Introduction to Clustering

Philip D. Waggoner

MACS 40800: Unsupervised Machine Learning

October 15, 2019

Lecture Outline

Lecture Outline

Clustering Basics

- Clustering methods attempt to group (or cluster) objects (i and i') based on some rule defining the similarity (or dissimilarity) between the objects

Clustering Basics

- Clustering methods attempt to group (or cluster) objects (i and i') based on some rule defining the similarity (or dissimilarity) between the objects

- Note that there is a distinction between clustering and classification/discrimination:

Clustering Basics

- Clustering methods attempt to group (or cluster) objects (i and i') based on some rule defining the similarity (or dissimilarity) between the objects

- Note that there is a distinction between clustering and classification/discrimination:
  - **Clustering**: the group labels are not known a priori

Clustering Basics

- Clustering methods attempt to group (or cluster) objects (i and i') based on some rule defining the similarity (or dissimilarity) between the objects

- Note that there is a distinction between clustering and classification/discrimination:
  - **Clustering**: the group labels are not known a priori
  - **Classification**: the group labels are known for a trained sample (we won't get into this iteration given the scope of the class)

Clustering Basics

- Clustering methods attempt to group (or cluster) objects (i and i') based on some rule defining the similarity (or dissimilarity) between the objects

- Note that there is a distinction between clustering and classification/discrimination:
  - **Clustering**: the group labels are not known a priori
  - **Classification**: the group labels are known for a trained sample (we won't get into this iteration given the scope of the class)

- Thus, the typical goal in clustering is to discover the "natural groupings" present in the data

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise)

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both “soft” and “hard”)

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both "soft" and "hard") $\rightsquigarrow$ maintain a set of clusters and assign points nearest to the cluster centroid (defined many ways, discussed more soon)

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both "soft" and "hard") $\rightsquigarrow$ maintain a set of clusters and assign points nearest to the cluster centroid (defined many ways, discussed more soon)
- Key difference between these techniques:

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both "soft" and "hard") $\rightsquigarrow$ maintain a set of clusters and assign points nearest to the cluster centroid (defined many ways, discussed more soon)
- Key difference between these techniques: **subdividing the data**

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both "soft" and "hard") $\rightsquigarrow$ maintain a set of clusters and assign points nearest to the cluster centroid (defined many ways, discussed more soon)
- Key difference between these techniques: **subdividing the data**
  - Hierarchical $\rightsquigarrow$ No

Clustering Basics

- Broadly, then, we can categorize clustering techniques into two camps:
  - **Hierarchical** (pairwise) $\rightsquigarrow$ move from signletons to progressively larger clusters based on similarity (we will get into similarity much more in a moment)
  - **Partitioning** (assignment, both "soft" and "hard") $\rightsquigarrow$ maintain a set of clusters and assign points nearest to the cluster centroid (defined many ways, discussed more soon)
- Key difference between these techniques: **subdividing the data**
  - Hierarchical $\rightsquigarrow$ No
  - Partitioning $\rightsquigarrow$ Yes

Clustering Basics

- Regardless of the technique, the **grouping** of data is central to reducing the dimensionality of the space, and so making the data more accessible and understandable

Clustering Basics

- Regardless of the technique, the **grouping** of data is central to reducing the dimensionality of the space, and so making the data more accessible and understandable
- But note that precisely how we think about grouping is dependent on the distribution of the data, which has implications for algorithm selection and validation

Clustering Basics

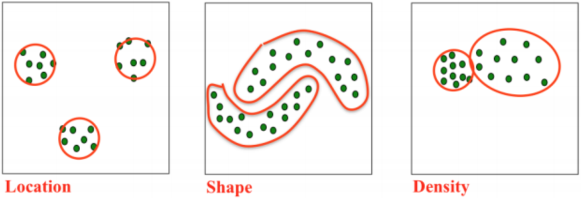
- Regardless of the technique, the **grouping** of data is central to reducing the dimensionality of the space, and so making the data more accessible and understandable
- But note that precisely how we think about grouping is dependent on the distribution of the data, which has implications for algorithm selection and validation
- In general, we can think of three main types of grouping:

Clustering Basics

- Regardless of the technique, the **grouping** of data is central to reducing the dimensionality of the space, and so making the data more accessible and understandable
- But note that precisely how we think about grouping is dependent on the distribution of the data, which has implications for algorithm selection and validation
- In general, we can think of three main types of grouping:
  - Location
  - Shape
  - Density

Clustering Basics

- Regardless of the technique, the **grouping** of data is central to reducing the dimensionality of the space, and so making the data more accessible and understandable
- But note that precisely how we think about grouping is dependent on the distribution of the data, which has implications for algorithm selection and validation
- In general, we can think of three main types of grouping:
  - Location
  - Shape
  - Density



