# BAN 5743

Dr. Goutam Chakraborty

# Outline

- Software used in BAN5743
- A quick recap of predictive modeling (machine learning) approach

# Software Used in This class

- Some base SAS and SAS EG
- A lot of SAS Enterprise Miner in class demos
  - Access via VMware Horizon client (desktop.okstate.edu)
    - Login using okey email and password
    - Use Spears lab image
    - Use windows search box to find SAS Enterprise Miner while in Spears lab
    - Save all work on M: drive (each of you have M drive access and should be able to save stuff there from Vmware – capacity 2 GB)
- Some SAS Viya  towards end of the semester and in some labs

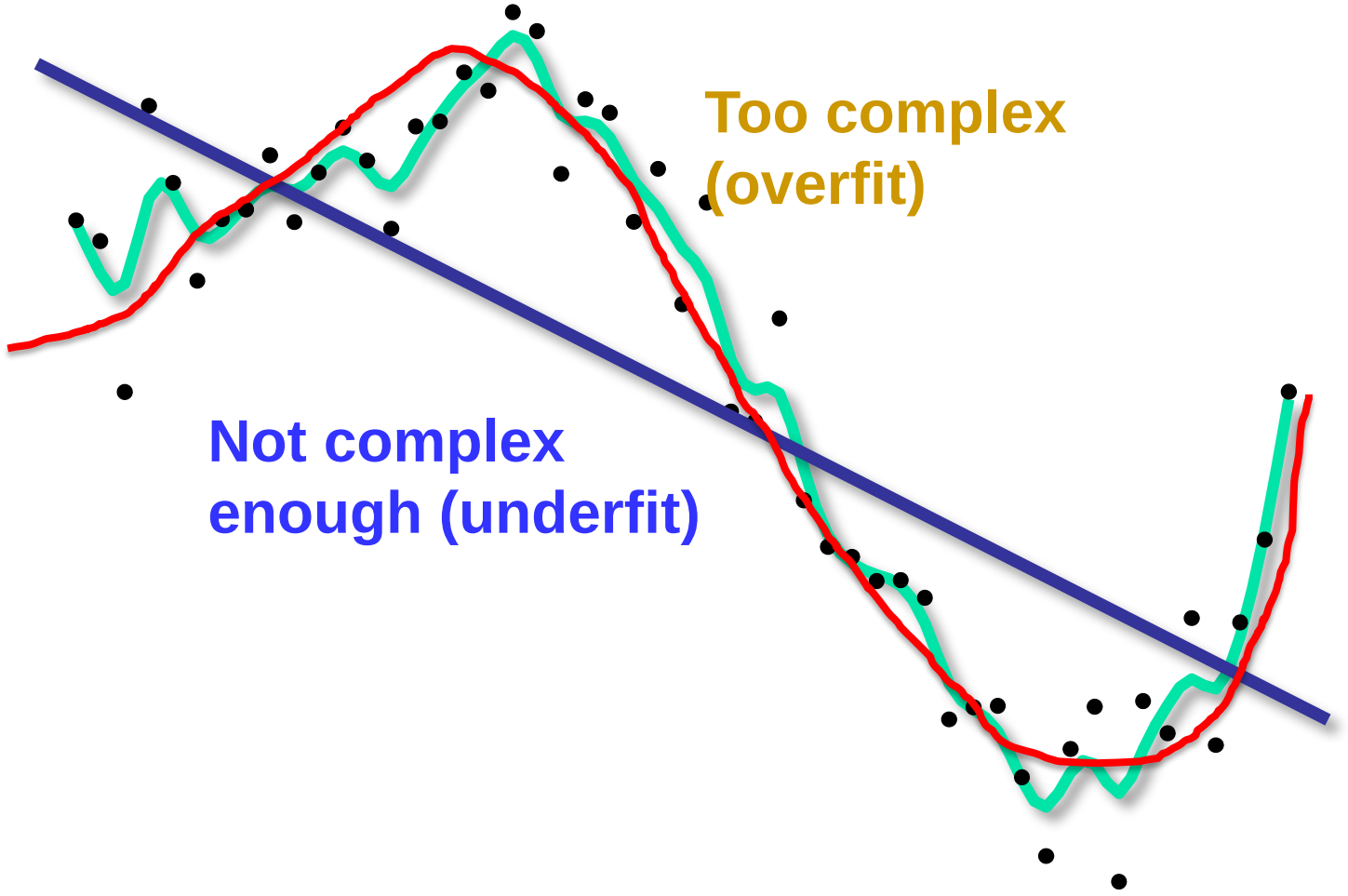
# My take on software

- SAS EM is a point-and-click interface and makes your life very easy to run complicated analysis
  - It's also an expensive software and many companies do not have it. Or, if they do, they have limited number of licenses
- So, while learning the tool is important, it's more critical that you learn the concepts/mechanics of each technique
  - Most of the times companies want to hire students from our program assuming that you are deeply trained (not just being able to do point-and-click)
    - So, dig in and learn concepts/mechanics
    - And, learn to use SAS, R and Python code to replicate SAS EM results

# Predictive Models: Statistical vs. Machine Learning Approach

- Statistical Approach
  - Explanation more important than prediction
  - Smaller data size, limited number of variables
  - Statistical significance via p-value is the dominant philosophy
- Machine Learning approach
  - Prediction more important than explanation
  - Large data size, many variables
    - Need different methods for variable selection
  - Model performance on unseen data is the dominant philosophy
    - Handled via data partitioning
    - Prevents model overfitting

# Model Complexity




...

