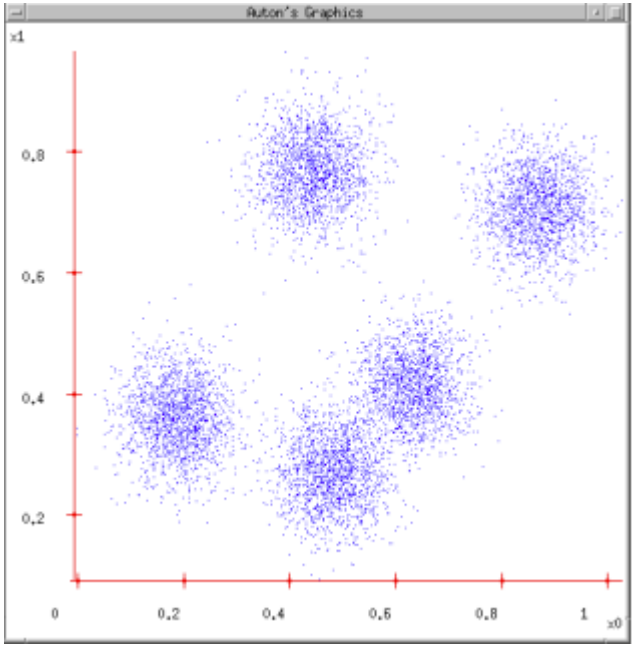
Clustering

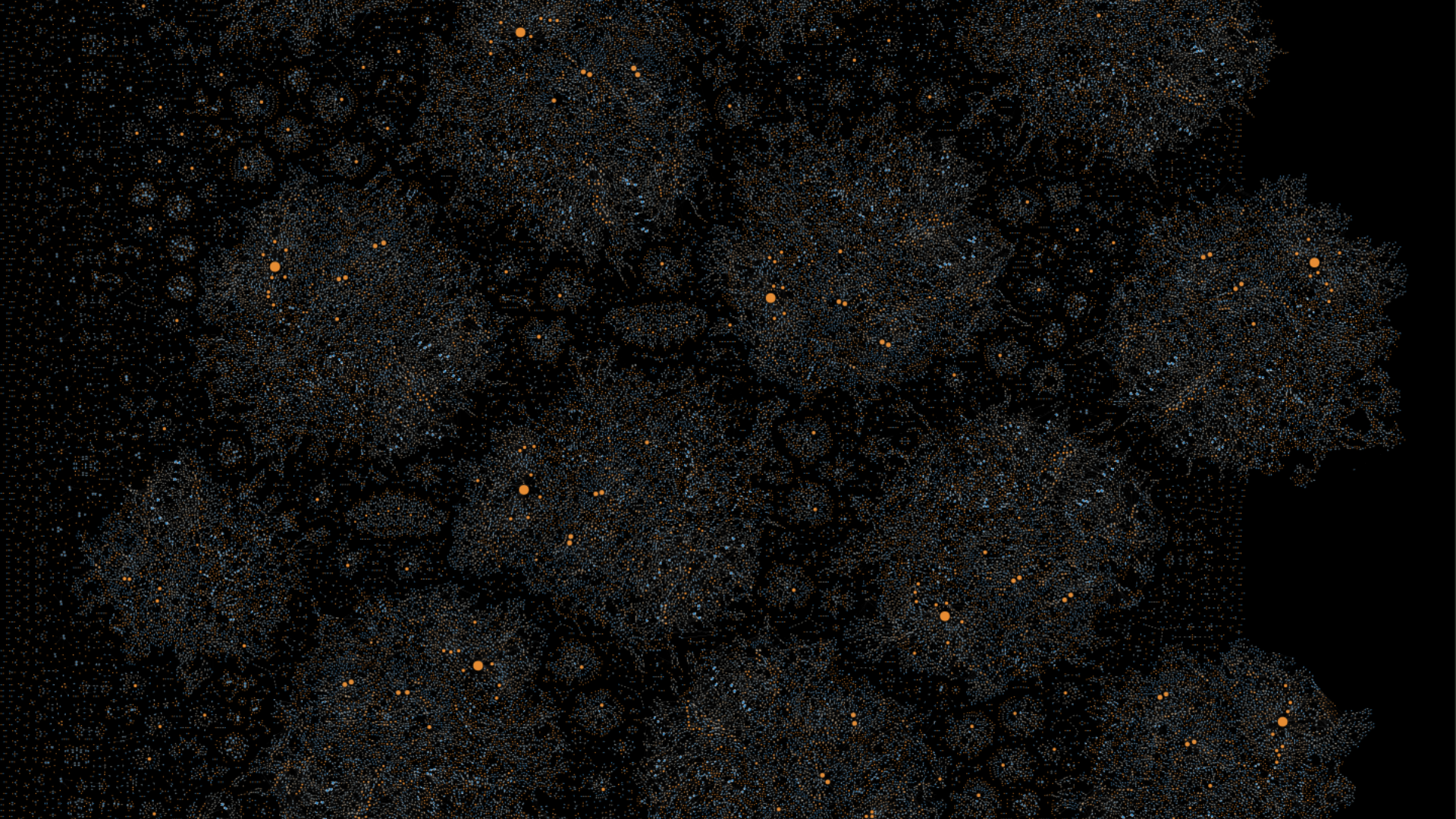
Overview and intro to clustering and Go

Ross Hendrickson
@savorywatt

Introduction

- Types of Learning
- Hierarchical clustering
- K-Means Clustering
- K- Nearest Neighbor Clustering
- Build a simple system that uses KNN to cluster users. What type of user is X?


Auton's Graphics
x1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
x0



Why Cluster?

- Medicine
- Market Research
- Image segmentation
- Community analysis
- Customers according to purchase histories
- Genes according to expression profile
- Users according to interests