Clustering Basics

- We will focus on three major types of clustering this week and next:

Clustering Basics

- We will focus on three major types of clustering this week and next:
  - Hierarchical agglomerative clustering

Clustering Basics

- We will focus on three major types of clustering this week and next:
  - Hierarchical agglomerative clustering
  - Hard partitioning (focus on k-means clustering, with a few cousins)

Clustering Basics

- We will focus on three major types of clustering this week and next:
    - Hierarchical agglomerative clustering
    - Hard partitioning (focus on k-means clustering, with a few cousins)
    - Soft (model-based) partitioning, can be probabilistic or fractional (focus on the EM algorithm and Gaussian mixture models, with a few cousins)

Lecture Outline

1. Clustering Basics
2. Diagnosing Clusterability
3. Conceptualizing and Calculating Distance
4. Clustering Applications
5. Problem Set 2

Diagnosing Clusterability

- Before we can cluster (or even decide whether clustering makes sense as a reasonable path forward), we should explore the feature space a bit

Diagnosing Clusterability

- Before we can cluster (or even decide whether clustering makes sense as a reasonable path forward), we should explore the feature space a bit
- There are several ways to do this:

Diagnosing Clusterability

- Before we can cluster (or even decide whether clustering makes sense as a reasonable path forward), we should explore the feature space a bit
- There are several ways to do this:
  - Informally (simple distribution plots)

Diagnosing Clusterability

- Before we can cluster (or even decide whether clustering makes sense as a reasonable path forward), we should explore the feature space a bit
- There are several ways to do this:
  - Informally (simple distribution plots)
  - Visually (VAT/ODI plots)

Diagnosing Clusterability

- Before we can cluster (or even decide whether clustering makes sense as a reasonable path forward), we should explore the feature space a bit
- There are several ways to do this:
  - Informally (simple distribution plots)
  - Visually (VAT/ODI plots)
  - Mathematically (sparse sampling)

A Brief Caveat: Informed Guessing

- Ambiguity underlies this process and all of UML

A Brief Caveat: Informed Guessing

- Ambiguity underlies this process and all of UML
- We can do our best to make sense of a complex feature space, but our data are unlabeled, and we have relatively weak (if any) priors on distributions, and so on

A Brief Caveat: Informed Guessing

- Ambiguity underlies this process and all of UML
- We can do our best to make sense of a complex feature space, but our data are unlabeled, and we have relatively weak (if any) priors on distributions, and so on
- This is a limitation of UML acorss the board; yet simultaneously the reason it is so important to combine UML with other data reduction and modeling processes

Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:

Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length

Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length
  - Petal Width

Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length
  - Petal Width
  - Sepal Length

Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length
  - Petal Width
  - Sepal Length
  - Sepal Width

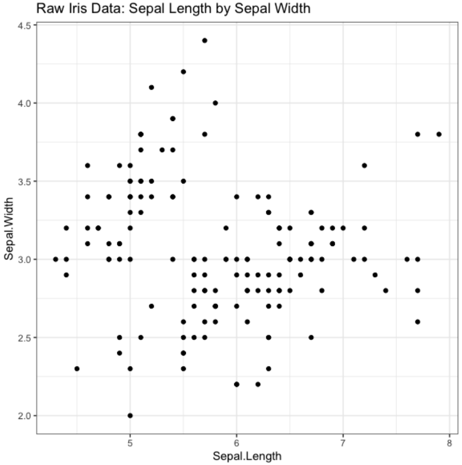
Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length
  - Petal Width
  - Sepal Length
  - Sepal Width
  - 3 Species: setosa, versicolor, and virginica

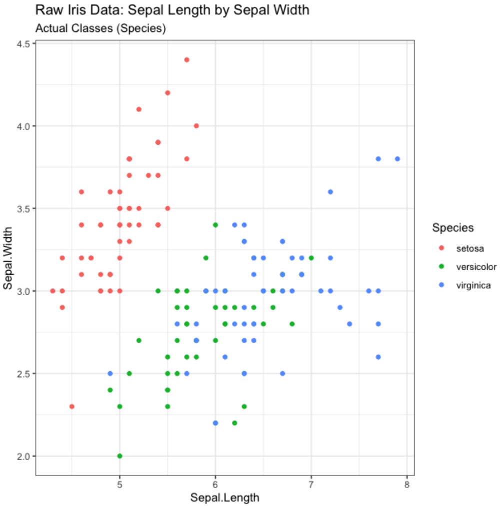
Diagnosing Clusterability: Informally

- Let's start with informal diagnosis using the famous Fisher Iris data set:
  - Petal Length
  - Petal Width
  - Sepal Length
  - Sepal Width
  - 3 Species: setosa, versicolor, and virginica
  - 150 observations (50 of each)