# Outline

- Machine learning approach to predictive modeling requires you **to first complete**
  - Data handling (data wrangling):
    - Missing values
    - Outliers
    - Transformations
  - Creating data split (training and validation)

# Data Exploration

- GIGO principle is very important in analytics projects
- It is imperative that as an analyst, you spend substantial time (before running any modeling) on checking, exploring and understanding data.
- At the minimum explore following:
  - Summary statistics (including minimum and maximum) for interval variables
    - Missing values, extreme values, skewed distribution, etc.
  - Categories (classes) for nominal variables
    - Rare classes create problems
  - Plots (many choices)
    - Univariate (understand nature of distribution of a variable)
    - Bivariate (explore relationship between 2 variables)
    - Multivariate (explore relationship between 3 or more variables)

# Overview of  Methods and Algorithms for Segmentation

Dr. Goutam Chakraborty

# Outline

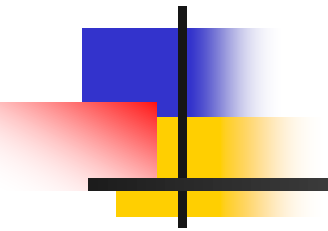
- Assumption about things you should know (or, go back to those notes and refresh your memory)
- Broad overview of different methods and alogrithms used for segmentation

# Make Sure you Refresh Your Knowledge About

- What is STP and where does segmentation fit in marketing strategy?
- How has segmentation concepts evolved over time?
- Criteria used to strategically evaluate effectiveness of segmentation?
- Bases vs. Descriptors of segments
- Different types of segmentation such as:
  - Targeting vs. Foundational, Needs, behavior or Value-based
- Types of segmentation variable used
- Classification of different types of segmentation methods

# Methods and Algorithms Used in Segmentation

- RFM (Recency, Frequency, Monetary) cells for behavioral segmentation
  - Very popular in Internet/Direct marketing world
  - Very simple algorithm – all you need is ability to sort data and do grouping
  - Quite a bit of arbitrariness in deciding number of cells
- Statistical and Machine Learning methods
  - Hierarchical clustering (many methods such as average, centroid, ward's …)
  - Non-Hierarchical clustering (most common: k-means)
  - Expectation-Maximization algorithm
  - Self Organizing map (SOM)
  - K-Nearest Neighbors algorithm (KNN)



# Pattern Recognition, Similarity and Distance

Dr. Goutam Chakraborty

# Outline

- Basic issues in pattern recognition via natural grouping of data
- Conceptual basis of similarity
- Distance metrics used to measure similarity

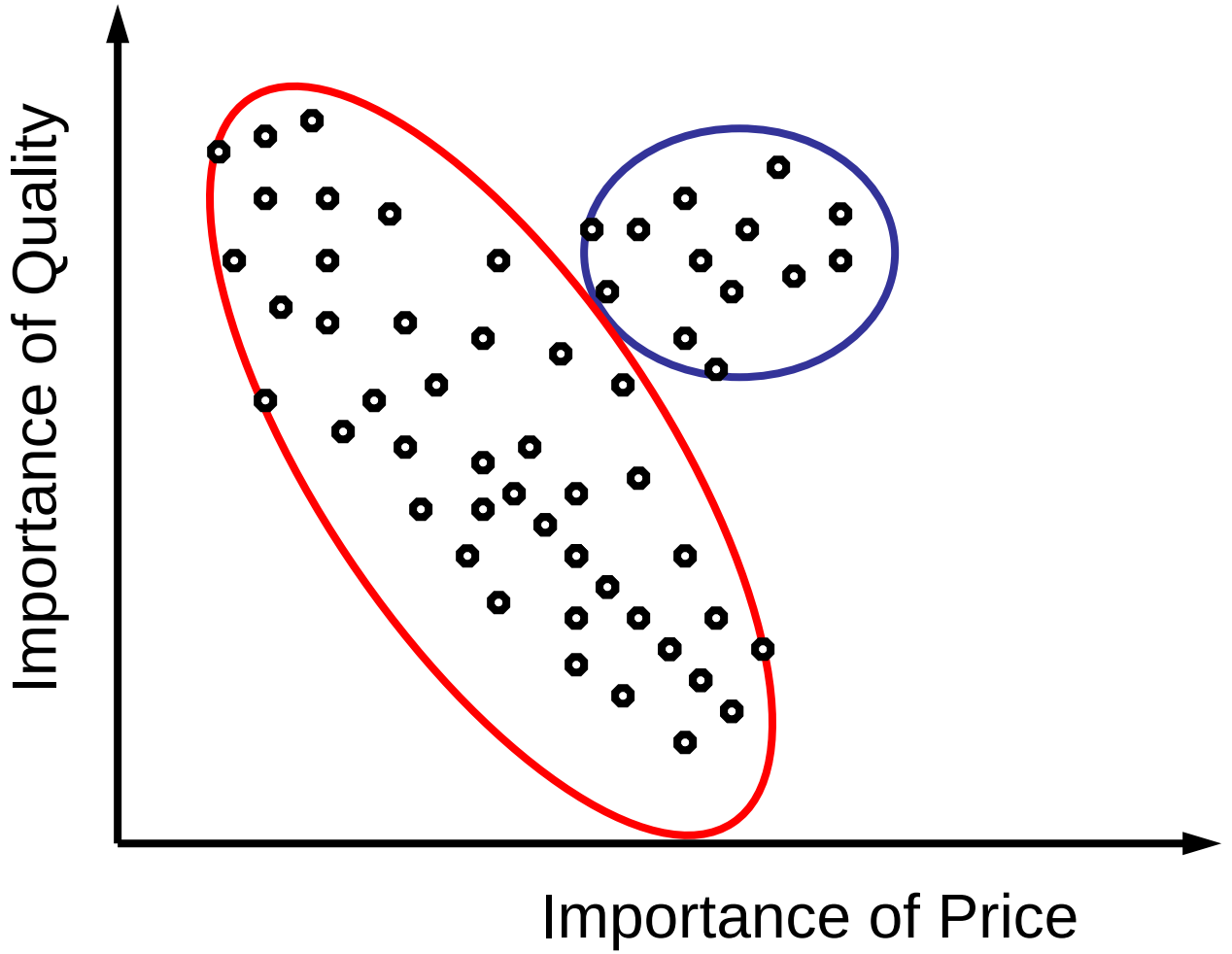
# Pattern Recognition or, Unsupervised Classification