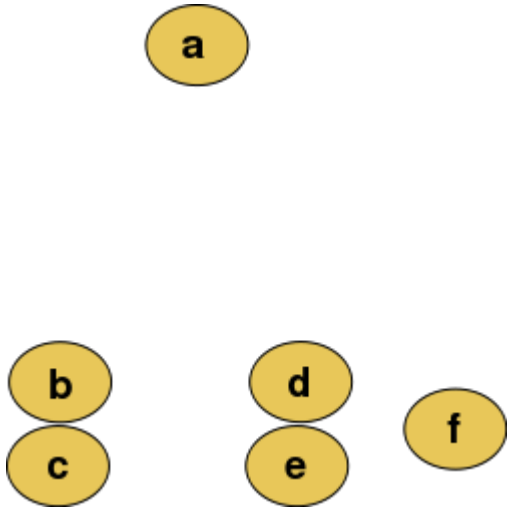
Types of Learning

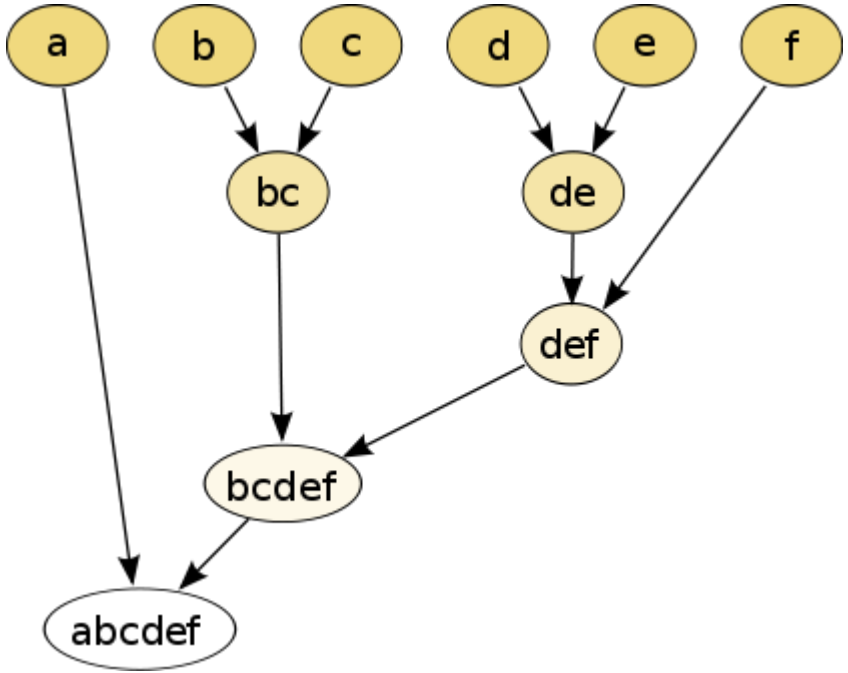
Supervised

Unsupervised

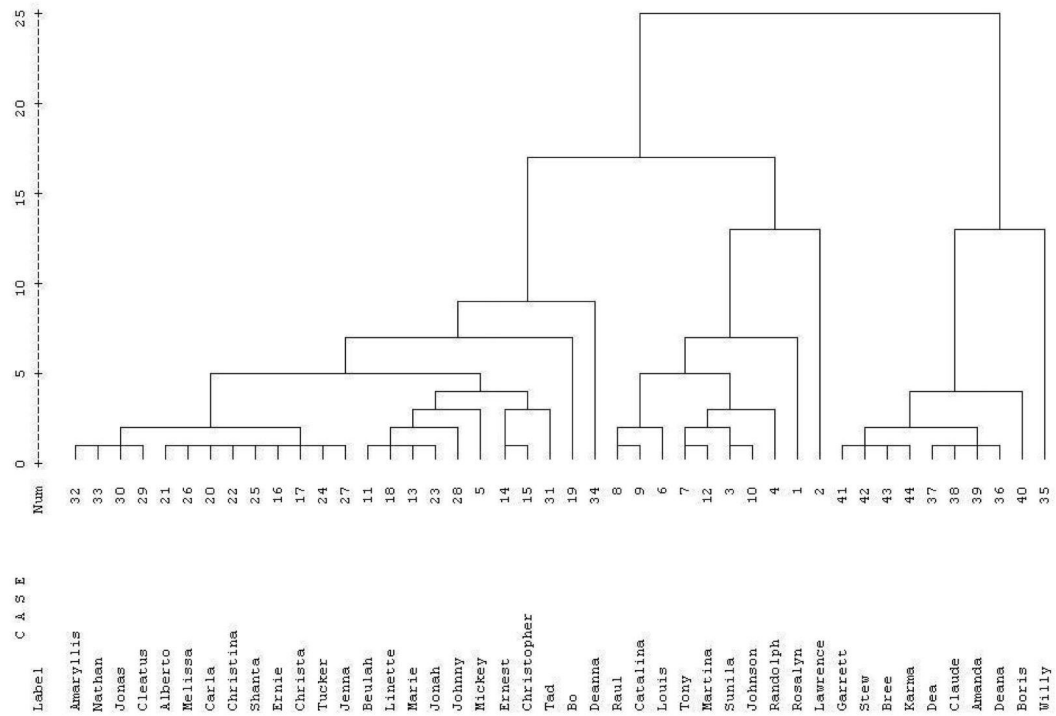
Hierarchical Clustering

- Examples of usage
- Core parts of the algorithm
- How it is useful

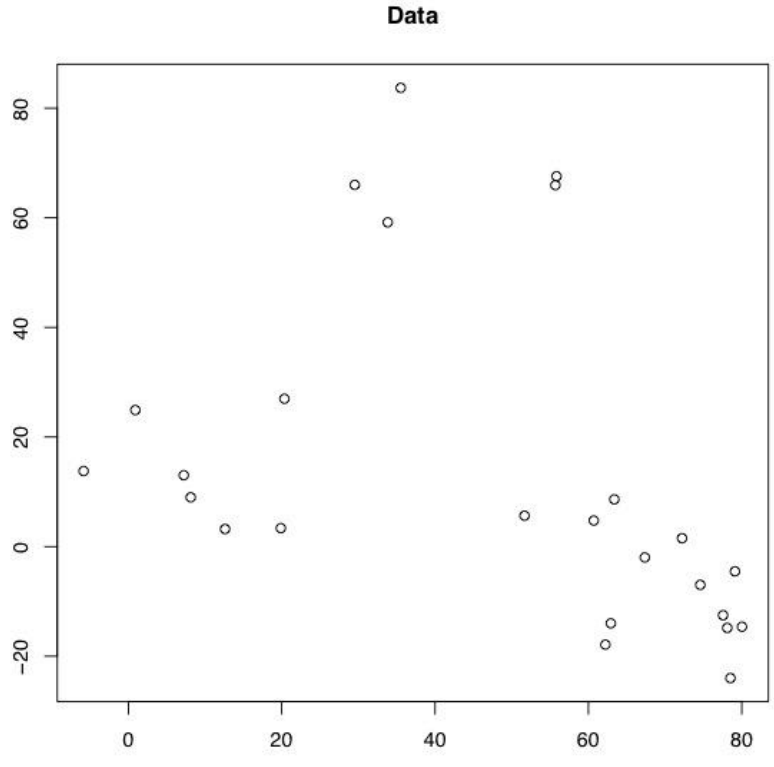

a
b
c
d
e
f


a
b
c
d
e
f
bc
de
def
bcdef
abcdef

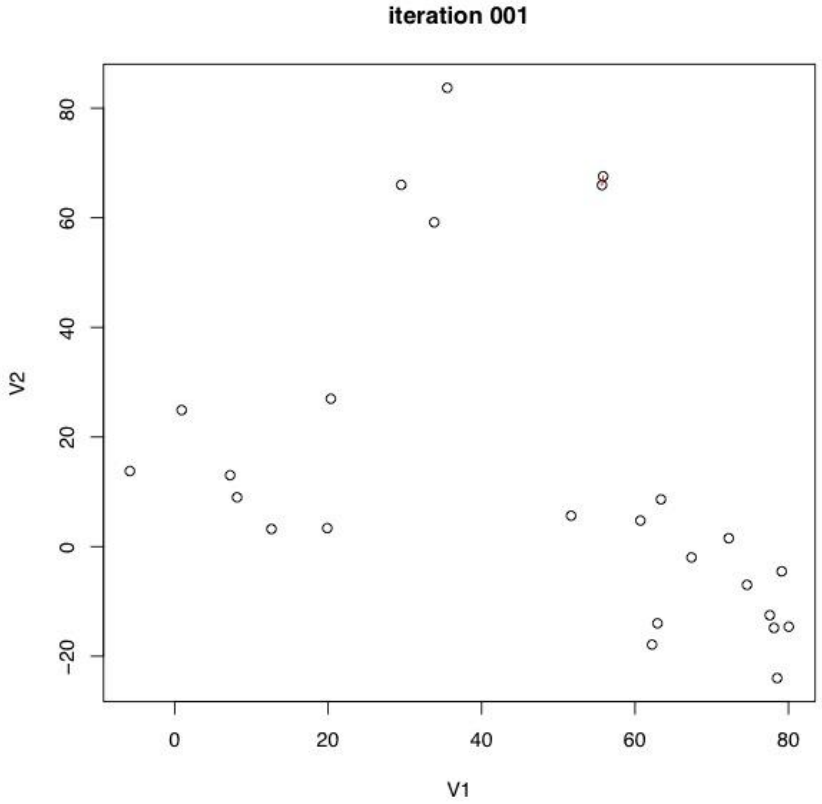
Dendogram


0
5
10
15
20
25
Num
C A S E
Label
32 Amaryllis
33 Nathan
30 Jonas
29 Cleatus
21 Alberto
26 Melissa
20 Carla
22 Christina
25 Shanta
16 Ernie
17 Christa
24 Tucker
27 Jenna
11 Beulah
18 Linette
13 Marie
23 Jonah
28 Johnny
5 Mickey
14 Ernest
15 Christopher
31 Tad
19 Bo
34 Deanna
8 Raul
9 Catalina
6 Louis
7 Tony
12 Martina
3 Sunila
10 Johnson
4 Randolph
1 Rosalyn
2 Lawrence
41 Garrett
42 Stew
43 Bree
44 Karma
37 Dea
38 Claude
39 Amanda
36 Deana
40 Boris
35 Willy