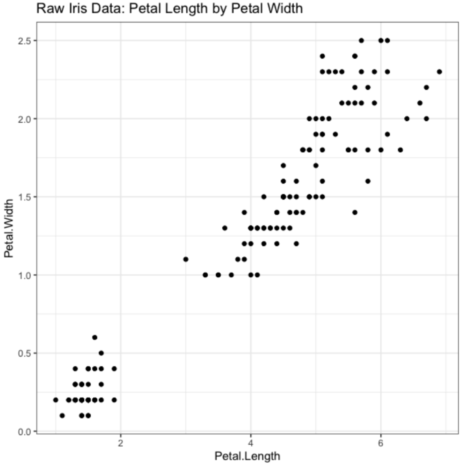
Diagnosing Clusterability: Informally (Sepal Length by Sepal Width)



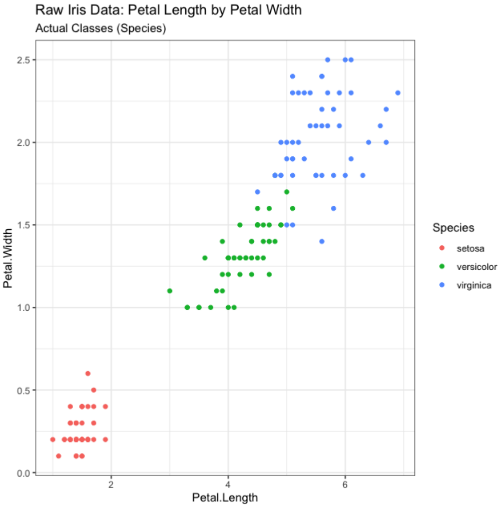
Diagnosing Clusterability: Informally (Sepal Length by Sepal Width)



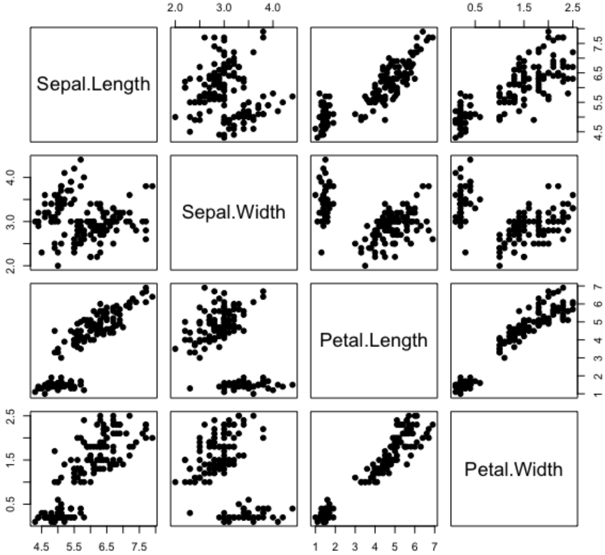
Diagnosing Clusterability: Informally (Petal Length by Petal Width)



Diagnosing Clusterability: Informally (Petal Length by Petal Width)



Diagnosing Clusterability: Informally (All Features)



Diagnosing Clusterability: VAT (ODI) Plots

- Next, let's get a little more precise on diagnosing clusterability using Visual Assessment of Tendency (VAT) plots, and still with the Fisher Iris data set

Diagnosing Clusterability: VAT (ODI) Plots

- Next, let's get a little more precise on diagnosing clusterability using Visual Assessment of Tendency (VAT) plots, and still with the Fisher Iris data set
- Originally dervied by Bezdek and Hathaway (2002), these VAT plots are also often called Ordered Dissimilarity Images (ODI)

Diagnosing Clusterability: VAT (ODI) Plots

- Next, let's get a little more precise on diagnosing clusterability using Visual Assessment of Tendency (VAT) plots, and still with the Fisher Iris data set

- Originally dervied by Bezdek and Hathaway (2002), these VAT plots are also often called Ordered Dissimilarity Images (ODI)

- **Dissimilarity**: first, visualize the dissimilarity matrix

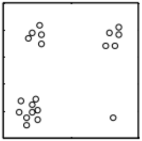
Diagnosing Clusterability: VAT (ODI) Plots

- Next, let's get a little more precise on diagnosing clusterability using Visual Assessment of Tendency (VAT) plots, and still with the Fisher Iris data set
- Originally dervied by Bezdek and Hathaway (2002), these VAT plots are also often called Ordered Dissimilarity Images (ODI)
- **Dissimilarity**: first, visualize the dissimilarity matrix
- **Ordered**: objects, o, that are spatially proximate (measured as k) are displayed in consecutive order (if k_i is near $k_{i'}$, then $o_1, o_2 \forall o \equiv o_{ki}, o_{ki'}$)