- Pattern recognition or, unsupervised classification of entities/objects is a basic human conceptual activity and a fundamental requirement for development of *scientific* knowledge.
  - *Cluster analysis* is the generic name for a wide variety of procedures that can be used to create an unsupervised classification of entities/objects.
  - It has been referred to as Q analysis, typology construction, classification analysis, unsupervised pattern recognition, and numerical taxonomy.
  - The essence of all these approaches is the unsupervised classification of data as suggested by “natural” groupings (pattern) of the data themselves.

# Natural Grouping in Data:
# A Geometrical View of an Ideal Pattern

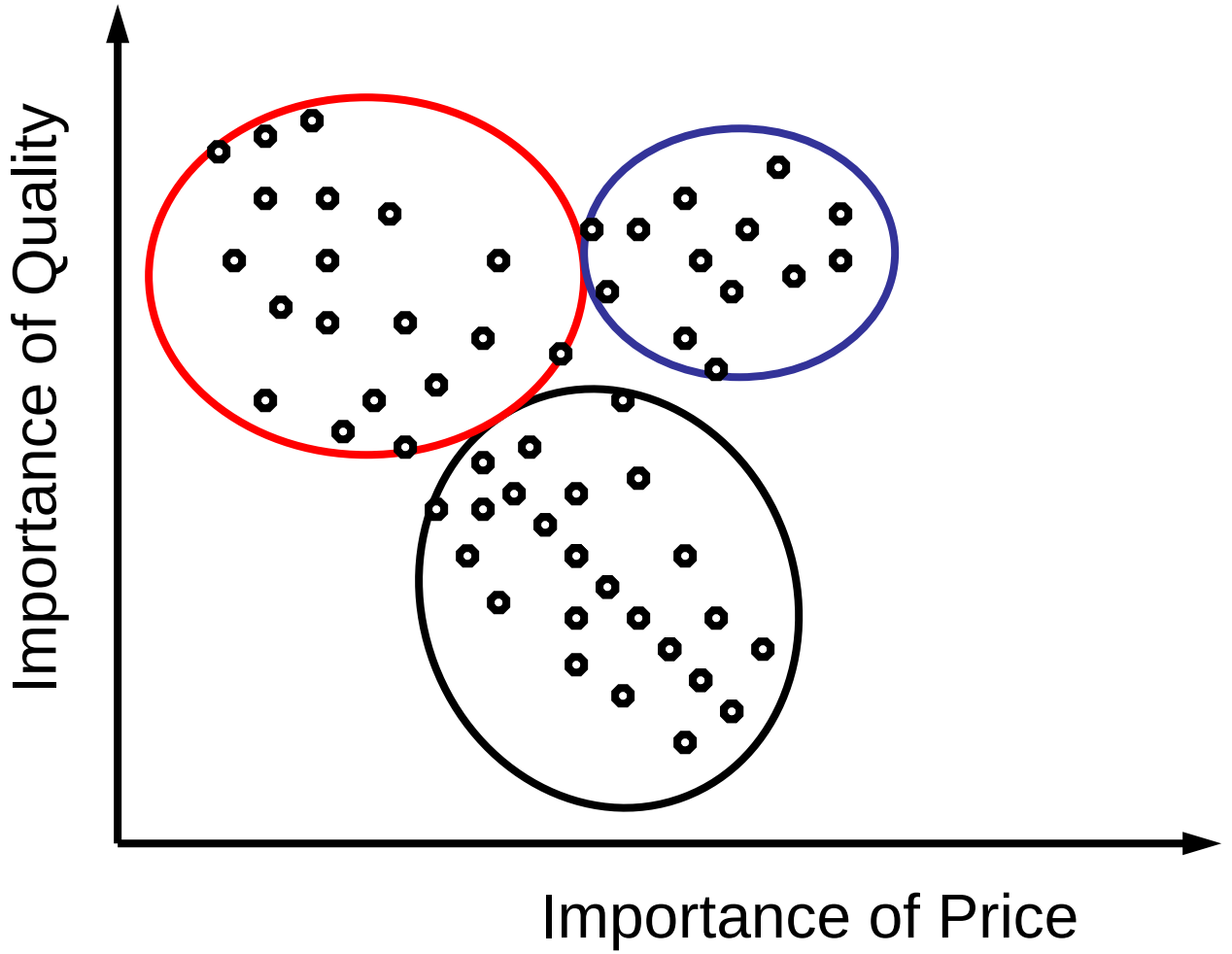
# Natural Grouping in Data: A Geometrical View of Pattern Likely to Be Found in Practice