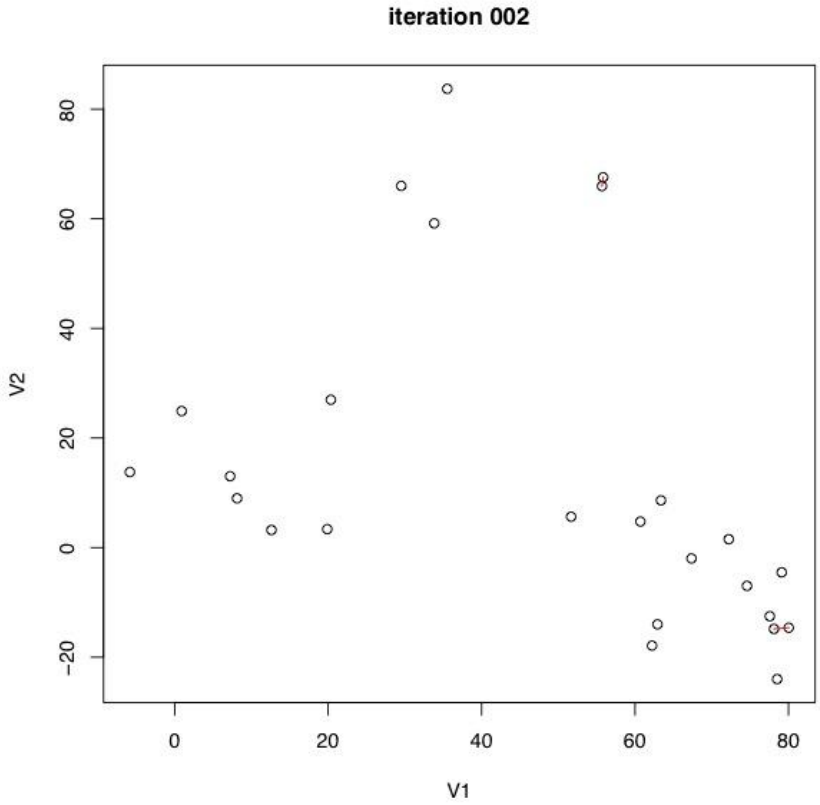
Example



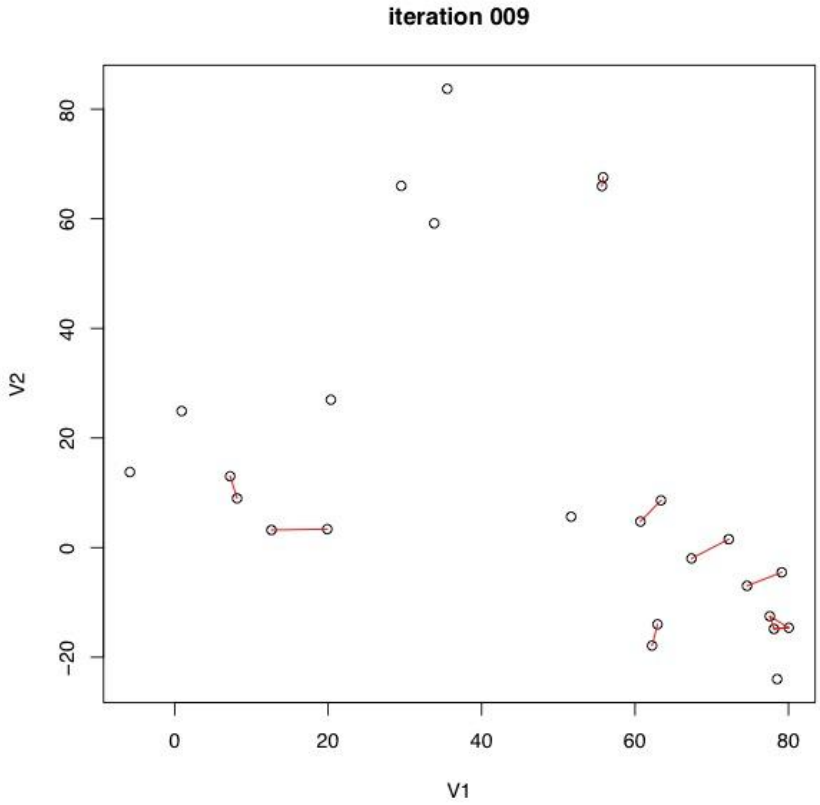
Example



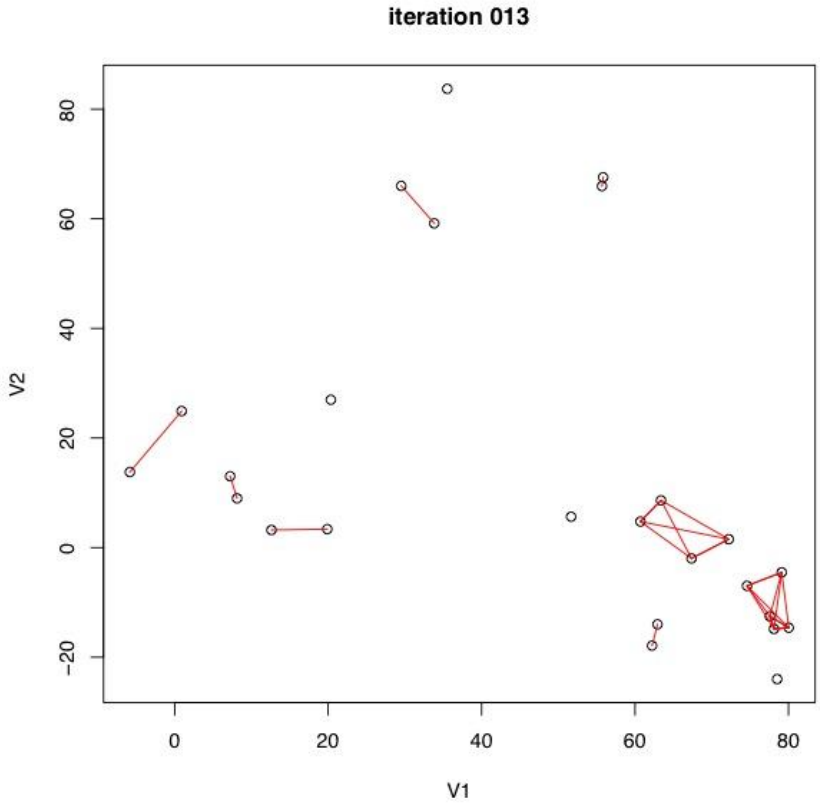
Example



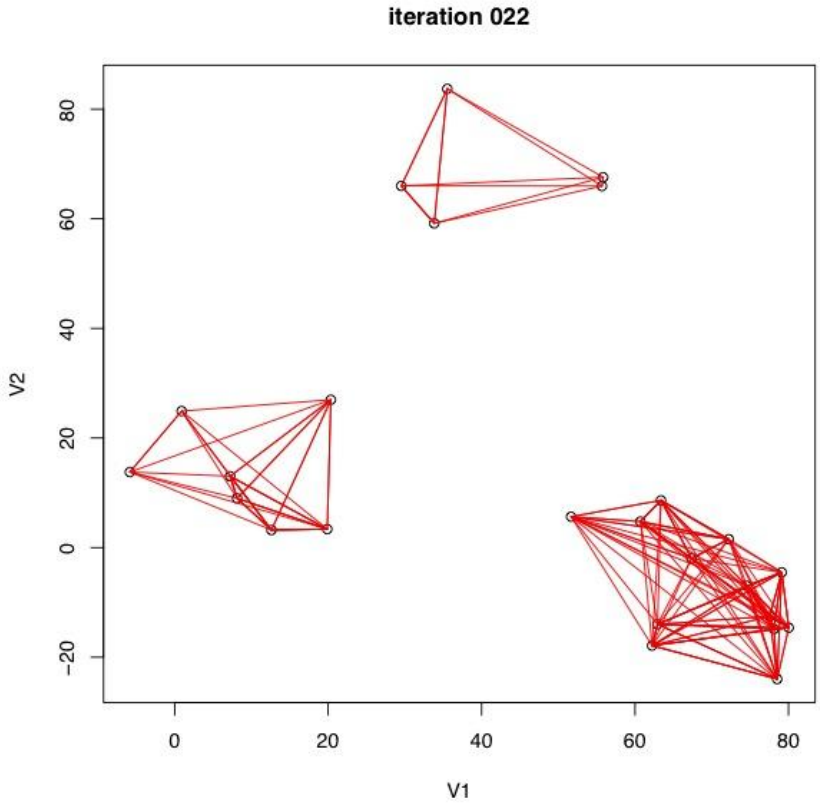
Example



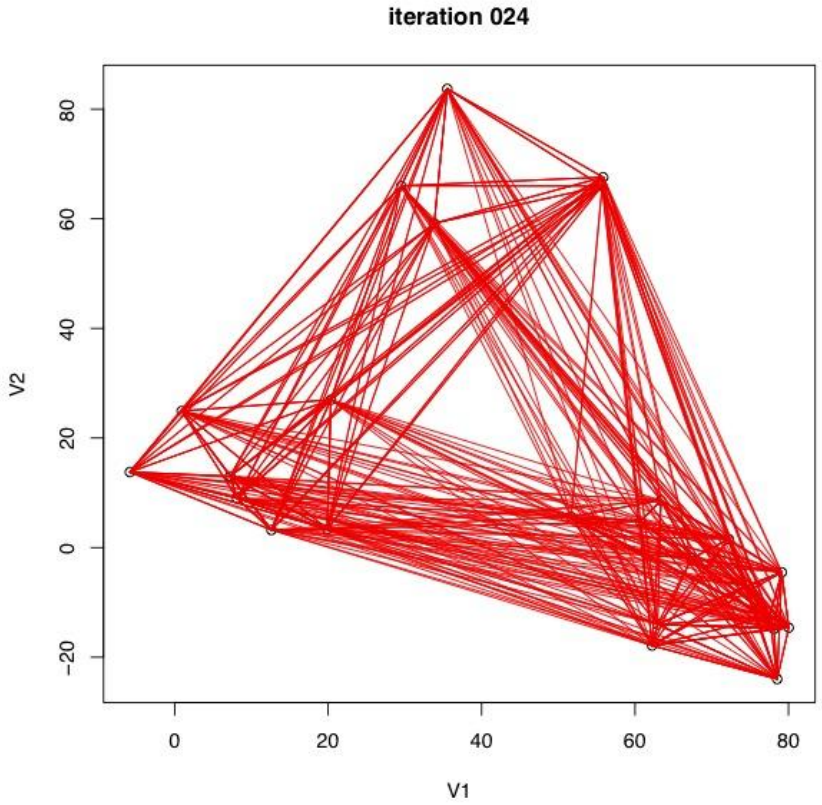
Example



Example