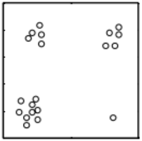
Diagnosing Clusterability: VAT (ODI) Plots

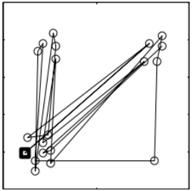
- Next, let's get a little more precise on diagnosing clusterability using Visual Assessment of Tendency (VAT) plots, and still with the Fisher Iris data set
- Originally dervied by Bezdek and Hathaway (2002), these VAT plots are also often called Ordered Dissimilarity Images (ODI)
- **Dissimilarity**: first, visualize the dissimilarity matrix
- **Ordered**: objects, o, that are spatially proximate (measured as k) are displayed in consecutive order (if k_i is near $k_{i'}$, then $o_1, o_2 \forall o \equiv o_{ki}, o_{ki'}$)
- The visual result becomes darker blocks along the diagonal reflect greater spatial similarity, compared to lighter shaded blocks, which inversely suggest greater dissimilarity

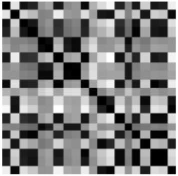
Order is Important



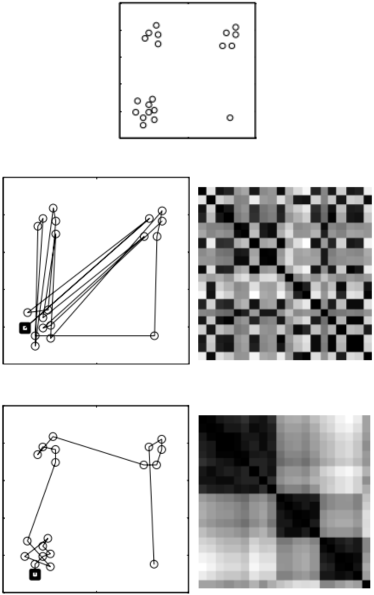
Order is Important







Order is Important



VAT (ODI): Iris Data

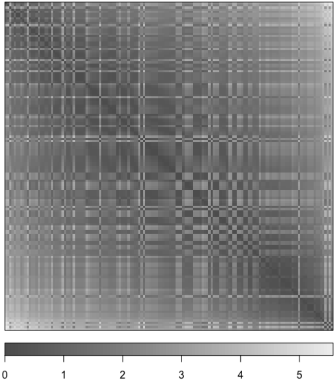


Figure: ODI: Sepal

VAT (ODI): Iris Data

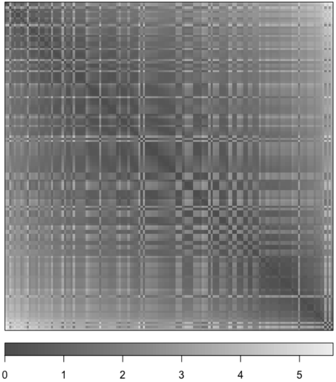


Figure: ODI: Sepal

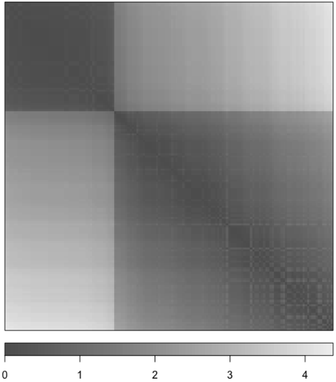


Figure: ODI: Petal

A Quick Comparison

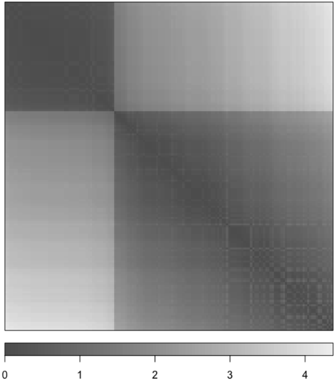


Figure: ODI: Petal

A Quick Comparison

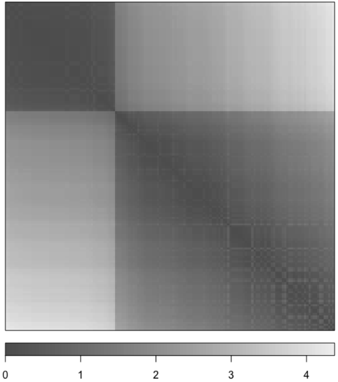


Figure: ODI: Petal

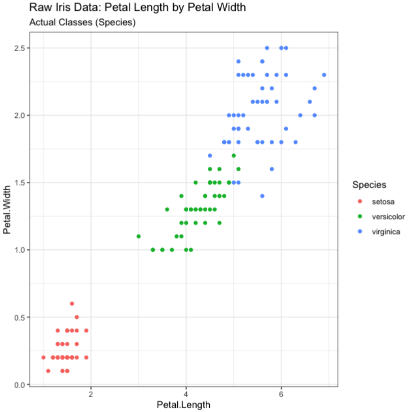


Figure: Raw Data: Petal

In R

Jump to R for a quick demo

Diagnosing Clusterability: Hopkins Statistic

- We can also mathematically explore what the ODI plots are showing using a simple, but powerful test statistic called the Hopkins statistic (we will stick with the Iris data here for consistency)