# How Many Groups and Which Observation Belongs in Which Group?



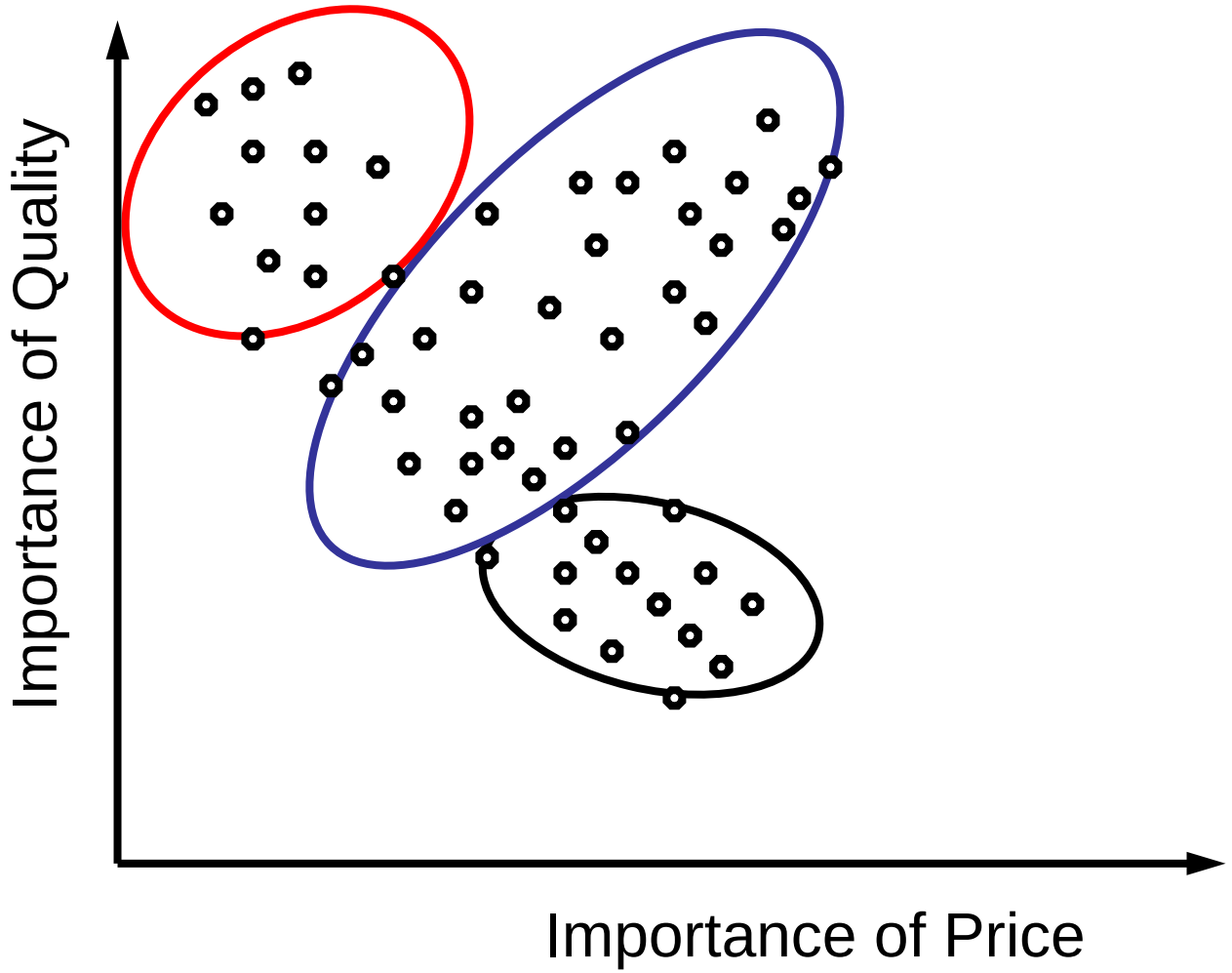

*continued...*

# How Many Groups and Which Observation Belongs in Which Group?




*continued...*

# How Many Groups and Which Observation Belongs in Which Group?



# Clustering, Segmentation, and a Deck of Cards!

# Concept of Similarity and Distance

- To identify natural groups, we must first define a measure of **similarity** (proximity) between objects/entities.
  - Assume variables (axes in space) are numeric.
  - Then, if two things are similar, they should be close to each other in the space.
    - That is, the distance between them should be ….
  - But, if two things are dissimilar, they should be well separated from each other in the space.
    - That is, the distance between them should be ….
  - A collection of similar things would therefore likely result in more cohesive (homogenous) groups than a collection of dissimilar things.

# Types of Distance Metrics Used in Clustering for Numeric Variables

- Euclidean distance
- City-block (Manhattan) distance
- Mahalanobis distance

# Examples of Distance Calculations – Three Customers, Two Variables

- Assume we have the following data from three customers about what they consider important in buying a product (rated on a scale with 1=not at all important and 11=extremely important).

| Customer | Price | Quality |
|---|---|---|
| A | 9 | 8 |
| B | 5 | 9 |
| C | 6 | 8 |

Euclidean distance between A and B is $\sqrt{(9-5)^2+(8-9)^2}$= 4.123

Euclidean distance between B and C is $\sqrt{(5-6)^2+(9-8)^2}$= 1.414

Euclidean distance between A and C is $\sqrt{(9-6)^2+(8-8)^2}$= 3

# Examples of Euclidean Distance Calculations

What if we also had data on a third variable, say importance of service, from the same three customers?

| Customer | Price | Quality | Service |
|---|---|---|---|
| A | 9 | 8 | 7 |
| B | 5 | 9 | 8 |
| C | 6 | 8 | 9 |

Euclidean distance between A and B is $\sqrt{(9-5)^2+(8-9)^2+(7-8)^2}$ = 4.243

Euclidean distance between B and C is $\sqrt{(5-6)^2+(9-8)^2+(8-9)^2}$ = 1.732

Euclidean distance between A and C is $\sqrt{(9-6)^2+(8-8)^2+(7-9)^2}$ = 3.606

# Similarity Measures for Categorical Variables

- A large number of coefficients have been used for measuring similarity between objects characterized by **only** categorical variables. Some of the commonly used measures are
  - Hamming distance
  - Simple matching coefficients
  - Nonmissing mismatches
  - Nonmissing matches
  - Jaccard's coefficients

# When Variables Used for Describing Objects Include Both Numeric and Categorical

This is a likely situation in many practical segmentation problems.