Example

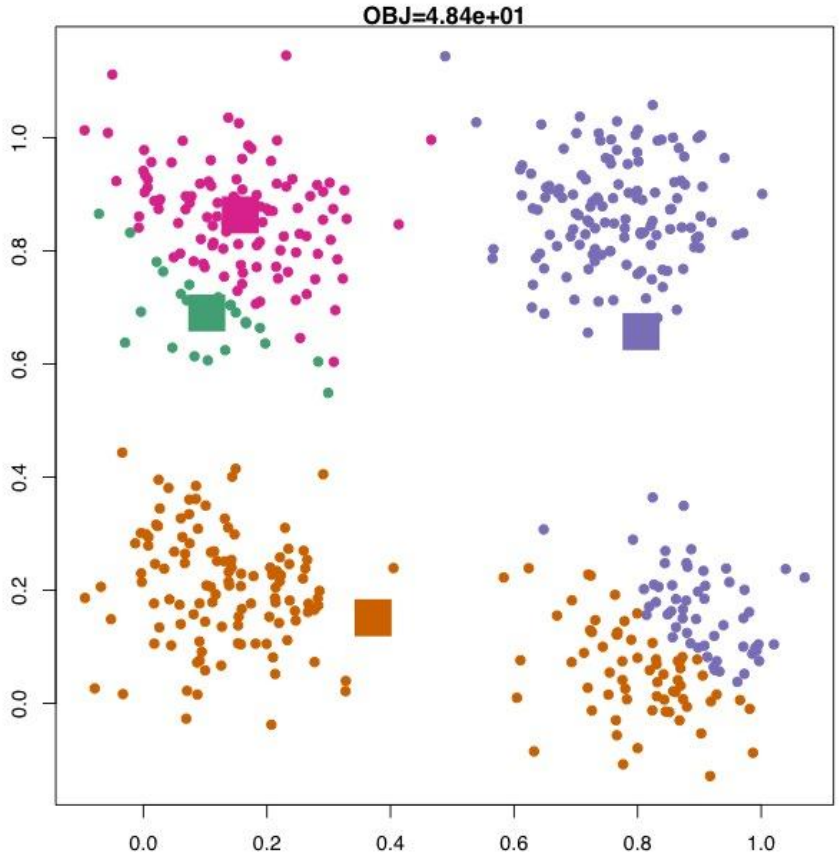


Code

K-Means Clustering

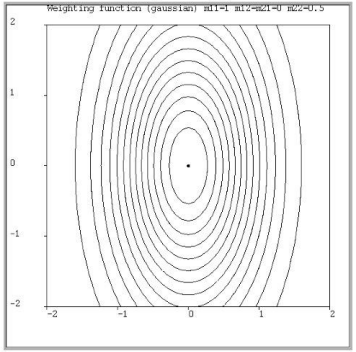
- Top Level
- Examples of usage
- Core parts of the algorithm
- How it is useful

k-means example

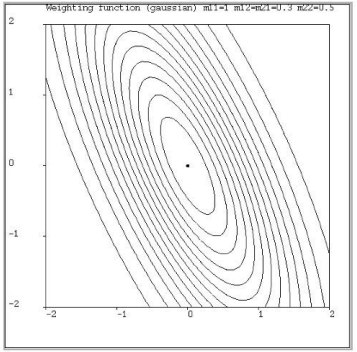