Diagnosing Clusterability: Hopkins Statistic

- We can also mathematically explore what the ODI plots are showing using a simple, but powerful test statistic called the Hopkins statistic (we will stick with the Iris data here for consistency)

- The Hopkins (or "H") statistic tests the the null hypothesis of spatial randomness in the data using a sparse sampling test

Diagnosing Clusterability: Hopkins Statistic

- We can also mathematically explore what the ODI plots are showing using a simple, but powerful test statistic called the Hopkins statistic (we will stick with the Iris data here for consistency)

- The Hopkins (or "H") statistic tests the the null hypothesis of spatial randomness in the data using a sparse sampling test

- It calculates the probability that a given dataset is generated by a uniform (random noise, with no clusters) distribution or not (non-random, with clustering likely)

Diagnosing Clusterability: Hopkins Statistic

- We can also mathematically explore what the ODI plots are showing using a simple, but powerful test statistic called the Hopkins statistic (we will stick with the Iris data here for consistency)

- The Hopkins (or "H") statistic tests the the null hypothesis of spatial randomness in the data using a sparse sampling test

- It calculates the probability that a given dataset is generated by a uniform (random noise, with no clusters) distribution or not (non-random, with clustering likely)

- This general procedure is called **sparse sampling**, of which H is one iteration

A Null Hypothesis Framework

- We specify a null hypothesis test,

A Null Hypothesis Framework

- We specify a null hypothesis test,

  H_0: the data is uniformly ("equally") distributed

A Null Hypothesis Framework

- We specify a null hypothesis test,

  H_0: the data is uniformly ("equally") distributed

  H_A: the data is not uniformly distributed

A Null Hypothesis Framework

- We specify a null hypothesis test,

  H_0: the data is uniformly ("equally") distributed

  H_A: the data is not uniformly distributed

- **Goal**: calculate the pairwise dissimilarity across all observations in the **actual** data and compare to a set of **simulated** data drawn from some random distribution (usually uniform) with the *same* standard deviation as the original data

A Null Hypothesis Framework

- We specify a null hypothesis test,

  H_0: the data is uniformly ("equally") distributed

  H_A: the data is not uniformly distributed

- **Goal**: calculate the pairwise dissimilarity across all observations in the **actual** data and compare to a set of **simulated** data drawn from some random distribution (usually uniform) with the *same* standard deviation as the original data

- In other words, we are creating a random, synthetic version of the original data set (the sampling "window"), and we are comparing to see whether these produce similar distributions (i.e., is the actual data random, compared to the synthetic data set, which we *know* is random?)

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D
- For each observation $q_i \in D'$, find it's nearest neighbor q_i'

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D
- For each observation $q_i \in D'$, find it's nearest neighbor q_i'
- Calculate the distance between q_i and q_i' and denote it $w_j = dist(q_i, q_i')$

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D
- For each observation $q_i \in D'$, find it's nearest neighbor q_i'
- Calculate the distance between q_i and q_i' and denote it $w_j = dist(q_i, q_i')$
- H is calculated as the mean distance in the actual data divided by the sum of the mean distances in the actual and simulated data

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D
- For each observation $q_i \in D'$, find it's nearest neighbor q_i'
- Calculate the distance between q_i and q_i' and denote it $w_j = dist(q_i, q_i')$
- H is calculated as the mean distance in the actual data divided by the sum of the mean distances in the actual and simulated data

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} u_j^d + \sum_{j=1}^{m} w_j^d} \tag{1}$$

Calculating the Hopkins Statistic

- Sample uniformly n observations, p_i, from our **actual** data, D
- For each observation $p_i \in D$, find the nearest neighbor p_i'
- Calculate the distance between p_i and p_i' and denote it as $u_j = dist(p_i, p_i')$
- Create a **simulated** dataset, D', drawn from a random, uniform distribution with n observations (q_i), with the same standard deviation as D
- For each observation $q_i \in D'$, find it's nearest neighbor q_i'
- Calculate the distance between q_i and q_i' and denote it $w_j = dist(q_i, q_i')$
- H is calculated as the mean distance in the actual data divided by the sum of the mean distances in the actual and simulated data

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} u_j^d + \sum_{j=1}^{m} w_j^d} \tag{1}$$