- Unfortunately, in such situations (involving a mix of numerical and categorical variables), complications arise because theoretically it is unclear what “distance” really means!
  - Often the practical solution is to convert the categorical variables into dummy or flag (1/0) variables and then use the dummy (flag) variables along with the other numeric variables in calculating distance metrics (such as Euclidean).
  - Or, convert categorical variables to a numeric variable using WOE (weight-of-evidence) coding.

# An Example with Numeric and Categorical Variables

Suppose we also know customers' marital status, and we would like to use that in our distance calculation.

| Customer | Price | Quality | Service | Marital Status |
|---|---|---|---|---|
| A | 9 | 9 | 7 | Single |
| B | 5 | 9 | 8 | Married |
| C | 6 | 8 | 9 | Divorced |

Convert marital status to dummy variables and use those in distance calculations.

| Customer | Price | Quality | Service | Single | Married | Divorced |
|---|---|---|---|---|---|---|
| A | 9 | 9 | 7 | 1 | 0 | 0 |
| B | 5 | 9 | 8 | 0 | 1 | 0 |
| C | 6 | 8 | 9 | 0 | 0 | 1 |

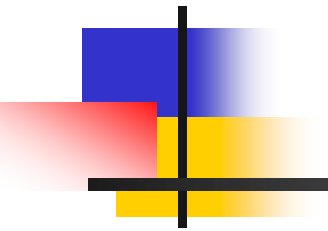
Euclidean distance between A and B is

$$\sqrt{(9-5)^2+(9-9)^2+(7-8)^2+(1-0)^2+(0-1)^2+(0-0)^2} = 4.359$$

# Challenges with Different Measurement Scales of Variables

What's the problem with using different measurement scales for variables used as input into clustering?

- Euclidean distance metric assume equal weight to each input variables.
  - But, in reality, the input variables with a wider scale of measurement (more variance) get weighted more in determining distances.
- So, the solution is standardization.
  - Many different ways of doing this.
  - *Range standardization* (where each input variable is scaled by first subtracting its minimum value and then dividing by its range) is often preferred for segmentation problems instead of *normal standardization* (0 mean, 1 standard deviation).



# Hierarchical Clustering

Dr. Goutam Chakraborty

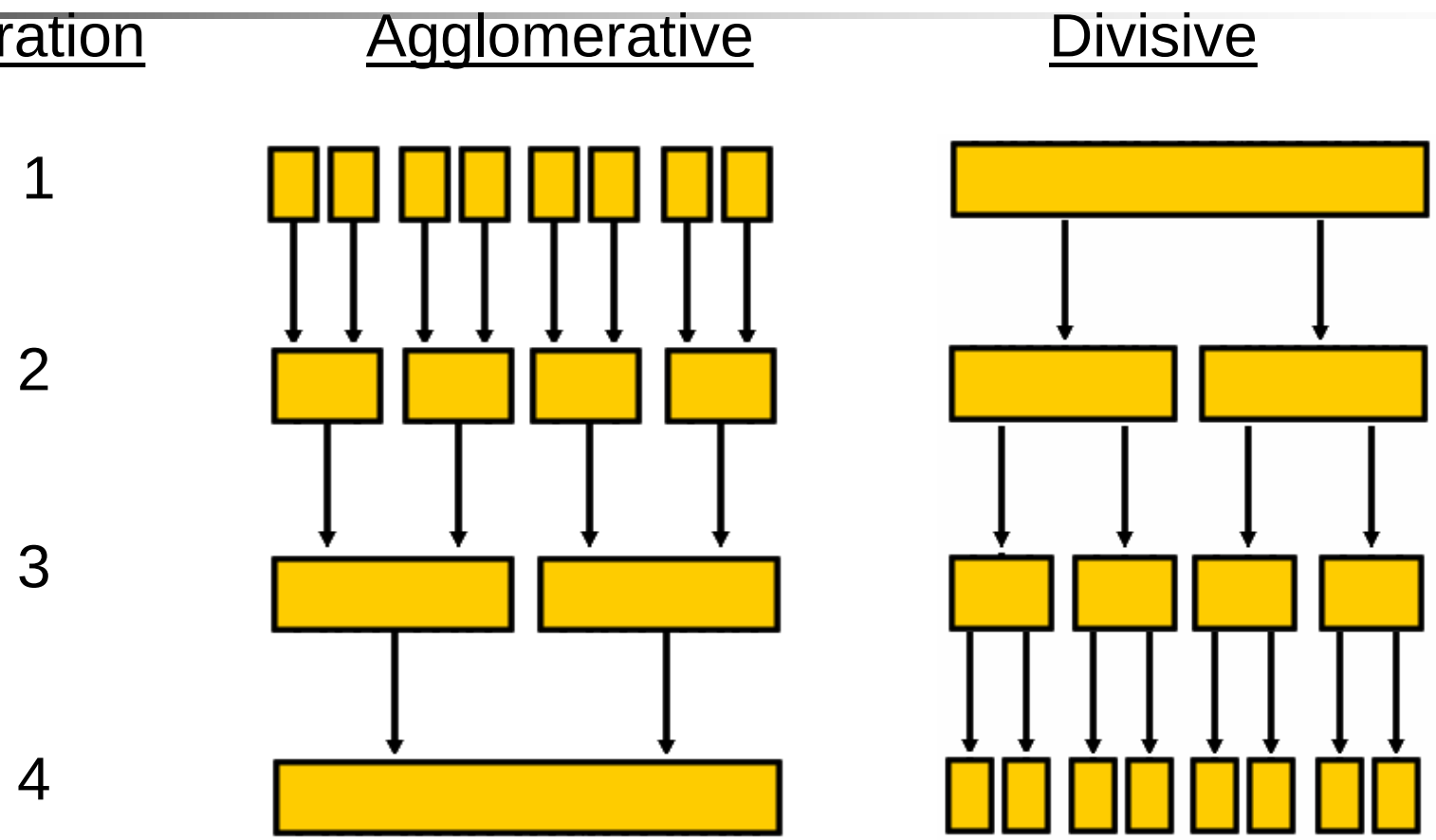
# Outline

- Differences between hierarchical and nonhierarchical clustering.
- Agglomerative versus divisive hierarchical clustering.
- Mechanics behind agglomerative hierarchical clustering.

