Introduction to Clustering and Classification

Psych 993

Methods for Clustering and Classification

Lecture 1

Today's Lecture

- Introduction to methods for clustering and classification

- Discussion of measures of distance:
  - Similarity
  - Dissimilarity
  - Statistical "distances"

General Overview

- Clustering methods are concerned with sorting objects into similar groups.
- The terms classification and clustering will be used synonymously throughout this course.
  - You will find that some authors provide a distinction between the two words.
  - Classification has been referred to the process of assigning objects to *known* groups.

Classification Examples

From Gordon (1999)

- Archaeologists have an interest in detecting similarities amongst artifacts, such as ornaments or stone implements found during excavations, as this would allow them to investigate the spatial distribution of artifact “types.”

Classification Examples

2. Plant ecologists collect information about the species of plant that are present in a set of quadrats, listing the species present in each quadrat and possibly also recording a measure of the abundance of each species of the quadrat. One of their aims is to arrange the quadrats in classes such that the members of each class possess some properties which distinguish them from members of other classes.

Classification Examples

3. Taxonomists are concerned with constructing classifications which summarize the relationships between taxonomic units of various kinds. The units are usually included in classes which are mutually exclusive and hierarchically nested, as illustrated by the Linnean system.

Classification Examples

4. Social network analysts investigate the interactions within a set of individuals, and are interested in identifying individuals who have similar aims or attributes.

Classification Examples

5. Those who enjoy sampling malt whiskies might be interested in seeing a classification of the distilleries producing these whiskies, as this would enable them to obtain an indication of the range of tastes available by selecting a small number of representatives from each class; using the classification, the whisky producers could identify their commercial competitors.

view: **All** | New Releases | Coming Soon

More results

1. 

HLM 6: Hierarchical Linear and Nonlinear Modeling
by Stephen W. Raudenbush, et al.
Publication Date: September 1, 2004

Our Price: $45.00

Add to cart
Add to Wish List

I Own It Not interested x|☆☆☆☆☆ Rate it
Recommended because you added Hierarchical Linear Models to your Shopping Cart (edit)

2. 

Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence
by Judith D. Singer, John B. Willett
Average Customer Review: ★★★★★
Publication Date: March 27, 2003

Our Price: $63.24
Used & new from $59.06

Add to cart
Add to Wish List

I Own It Not interested x|☆☆☆☆☆ Rate it
Recommended because you added Hierarchical Linear Models to your Shopping Cart (edit)

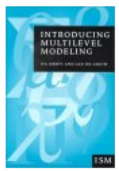
3. 

Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling
by Tom A B Snijders, Roel Bosker
Publication Date: December 7, 1999

Our Price: $47.95
Used & new from $40.01

Add to cart
Add to Wish List

I Own It Not interested x|☆☆☆☆☆ Rate it
Recommended because you added Hierarchical Linear Models to your Shopping Cart (edit)

4. 

Introducing Multilevel Modeling (Introducing Statistical Methods series)
by Ita G G Kreft, Jan de Leeuw
Average Customer Review: ★★★★☆
Publication Date: June 1, 1998

Our Price: $44.95
Used & new from $44.95

Add to cart
Add to Wish List

I Own It Not interested x|☆☆☆☆☆ Rate it
Recommended because you added Hierarchical Linear Models to your Shopping Cart (edit)

5. 

Multilevel Modeling (Quantitative Applications in the Social Sciences)
by Douglas A. Luke
Publication Date: July 8, 2004