- In general, $H > 0.5$ leads to rejection of H_0, suggesting the data are non-random, and are "clusterable"

Lecture Outline

Similarity and Dissimilarity

- Distance data are relational, where greater values reflect can either reflect greater similarity or dissimilarity across observations or features

Similarity and Dissimilarity

- Distance data are relational, where greater values reflect can either reflect greater similarity or dissimilarity across observations or features
- Therefore, most efforts to produce a simple structure from a complex data set require a **measure** of distance or "similarity"

Similarity and Dissimilarity

- Distance data are relational, where greater values reflect can either reflect greater similarity or dissimilarity across observations or features
- Therefore, most efforts to produce a simple structure from a complex data set require a **measure** of distance or “similarity”
- Important considerations include the nature of the variables, scales of measurement, and domain expertise

Similarity and Dissimilarity

- Distance data are relational, where greater values reflect can either reflect greater similarity or dissimilarity across observations or features
- Therefore, most efforts to produce a simple structure from a complex data set require a **measure** of distance or "similarity"
- Important considerations include the nature of the variables, scales of measurement, and domain expertise
- When *items* are clustered, proximity is usually indicated by some sort of distance

Similarity and Dissimilarity

- Distance data are relational, where greater values reflect can either reflect greater similarity or dissimilarity across observations or features
- Therefore, most efforts to produce a simple structure from a complex data set require a **measure** of distance or "similarity"
- Important considerations include the nature of the variables, scales of measurement, and domain expertise
- When *items* are clustered, proximity is usually indicated by some sort of distance
- By contrast, *features* are usually grouped on the basis of correlations (but also for observations)

Standardization

- First, *always* standardize when clustering

Standardization

- First, *always* standardize when clustering
- Distance calculations are strongly influenced by unit measurement and magnitude

Standardization

- First, *always* standardize when clustering
- Distance calculations are strongly influenced by unit measurement and magnitude
- Suppose you had two features: weight (lbs.) and household income (dollars)

Standardization

- First, *always* standardize when clustering
- Distance calculations are strongly influenced by unit measurement and magnitude
- Suppose you had two features: weight (lbs.) and household income (dollars)
- In such a case, different units of measurement *and* distributions will always return biased results

Standardization

- First, *always* standardize when clustering
- Distance calculations are strongly influenced by unit measurement and magnitude
- Suppose you had two features: weight (lbs.) and household income (dollars)
- In such a case, different units of measurement *and* distributions will always return biased results
- Therefore, prior to any clustering (or distance caluclation), always standardize inputs to ensure they are on the same scale (most commonly setting $\mu = 0$ and $\sigma = 1$)

Standardization

- First, *always* standardize when clustering
- Distance calculations are strongly influenced by unit measurement and magnitude
- Suppose you had two features: weight (lbs.) and household income (dollars)
- In such a case, different units of measurement *and* distributions will always return biased results
- Therefore, prior to any clustering (or distance caluclation), always standardize inputs to ensure they are on the same scale (most commonly setting $\mu = 0$ and $\sigma = 1$)
- The result is effectly "unit-less" inputs

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p, q)$, and we want to calculate a measurement of distance, $d(p, q)$, that accurately reflects this feature space, such that $d(p, q) \rightarrow \delta(p, q)$

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p,q)$, and we want to calculate a measurement of distance, $d(p,q)$, that accurately reflects this feature space, such that $d(p,q) \rightarrow \delta(p,q)$

- Generally a distance measure $d(p,q)$ between two points p and q satisfies the following properties, where g is any other intermediate point:

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p, q)$, and we want to calculate a measurement of distance, $d(p, q)$, that accurately reflects this feature space, such that $d(p, q) \to \delta(p, q)$

- Generally a distance measure $d(p, q)$ between two points p and q satisfies the following properties, where g is any other intermediate point:

    - $d(p, q) = d(q, p)$

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p, q)$, and we want to calculate a measurement of distance, $d(p, q)$, that accurately reflects this feature space, such that $d(p, q) \rightarrow \delta(p, q)$

- Generally a distance measure $d(p, q)$ between two points p and q satisfies the following properties, where g is any other intermediate point:

  - $d(p, q) = d(q, p)$

  - $d(p, q) > 0$, if $p \neq q$

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p, q)$, and we want to calculate a measurement of distance, $d(p, q)$, that accurately reflects this feature space, such that $d(p, q) \rightarrow \delta(p, q)$

- Generally a distance measure $d(p, q)$ between two points p and q satisfies the following properties, where g is any other intermediate point:

    - $d(p, q) = d(q, p)$
    - $d(p, q) > 0$, if $p \neq q$
    - $d(p, q) = 0$, if $p = q$