# Types of Clustering

- Hierarchical clustering
  - Average, Centroid, Ward's method, others
- Nonhierarchical clustering (or, partitive clustering)
  - K-means, SOM, EM and others

# Two Types of Hierarchical Clustering



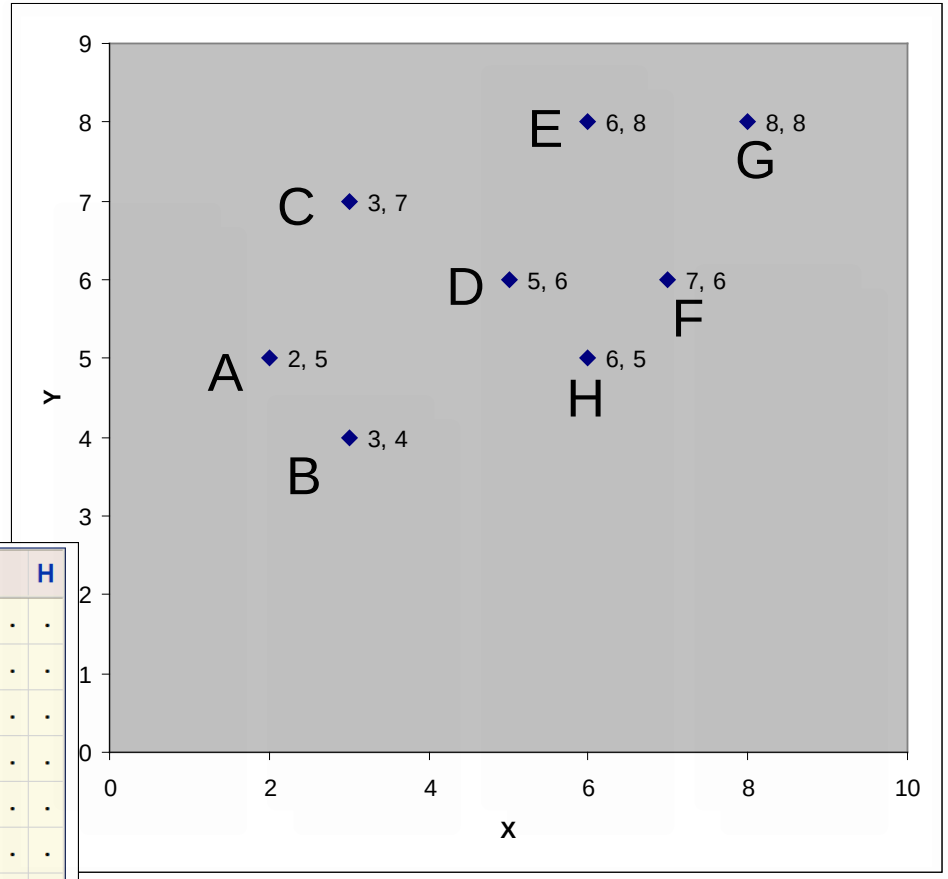

Source: Applied Clustering Techniques Course Notes, SAS Institute, Inc., Cary, NC.

# Numerical Example of Clustering (Agglomerative, Single Linkage)

- Data from eight subjects (A, B, C, D, E, F, G, H) on two variables, X and Y.

| ID | X | Y |
|---|---|---|
| A | 2 | 5 |
| B | 3 | 4 |
| C | 3 | 7 |
| D | 5 | 6 |
| E | 6 | 8 |
| F | 7 | 6 |
| G | 8 | 8 |
| H | 6 | 5 |

| ID | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | 0.00000 | . | . | . | . | . | . | . |
| B | 1.41421 | 0.00000 | . | . | . | . | . | . |
| C | 2.23607 | 3.00000 | 0.00000 | . | . | . | . | . |
| D | 3.16228 | 2.82843 | 2.23607 | 0.00000 | . | . | . | . |
| E | 5.00000 | 5.00000 | 3.16228 | 2.23607 | 0.00000 | . | . | . |
| F | 5.09902 | 4.47214 | 4.12311 | 2.00000 | 2.23607 | 0.00000 | . | . |
| G | 6.70820 | 6.40312 | 5.09902 | 3.60555 | 2.00000 | 2.23607 | 0.00000 | . |
| H | 4.00000 | 3.16228 | 3.60555 | 1.41421 | 3.00000 | 1.41421 | 3.60555 | 0 |



# Agglomerative Clustering Steps

| Step | Minimum Distance Between Unclustered IDs | ID Pair | Cluster Membership | Number of Clusters | Overall Heterogenity Measure (Average Within Cluster Distance) |
|---|---|---|---|---|---|
| Initial | | | A B C D E F G H | 8 | 0 |
| | | | | | |
| 1 | 1.414 | A,B | (**AB**) C D E F G H | 7 | 1.414 |
| 2 | 1.414 | H,D | (**AB**) C (**DH**) E F G | 6 | 1.414 |
| 3 | 1.414 | H,F | (**AB**) C (**DHF**) E G | 5 | 1.561 |
| 4 | 2 | G,E | (**AB**) C (**DHF**) (**GE**) | 4 | 1.648 |
| 5 | 2.236 | F,E | (**AB**) C (**DHGEF**) | 3 | 2.266 |
| 6 | 2.236 | A,C | (**ABC**) (**DHGEF**) | 2 | 2.32 |
| 7 | 2.236 | D,C | (**ABCDHGEF**) | 1 | 3.343 |