Our Price: $16.95

Add to cart

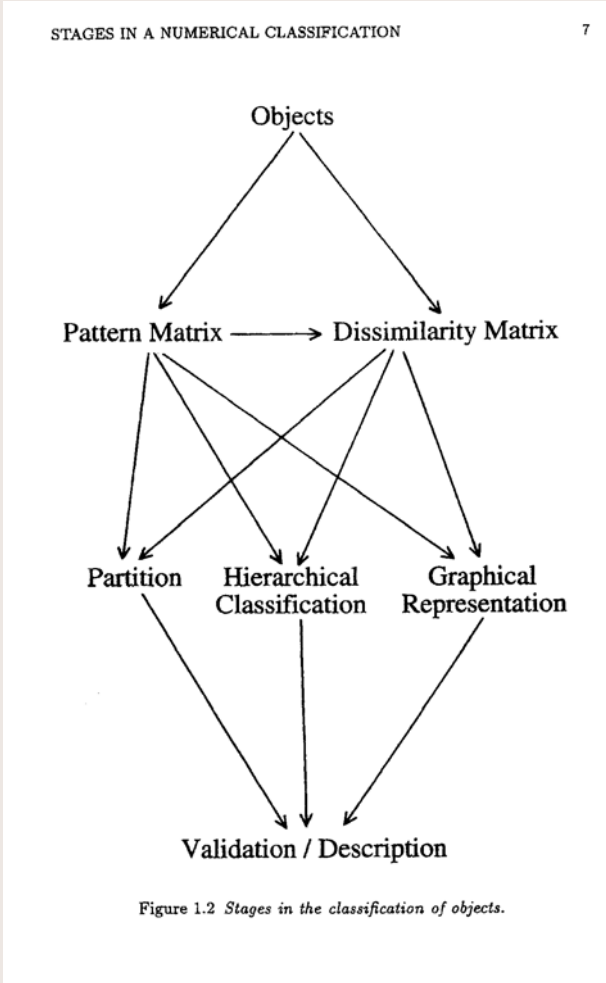
Classification Examples

- Do you have any examples where you thought classification would be appropriate?

Aims of Classification

- Reduce complexity in data.
  - Summarize the types of objects and the groups they represent
  - Data mining.
- Understand phenomena under study.
- Reduce data heterogeneity.
  - Are there groups of outliers in your data?
- Any other aims you can think of?

Stages in Classification (Gordon, 1999)



Common Questions Regarding Classification

1. How should the objects for analysis be selected?
2. Which variables should be used to describe the objects?
3. Should any standardization or differential weighting of variables be undertaken?

Common Questions Regarding Classification

4. How should a relevant measure of dissimilarity be constructed from a pattern matrix?
5. Which clustering and graphical procedures should be used in the analysis of the data?
6. How should the results of the study be summarized?

Magnitude of the Problem

- Clustering algorithms make use of measures of similarity (or alternatively, dissimilarity) to define and group variables or observations.
- Clustering presents a host of technical problems.
- For a reasonable sized data set with *n* objects (either variables or individuals), the number of ways of grouping *n* objects into *k* groups is a Sterling number of the second kind:

$$\left(\frac{1}{k!}\right)\sum_{j=0}^{k} -1^{j-k}\binom{k}{j} \quad j^{n}$$

What This Means

- Imagine you have a deck of cards.
- How many ways could you cluster the cards into four groups?

- 845,099,323,305,172,000,000,000,000,000

- How many of these clusters are meaningful?

Numerical Problems

- In theory, one way to find the best solution is to try each possible grouping of all of the objects – an optimization process called integer programming.
- It is difficult, if not impossible, to do such a method given the state of today's computers (although computers are catching up with such problems).

Numerical Problems

- Rather than using such brute-force type methods, a set of heuristics have been developed to allow for fast clustering of objects in to groups.
- Such methods are called heuristics because they do not guarantee that the solution will be optimal (best), only that the solution will be better than most.

Clustering Method Inputs

- The inputs into clustering methods are in the form of measures of similarities or dissimilarities.
- The result of the heuristic depends in large part on the measure of similarity/dissimilarity used by the procedure.

Measures of Distance

- Care must be taken with choosing the metric by which similarity is quantified.
- Important considerations include:
  - The nature of the variables (e.g., discrete, continuous, binary).
  - Scales of measurement (nominal, ordinal, interval, or ratio).
  - The nature of the matter under study.

Clustering Variables vs. Clustering Observations

- When variables are to be clustered, oft used measures of similarity include correlation coefficients (or similar measures for non-continuous variables).

- When observations are to be clustered, distance metrics are often used.

Euclidean Distance

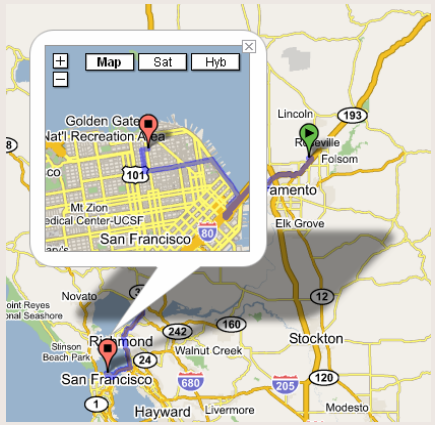
- Euclidean distance is a frequent choice of a distance metric:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_p-y_p)^2} \\ &= \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} \end{aligned}$$

- This distance is often chosen because it represents an understandable metric.
- But sometimes, it may not be best choice.

Euclidean Distance?

- Imagine I wanted to know how many miles it was from my old house in Sacramento to Lombard Street in San Francisco…



- Knowing how far it was on a straight line would not do me too much good, particularly with the number of one-way streets that exist in San Francisco.