Similarity and Dissimilarity

- We begin with some configuration of actual data, $\delta(p, q)$, and we want to calculate a measurement of distance, $d(p, q)$, that accurately reflects this feature space, such that $d(p, q) \to \delta(p, q)$
- Generally a distance measure $d(p, q)$ between two points p and q satisfies the following properties, where g is any other intermediate point:
  - $d(p, q) = d(q, p)$
  - $d(p, q) > 0$, if $p \neq q$
  - $d(p, q) = 0$, if $p = q$
  - $d(p, q) \leq d(p, g) + d(g, q)$

Spatial Measures Starting with Minkowski

Spatial Measures Starting with Minkowski

$$d_m(p, q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \qquad (2)$$

where, setting $m \geq 1$ defines some true distance

Spatial Measures Starting with Minkowski

$$d_m(p, q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \tag{2}$$

where, setting $m \geq 1$ defines some true distance

- $m = 1$: Manhattan ("city block") Distance:

Spatial Measures Starting with Minkowski

$$d_m(p, q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \quad (2)$$

where, setting $m \geq 1$ defines some true distance

- $m = 1$: Manhattan ("city block") Distance:

$$d_{manhattan}(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

Spatial Measures Starting with Minkowski

$$d_m(p, q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \tag{2}$$

where, setting $m \geq 1$ defines some true distance

- $m = 1$: Manhattan ("city block") Distance:

$$d_{manhattan}(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

- Canberra Distance (weighted version of Manhattan):

Spatial Measures Starting with Minkowski

$$d_m(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \tag{2}$$

where, setting $m \geq 1$ defines some true distance

- $m = 1$: Manhattan ("city block") Distance:

$$d_{manhattan}(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

- Canberra Distance (weighted version of Manhattan):

$$d_{canberra}(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Spatial Measures

$$d_m(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \tag{3}$$

where, setting $m \geq 1$ defines some true distance

- $m = 2$: Euclidean Distance:

Spatial Measures

$$d_m(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \quad (3)$$

where, setting $m \geq 1$ defines some true distance

- $m = 2$: Euclidean Distance:

$$d_{euclidean}(p,q) = (\sum_{i=1}^{n} (p_i - q_i)^2)^{\frac{1}{2}}$$

or, more compactly (and commonly),

Spatial Measures

$$d_m(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^m)^{\frac{1}{m}} \tag{3}$$

where, setting $m \geq 1$ defines some true distance

- $m = 2$: Euclidean Distance:

$$d_{euclidean}(p,q) = (\sum_{i=1}^{n} (p_i - q_i)^2)^{\frac{1}{2}}$$

or, more compactly (and commonly),

$$d_{euclidean}(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Correlation Measures

- Correlation measures calculate similarity between observations based on...

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!
- Meaning, two observations could be correlated across features, but far apart in space, suggesting that if we want to cluster based on *attribute* similarity, then correlation measures are ideal, over spatial distance

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!
- Meaning, two observations could be correlated across features, but far apart in space, suggesting that if we want to cluster based on *attribute* similarity, then correlation measures are ideal, over spatial distance
- Pearson Distance:

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!

- Meaning, two observations could be correlated across features, but far apart in space, suggesting that if we want to cluster based on *attribute* similarity, then correlation measures are ideal, over spatial distance

- Pearson Distance:

$$d_{pearson}(p, q) = 1 - \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \sum_{i=1}^{n}(q_i - \bar{q})^2}}$$

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!
- Meaning, two observations could be correlated across features, but far apart in space, suggesting that if we want to cluster based on *attribute* similarity, then correlation measures are ideal, over spatial distance
- Pearson Distance:

$$d_{pearson}(p, q) = 1 - \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \sum_{i=1}^{n}(q_i - \bar{q})^2}}$$

- Eisen Cosine Distance (generalized case of Pearson)):

Correlation Measures

- Correlation measures calculate similarity between observations based on...**correlation**!

- Meaning, two observations could be correlated across features, but far apart in space, suggesting that if we want to cluster based on *attribute* similarity, then correlation measures are ideal, over spatial distance

- Pearson Distance:

$$d_{pearson}(p, q) = 1 - \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \sum_{i=1}^{n}(q_i - \bar{q})^2}}$$

- Eisen Cosine Distance (generalized case of Pearson)):

$$d_{eisen}(p, q) = 1 - \frac{|\sum_{i=1}^{n} p_i q_i|}{\sqrt{\sum_{i=1}^{n} p_i^2 \sum_{i=1}^{n} q_i^2}}$$

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs.

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?
- Suppose we wanted to cluster based on whether someone is a female or not (categorical, $(0, 1)$) and their income (continuous)

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?
- Suppose we wanted to cluster based on whether someone is a female or not (categorical, $(0, 1)$) and their income (continuous)
- Why couldn't we used one of the previous distance metrics, like Euclidean distance in this case?