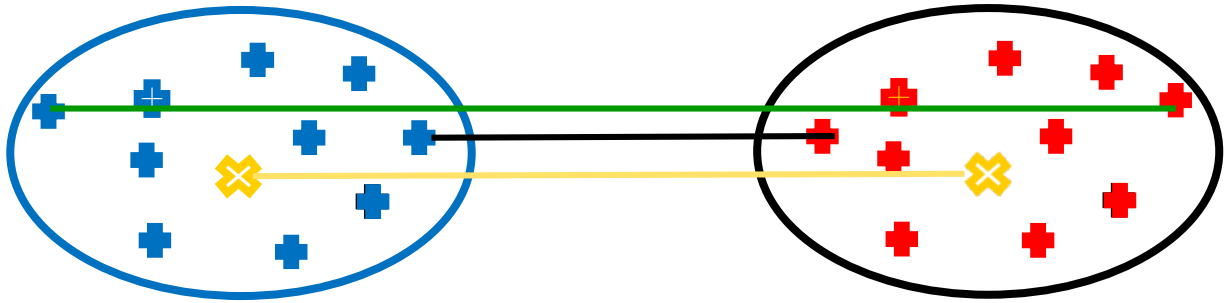
| ID | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | 0.00000 | . | . | . | . | . | . | . |
| B | 1.41421 | 0.00000 | . | . | . | . | . | . |
| C | 2.23607 | 3.00000 | 0.00000 | . | . | . | . | . |
| D | 3.16228 | 2.82843 | 2.23607 | 0.00000 | . | . | . | . |
| E | 5.00000 | 5.00000 | 3.16228 | 2.23607 | 0.00000 | . | . | . |
| F | 5.09902 | 4.47214 | 4.12311 | 2.00000 | 2.23607 | 0.00000 | . | . |
| G | 6.70820 | 6.40312 | 5.09902 | 3.60555 | 2.00000 | 2.23607 | 0.00000 | . |
| H | 4.00000 | 3.16228 | 3.60555 | 1.41421 | 3.00000 | 1.41421 | 3.60555 | 0 |

...

# Distance Between Clusters in Agglomerative Algorithm

- Single linkage (nearest neighbor)
- Complete linkage (farthest neighbor)
- Average linkage
- Centroid method
- Ward’s method

# What's the Distance Between the Clusters? (This help us in Joining Clusters)



Single Linkage : Shortest Distance

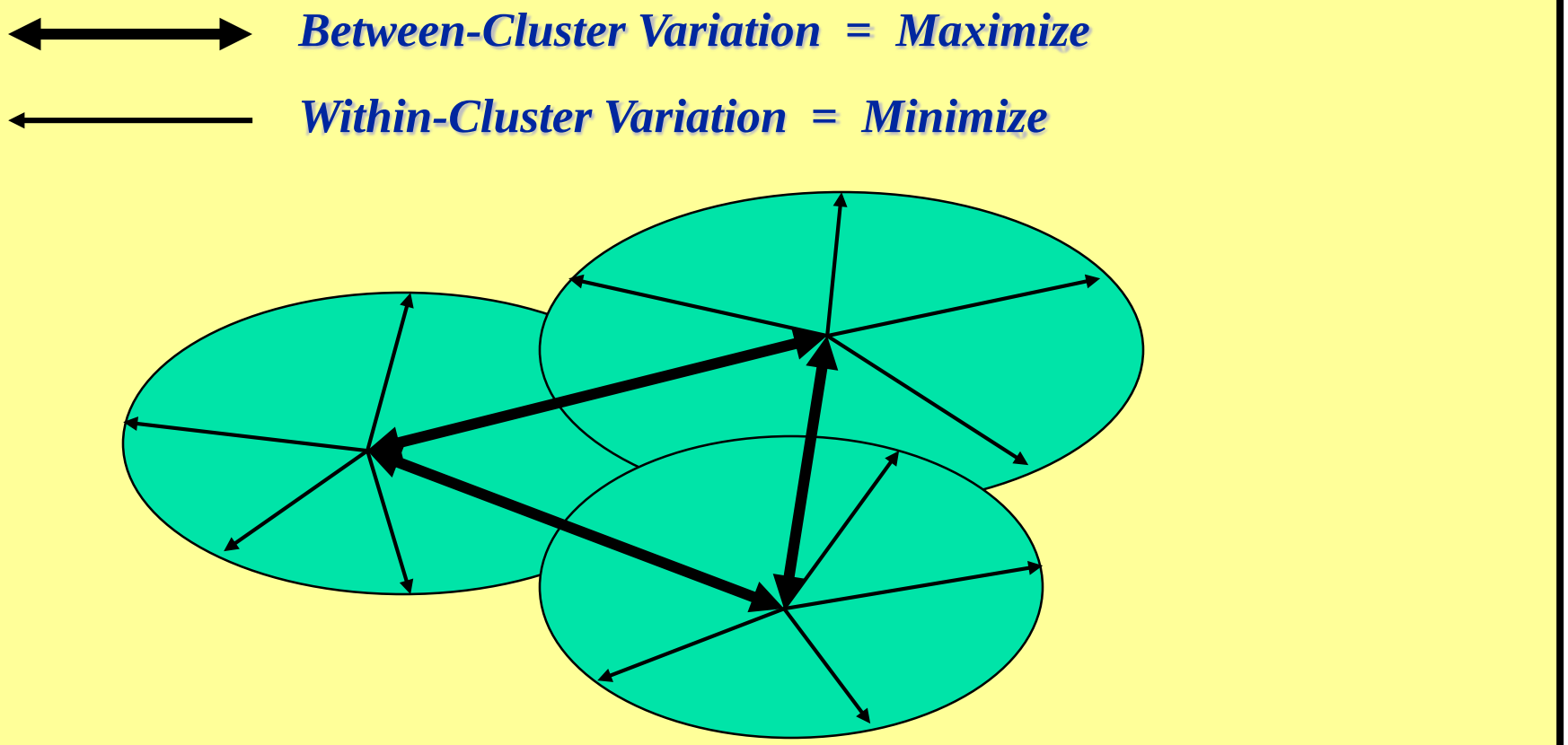
Complete Linkage : Maximum Distance

Average Linkage : Average of all pairwise distances

Centroid Method: Distance between cluster centers

...