Other Distance Metrics

- Other popular distance metrics include the Minkowski metric:

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$$

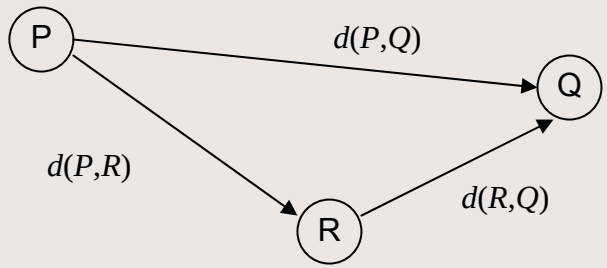
- The key to this metric is the choice of *m*:
  - If *m* = 2, this provides the Euclidean distance.
  - If *m* = 1, this provides the "city-block" distance.

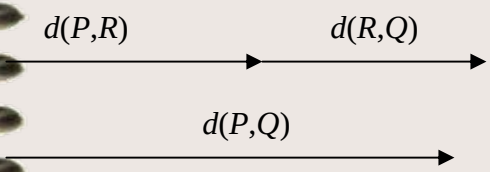
Preferred Distance Properties

- It is often desirable to use a distance metric that meets the following properties:
  - $d(P,Q) = d(Q,P)$
  - $d(P,Q) > 0$ if $P \neq Q$
  - $d(P,Q) = 0$ if $P = Q$
  - $d(P,Q) \leq d(P,R) + d(R,Q)$
- The Euclidean and Minkowski metrics satisfy these properties.

Triangle Inequality

- The fourth property is called the triangle inequality, which often gets violated by non-routine measures of distance.
- This inequality can be shown by the following triangle (with lines representing Euclidean distances):





Binary Variables

- In the case of binary-valued variables (variables that have a 0/1 coding), many other distance metrics may be defined.
- The Euclidean distance provides a count of the number of mismatched observations:

	Variables				
	1	2	3	4	5
Item i	1	0	0	1	1
Item k	1	1	0	1	0

- Here, $d(i,k) = 2$
- This is sometimes called the Hamming Distance.

Other Binary Distance Measures

- There are a number of other ways to define the distance between a set of binary variables (as shown on p. 674).
- Most of these measures reflect the varied importance placed on differing cells in a 2 x 2 table.

General Distance Measure Properties

- Use of measures of distance that are monotonic in their ordering of object distances will provide identical results from clustering heuristics.
- Many times this will only be an issue if the distance measure is for binary variables *or* the distance measure does not satisfy the triangle inequality.

Statistical Distances

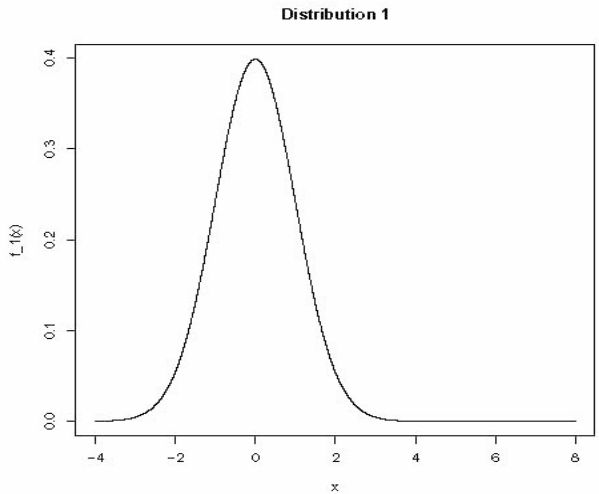
- Statistical distances are often used in model-based procedures for clustering.
  - The main metric of a statistical distance is a likelihood value.
- These distances often satisfy the triangle inequality.

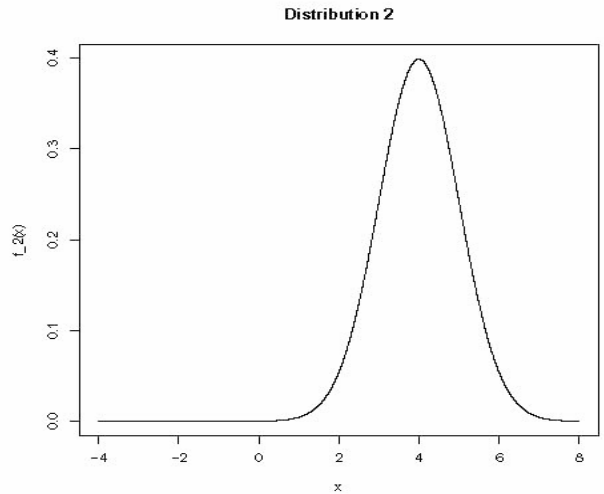
Likelihoods

- A simple way of thinking about the statistical likelihood of an observation, denoted $L(x)$, is to consider your standard normal distribution.