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?
- Suppose we wanted to cluster based on whether someone is a female or not (categorical, $(0, 1)$) and their income (continuous)
- Why couldn't we used one of the previous distance metrics, like Euclidean distance in this case?
- Why couldn't we use classes ($0, 1$, or $=, \neq$)?

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?
- Suppose we wanted to cluster based on whether someone is a female or not (categorical, $(0, 1)$) and their income (continuous)
- Why couldn't we used one of the previous distance metrics, like Euclidean distance in this case?
- Why couldn't we use classes $(0, 1$, or $=, \neq)$?
- The problem:

Mixed Data Distance Measure

- These measures are great for continuous, numeric inputs. But what about if we have mixed data on which we would like to cluster (which is quite common in the social sciences)?
- Suppose we wanted to cluster based on whether someone is a female or not (categorical, $(0, 1)$) and their income (continuous)
- Why couldn't we used one of the previous distance metrics, like Euclidean distance in this case?
- Why couldn't we use classes ($0, 1$, or $=, \neq$)?
- The problem: **need a distance measure to transform both types of data, without information loss or meaningless values**

Gower's Distance

Gower's Distance

$$d_{gower}(p, q) = \frac{\sum_{k=1}^{n} w_{pqk} s_{pqk}}{\sum_{k=1}^{n} w_{pqk}} \tag{4}$$

where

w_{pqk} is the weight for feature, k, between observations p and q

s_{pqk} is the distance between p and q on feature k

Gower's Distance

$$d_{gower}(p, q) = \frac{\sum_{k=1}^{n} w_{pqk} s_{pqk}}{\sum_{k=1}^{n} w_{pqk}} \tag{4}$$

where

w_{pqk} is the weight for feature, k, between observations p and q

s_{pqk} is the distance between p and q on feature k

- Gower's measure essentially captures the weighted average of the distances on the different features

Gower's Distance

- The key to this measure is, unlike numeric distance metrics, calculation of s_{pqk} does not apply the same formula to all features

Gower's Distance

- The key to this measure is, unlike numeric distance metrics, calculation of s_{pqk} does not apply the same formula to all features
- For categorical features, we use allow for an $[=, \neq]$ comparison, while for numeric features we caluclate the absolute difference

Gower's Distance

- The key to this measure is, unlike numeric distance metrics, calculation of s_{pqk} does not apply the same formula to all features
- For categorical features, we use allow for an $[=, \neq]$ comparison, while for numeric features we caluclate the absolute difference
- Importantly, all measures, s_{pqk}, are scaled to range between $[0, 1]$, where for categorical features, $s_{pqk} = 0$ when p and q are equal, and 1 otherwise,

$$s_{pqk} = \begin{cases} 0 & \text{if } p_k = q_k \\ 1 & \text{if } p_k \neq q_k \end{cases} \tag{5}$$

Gower's Distance

- The key to this measure is, unlike numeric distance metrics, calculation of s_{pqk} does not apply the same formula to all features
- For categorical features, we use allow for an $[=, \neq]$ comparison, while for numeric features we caluclate the absolute difference
- Importantly, all measures, s_{pqk}, are scaled to range between $[0, 1]$, where for categorical features, $s_{pqk} = 0$ when p and q are equal, and 1 otherwise,

$$s_{pqk} = \begin{cases} 0 & \text{if } p_k = q_k \\ 1 & \text{if } p_k \neq q_k \end{cases} \tag{5}$$

- And numeric features are scaled by dividing the absolute difference by the range of the feature,

$$s_{pqk} = \frac{|p_k - q_k|}{max(k) - min(k)} \tag{6}$$

Lecture Outline

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower?

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys
  - Geopolitical studies

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys
  - Geopolitical studies
  - Market preferences

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys
  - Geopolitical studies
  - Market preferences
  - Economies

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys
  - Geopolitical studies
  - Market preferences
  - Economies
  - Social media users

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
    - Respondents on large surveys
    - Geopolitical studies
    - Market preferences
    - Economies
    - Social media users
    - Perceptual studies

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
  - Respondents on large surveys
  - Geopolitical studies
  - Market preferences
  - Economies
  - Social media users
  - Perceptual studies
  - Geographic trends

What can I cluster?

- We have now introduced the idea of clustering, as well as key things to consider prior to clustering
- But we may be thinking, what can I cluster thats not a flower? $\rightsquigarrow$ a lot of things...
    - Respondents on large surveys
    - Geopolitical studies
    - Market preferences
    - Economies
    - Social media users
    - Perceptual studies
    - Geographic trends
    - And a lot more...

Lecture Outline

Problem Set #2: Clustering (part 1)

- **Due Saturday, 10/19, by 12 noon**