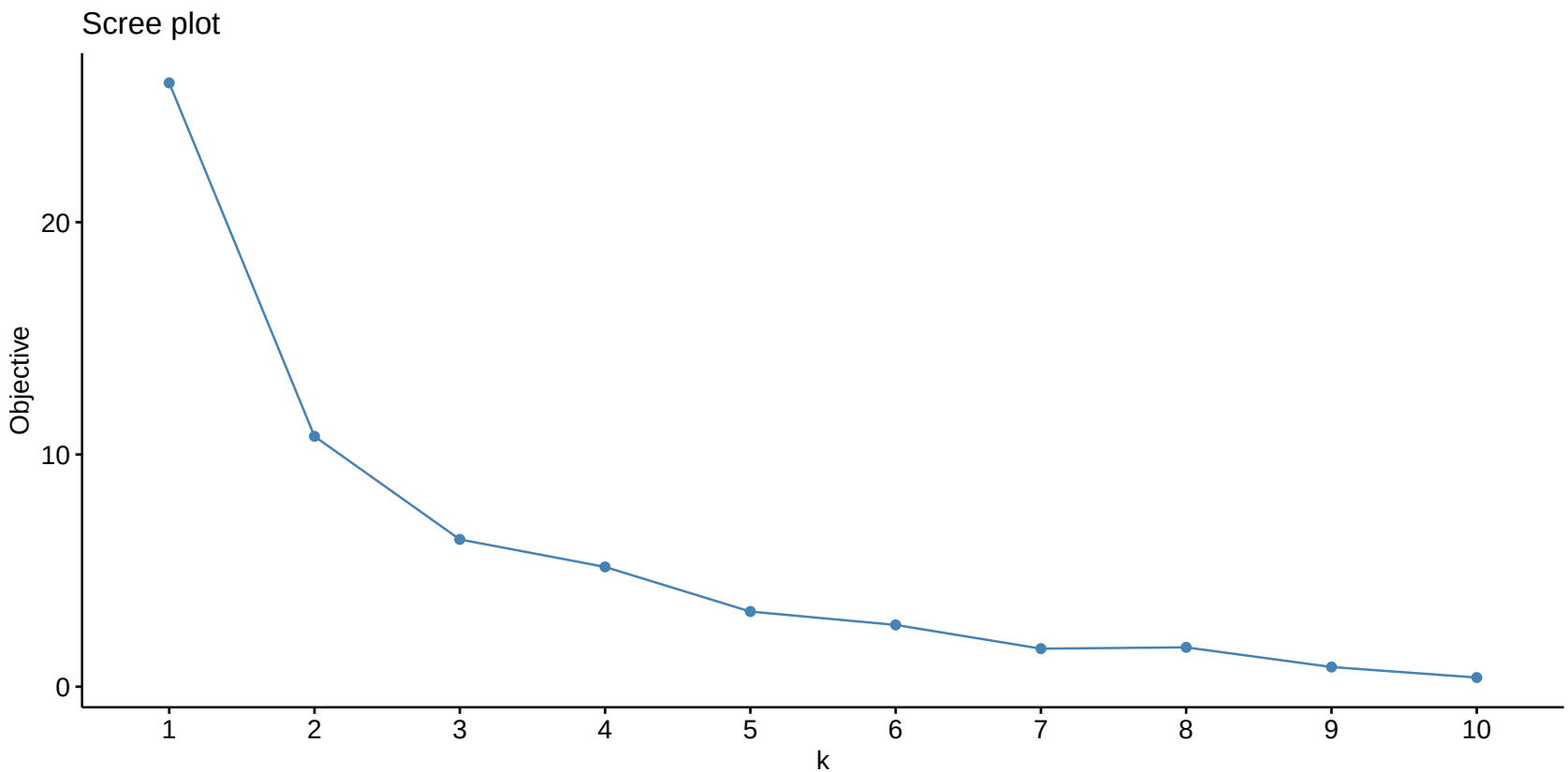
k-means (example - starting data)

These are data from Census's ACS income by degree field survey:

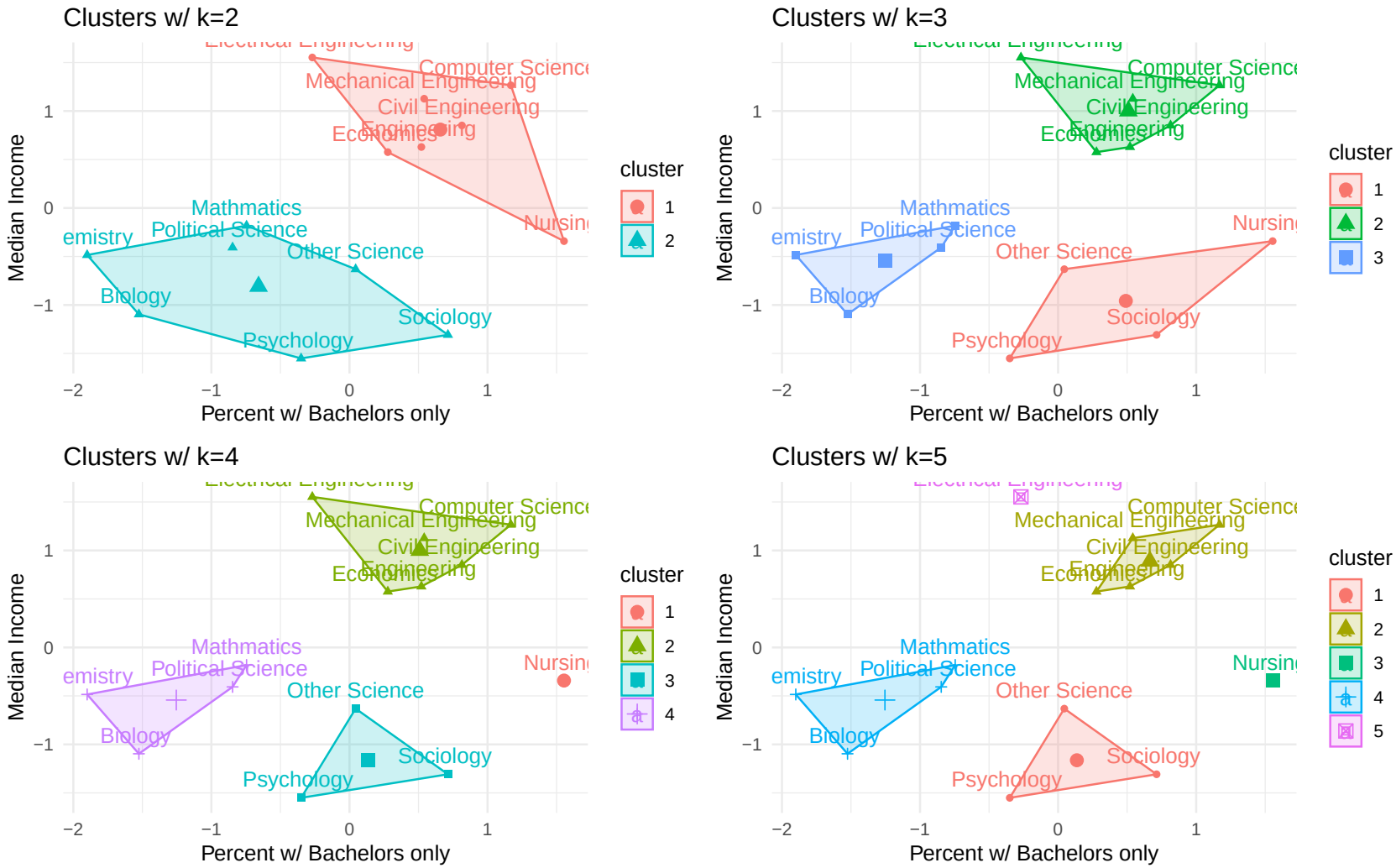


k-means (example - objective vs. k)

Below is the value of the objective function for several choices of k, notice how it levels off:



k-means (example - clustering results)



Conclusion

- *Supervised learning* aims to predict a numeric outcome or classify data-points after learning the patterns present in the available data
- *Unsupervised learning* finds patterns in the available data without specifying a target outcome
  - k-means clustering is an example of an unsupervised learning method
- Today's lab will be longer than normal and will cover both the basics of Python as well as how to utilize k-means clustering in Python with an application to image segmentation