Likelihoods

- This likelihood is represented by using the probability density function of a statistical distribution (the functional form of the distribution).
- To demonstrate, consider the plot of the following two distributions:





Likelihoods

- Both of these distributions are normal, but Distribution 1 is N(0,1) and Distribution 2 is N(4,1).
- Functionally, the likelihood for an observation x of being in Distribution 1 is given by the normal density:

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$$

- Functionally, the likelihood for an observation x of being in Distribution 2 is given by the normal density:

$$f_2(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-(x-4)^2/2\right)$$

- Imagine we encounter an observation, $x = 3$, what would the likelihood be for this observation coming from Distribution 1 or 2?

Likelihoods

- The likelihood of this observation coming from Distribution 1 is given by:

$$f_1(3) = \frac{1}{\sqrt{2\pi}} \exp\left(-3^2/2\right) = 0.004$$

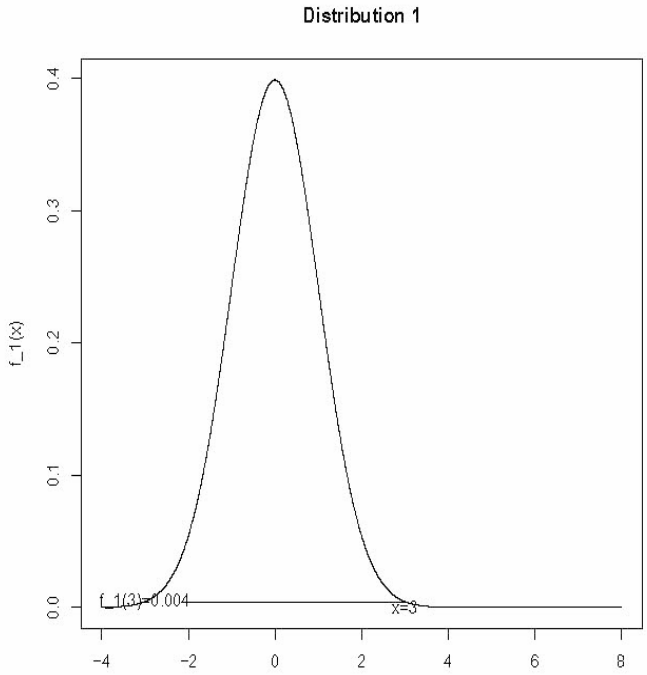
- The likelihood of this observation coming from Distribution 2 is given by:

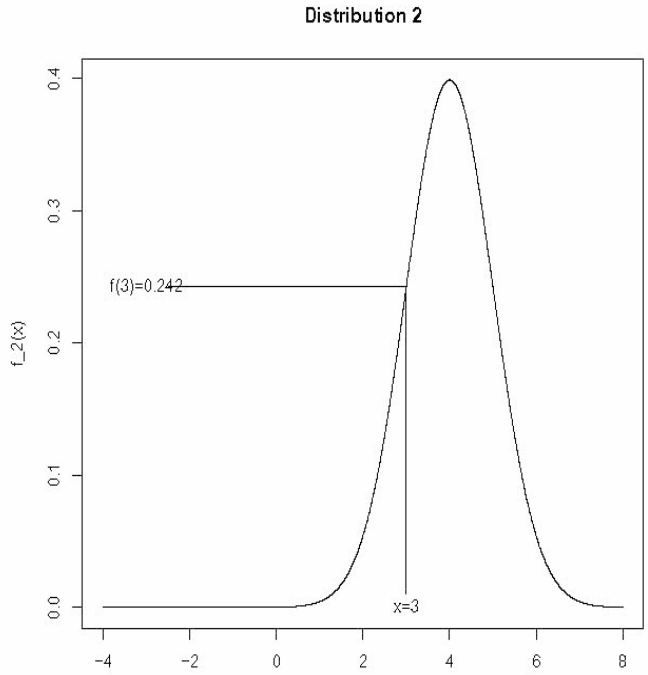
$$f_2(3) = \frac{1}{\sqrt{2\pi}} \exp\left(-(3-4)^2/2\right) = 0.242$$

- Therefore, we would say that this observation *most likely* came from Distribution 2.

Likelihoods

- The process we just saw is akin to determine the value of the $y - axis$ for the point x:





Other Statistical Distances

- Functions of statistical distances, however, do not usually satisfy the triangle inequality.
  - For instance, the Kullback-Liebler Distance is defined for discrete distributions as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

  - And for continuous distributions as:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

Kullback-Leibler Continued

$$D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P),$$