Some notable distance metrics

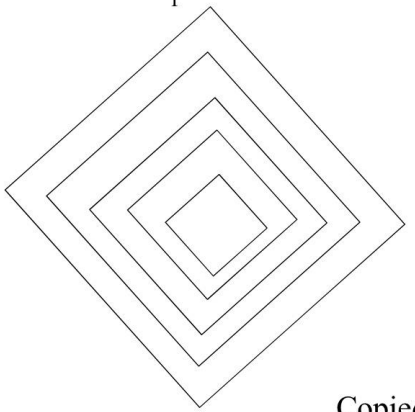
Scaled Euclidean (diagonal covariance)



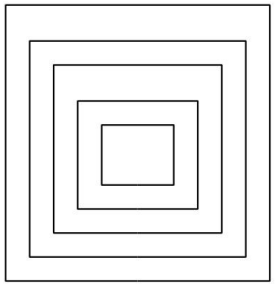
Mahalanobis (full covariance)



L_1 norm



L_∞(max) norm



Copied from Andrew Moore

KNN

- Top Level
- Examples of usage
- Core parts of the algorithm
- How it is useful

KNN

Learning by analogy:

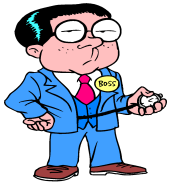
Tell me who your friends are and I'll tell you who you are

A new example is assigned to the most common class among the (K) examples that are most similar to it.

