Fuzzy Clustering - Harrison Kiang

## Introduction

Many K-partitioning clustering algorithms such as K-Medians, K-Modes, K-Medoids, etc. depend on object cluster assignment and centroid update. That is, to find the cluster assignments of points , to clusters , each iteration has a set of common steps:

1. Assign each point to the nearest cluster
2. For every cluseter , calculate the new position of the respective centroid.

In the cluster analysis part of [CS 412: Introduction to Data Mining](https://www.coursera.org/learn/cs-412/home/welcome) by professor [Jiawei Han](http://hanj.cs.illinois.edu/) from the [University of Illinois at Urbana-Champaign](http://illinois.edu/), there is a lecture in the "Paritioning-Based clustering Methods" section discussing "The K-Medians and K-Modes Clustering Methods". Here is the link: [3.5 The K-Medians and K-Modes Clustering Methods](https://www.coursera.org/learn/cs-412/lecture/pShI2/3-5-the-k-medians-and-k-modes-clustering-methods). An alternative for K-Modes, fuzzy K-Modes, was mentioned: "A **fuzzy K-Modes** method is proposed to calculate a **fuzzy cluster membership value** for each object to each cluster". Unfortunately, this important concept of fuzzy clustering is not covered in detail.

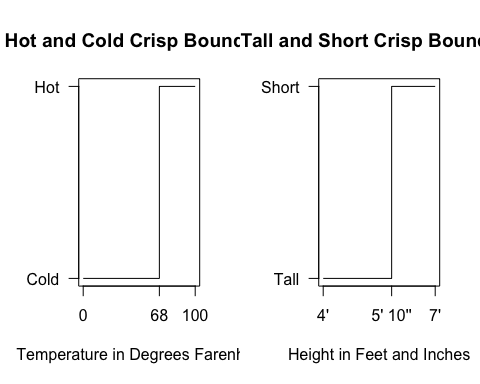
The purpose of this paper is to serve as an overview as to what fuzzy clustering is, and include a comparison of fuzzy clustering and hard clustering. An algorithm employing fuzzy clustering called the Fuzzy c-Means Algorithm will also be discussed.

## Fuzzy Logic

If you ask someone to describe how the temperature feels outside, they may either say it's hot or it's cold. Here, there is no precise mathematical definition for "hot" or "cold". You could also ask someone to describe a person's height. They may say that a person is tall or short. Here, there is no precise mathematical definition for "tall" or "short".

In both cases, it's possible to define a crisp bounday between what is defined as hot or cold, and between what is defined as tall or short. One could say that any temperature less than 68F is cold, and any temperature greater than or equal to 68F is hot. One could also say that any height less than 5 feet 10 inches is short, and any height greater or equal to 5 feet 10 inches is tall. Here, 68 and 5 feet 10 inches are the crisp boundaries for temperature and height, respectively.

par(mfrow=c(1,2))  
x <- c(0,68,100)  
y <- c(0,1,1)  
plot(x,y, type="s", main = "Hot and Cold Crisp Boundary", xlab = "Temperature in Degrees Farenheit", ylab="", xaxt="n", yaxt="n")  
axis(1, at=c(0,68,100), labels=c(0,68,100))  
axis(2, at=c(0,1), labels=c("Cold", "Hot"), las=1)  
x <- c(121.92,177.8,213.36)  
y <- c(0,1,1)  
plot(x,y, type="s", main = "Tall and Short Crisp Boundary", xlab = "Height in Feet and Inches", ylab="", xaxt="n", yaxt="n")  
axis(1, at=c(121.92,177.8,213.36), labels=c("4'","5' 10''", "7'"))  
axis(2, at=c(0,1), labels=c("Tall", "Short"), las=1)



Temperature and height are both interperable fields in that some people might interpret a given temperature or height differently. In other words, some people might classify a person as tall, and others might classify that same person as short. Likewise, some people might consider a given temperature as cold, and others might classify that same temperature as hot.

There are certain heights in which the vast majority of people would agree to be either "short" or "tall" (likewise for "short" and "tall"). Let's say that everyone agrees that people 5 feet and 2 inches and below in height are short, and people 5 feet and 10 inches and above in height are tall. Let's say that everyone agrees that any temperature less than or equal to 55F is cold, and any temperature greater than or equal to 68F. The area of disagreement, or the "grey area", is where the value of the sample is uncertain. That is, for samples within this "grey area", . A value of 0 or 1 means that everyone agrees on the attribute value. A value in between 0 and 1 indicates a degree of uncertainty. Such a distribution for both temperature and height with "grey areas" is shown below:

From here, you can make decision rules.

Here is an excellent video with greater detail on on fuzzy logic with many real world examples:

## Fuzzy Clustering vs Hard Clustering

Fuzzy clustering, or soft clustering, is where data points have the potential to belong to multiple clusters. Unlike clustering, or hard clustering, where points belong to exactly one cluster, points have the ability to be members of multiple clusters.

The membership between points and clusters is stored in a x matrix where is the degree to which belongs to cluster . Note that for every row entry in . This means that . That is, the degree to which points belong to all clusters sum to 1, and therefore the degree to which all points belong to all clusters is equal to the number of points. A sample membership matrix is shown below:

Here, point is only a member of cluster , point is equally a meber of both clusters and , and point is a stronger member of compared to .

In hard clustering, such a membership matrix might look like this:

Here, point is only a member of cluster , point is only a meber of cluster , and point is only a member of . Note in both variants that for every row entry in , , and therefore .

## Fuzzy C-Means Clustering (FCM) Algorithm

Fuzzy C-Means Clustering is very similar to the K-Means algorithm. It takes the following as inputs:

1. The number of clusters
2. The points to cluster

The basic outline of the Fuzzy C-Means Algorithm is as follows:

Here is a link to an [Interactive Demo](https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletFCM.html) for Fuzzy C-Means.

## Resources

<https://en.wikipedia.org/wiki/Fuzzy_clustering#Comparison_to_Hard_Clustering>

<https://www.youtube.com/watch?v=P8wY6mi1vV8>

<https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html>

<https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html>

<https://homes.di.unimi.it/valenti/SlideCorsi/Bioinformatica05/Fuzzy-Clustering-lecture-Babuska.pdf>