# Ward's Method (Variance Based)



# Which Is the Best Method?

- And the winner is….

# Four Fundamental Questions in Hierarchical Clustering

Q1: Which variables should be used for clustering?

Q2: How do we define similarity between observations?

Q3: How should clusters be joined?

Q4: How many clusters?

# Demo of Hierarchical Clustering (Ward's)

Dr. Goutam Chakraborty

# Outline

- Use Ward's method to conduct cluster analysis.
- Interpret cluster analysis results from running Ward's method.
- Interpret cluster profiles by comparing the variable means of each cluster to others, as well as to the overall mean.
  - First, we will handle profiling clusters with base variables
  - Later, we will profile with descriptors and other managerially relevant variables

# SAS EG Steps

- Use Analyze > Multivariate > Cluster > Select *reliab_dstz …talk_dir_dstz* as Analysis variables, *cust_id* as Identifying label  and *all variables* as copy variables > Cluster method as Ward's Minimum variance > Results >Check simple summary statistics, Check Tree output
- I will then modify SAS EG code to do following:
  - ODS Graphics ON/Imagemap=on;
  - Plots (unpack)=PSF
  - Plots (unpack)=PST2
  - Plots (unpack) =CCC;
- Look at results, select number of cluster and then modify SAS Code again to change *Nclusters*

# Summary of Clustering Using Ward's Method

Ward's minimum variance method results summary:

- Cluster history shows no single observation joining with clusters at the late stages.
- The number of clusters suggested by the pseudo *t*-square plot is 6+1=7 or 8+1=9.
- CCC all negative (cannot be used!)
- Pseudo F shows no peak!
- The number of clusters suggested by a *big drop* in RSQ (or, *increase* in SPRSQ) is 6, 4, or 3.
- SAS Enterprise Guide does not print average cluster distance for Ward's method. So, we will make a managerial decision **to go with 4 clusters for now**!

# Profiling Clusters : The Big Questions

Several types of questions are often asked in profiling:

- How is the average member *of one cluster* different from an average member *of a different cluster*?
- How is the average member *of any cluster* different from the average member *of the entire data*?
- How does the *distribution* of a variable *within a clus*ter compare to the *distribution* of the same variable in the *entire data*?
- Which variables are *most important predictors* for each cluster?

# Profiling Clusters with Bases

- Profiling involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
  - For **numeric** variables, this involves
    - comparing the mean of each variable across clusters
    - comparing the mean of each variable in a cluster with the mean for the same variable for the entire data
    - comparing the distribution (histogram) of each variable in a cluster with the distribution of the same variable for the entire data
  - For **categorical** variables, this involves comparing % members in each category within a cluster with the % members in the same category for the entire data

# SAS EG Step

- Use disjoint cluster data from proc tree
- Describe > One way frequencies > Select cluster as Analysis Variable
- ANOVA > One-Way ANOVA > Select all _*dst*z variables as Dependent Variables > Cluster as Independent Variable > Means > breakdown > select Mean > Run
  - Optional: Copy and paste table at the end of SAS Output in Excel

# Cluster Profiles using Double Standardized Base Variable Means

| CLUSTER | Reliab dstz | Time dstz | Av_br dstz | Av_spec dstz | Price dstz | Credit dstz | Av_pay dstz | Return dstz | Warranty dstz | Talk_dir dstz |
|---|---|---|---|---|---|---|---|---|---|---|
| **Overall(.)** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1.0** | -0.4 | -0.5 | 0.1 | -0.2 | -0.1 | 0.2 | 1.0 | 0.1 | 0.0 | -0.2 |
| **2.0** | 0.1 | 0.0 | 0.6 | 0.4 | 0.3 | 0.1 | -0.7 | 0.5 | 0.3 | -0.2 |
| **3.0** | 0.2 | 0.4 | -1.0 | 0.0 | 0.1 | 0.1 | -0.5 | 0.5 | 0.5 | 0.4 |
| **4.0** | 0.5 | 0.5 | 0.2 | 0.0 | -0.2 | -0.3 | -0.5 | -0.9 | -0.6 | 0.3 |

# Profiles Using Double-Standardized Indexes

| CLUSTER | Reliab index | Time index | Av_br index | Av_spec index | Price index | Credit index | Av_pay index | Return index | Warranty index | Talk_dir index |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall(.) | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 1.0 | 60 | 51 | 107 | 79 | 89 | 117 | 195 | 106 | 102 | 81 |
| 2.0 | 105 | 104 | 163 | 140 | 130 | 108 | 32 | 147 | 131 | 76 |
| 3.0 | 121 | 136 | 0 | 97 | 115 | 111 | 52 | 149 | 146 | 141 |
| 4.0 | 147 | 147 | 115 | 98 | 83 | 68 | 53 | 6 | 36 | 127 |

**Cluster 1:** Care most about availability of payment options and credit.

**Cluster 2:** Care most about availability of brands, availability of detailed specifications, price, return policy, and warranty policy.

**Cluster 3:** Care most about return, warranty, and ability to talk directly to salesperson, timeliness and reliability.

**Cluster 4:** Care most about reliability, timeliness of delivery and talk directly to salesperson.