Basic algorithm

- To determine the class of a new example E:
  - Calculate the distance between E and all examples in the training set
  - Select K-nearest examples to E in the training set
  - Assign E to the most common class among its K-nearest neighbors

Basic Example

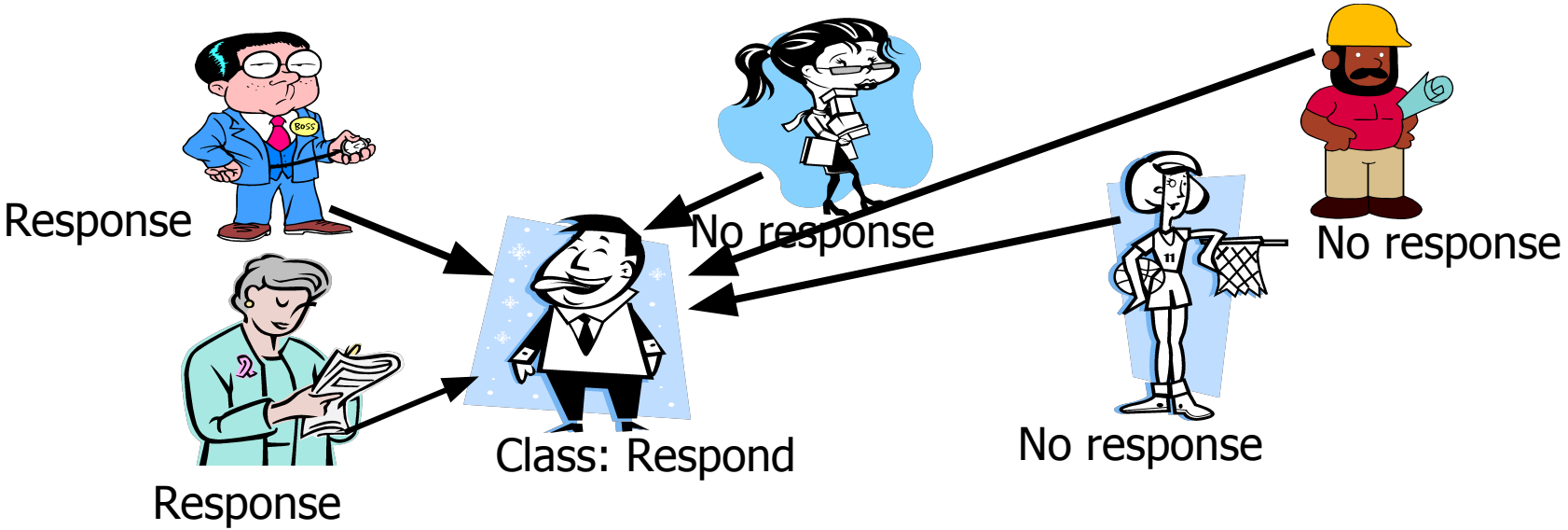


John:
Age=35
Income=95K
No. of credit cards=3

Rachel:
Age=41
Income=215K
No. of credit cards=2

- "Closeness" is defined in terms of the *Euclidean* distance between two examples.
  - The Euclidean distance between $X=(x_1, x_2, x_3,\ldots x_n)$ and $Y=(y_1,y_2, y_3,\ldots y_n)$ is defined as:

Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?


Response
No response
No response
Response
Class: Respond
No response

Strengths

- Simple to implement and use
- Comprehensible – easy to explain prediction
- Robust to noisy data by averaging k-nearest neighbors.
- Some appealing applications (will discuss next in personalization)

Weakness

- Need a lot of space to store all examples.
- Takes more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples).
- Distance between neighbors could be dominated by some attributes with relatively large numbers (e.g., income in our example). Important to normalize some features (e.g., map numbers to numbers between 0-1)

  Example: Income

  Highest income = 500K

  Davis's income is normalized to 95/500, Rachel income is normalized to 215/500, etc.)

Golearn

- SVM
- Linear Regression
- KNN Classification
- KNN Regression
- Neural Network
- Naive Bayes

Example

https://github.com/sjwhitworth/golearn

Live

- go over hcluster
- go over knn implementation
- write main.go
- boom
- load csv
- distance pearson
- benchmark knn
- add in params to point
- predict

Ack

Workiva

Programming Collective Intelligence

Leon Bottou - CMU

David M. Blei - Princeton