Project 4 Report

# Project Objective

Project 4 will apply 2 clustering techniques, K-means and EM, to segmenting image. Some more requirements with this project are:

* Use random initialization for K-means
* Use K-means as an initialization for EM
* Compare 2 algorithms on their performance and the sensitivity to the initialization
* Plot the objective function vs. the iteration number for both K-means and EM

In this project, Matlab is used as the programming language. Two moisac images are given as the input to segment. And two ground-truth maps are given as the foundation for evaluate the performance.

# Theoretical Background

## Clustering

Clustering is a field which the objective is to segment an image into different regions based on their features. For example,



To cluster or segment an image, the well-known methods usually use the image’s feature space as the input, where each point reflects the feature vector at a pixel. In this project, we introduce 2 common algorithms: K-means (hard clustering) and EM (soft clustering).

## K-means

### K-means formulation

 is a given set of unlabeled data samples in a d-dimensional space. A partition of D, denoted as, is a way to divide D into subsets such that. The problem of clustering is formulated as



In which, *f(.)* is the objective function (criterion, cost function,…)

K-means approach assumes that the number of classes, k, is given. The objective function in K-means is so-called intra-class divergence, which is:



Two basic steps in K-means are:

* Assume the cluster centers are known, and allocate each data point to the closest cluster center.
* Assume the allocation is known, and compute a new set of cluster centers. Each center is the mean of the points allocated to that cluster.

These steps are repeated many times until the intra-class divergence doesn’t change significantly or the centers don’t change.

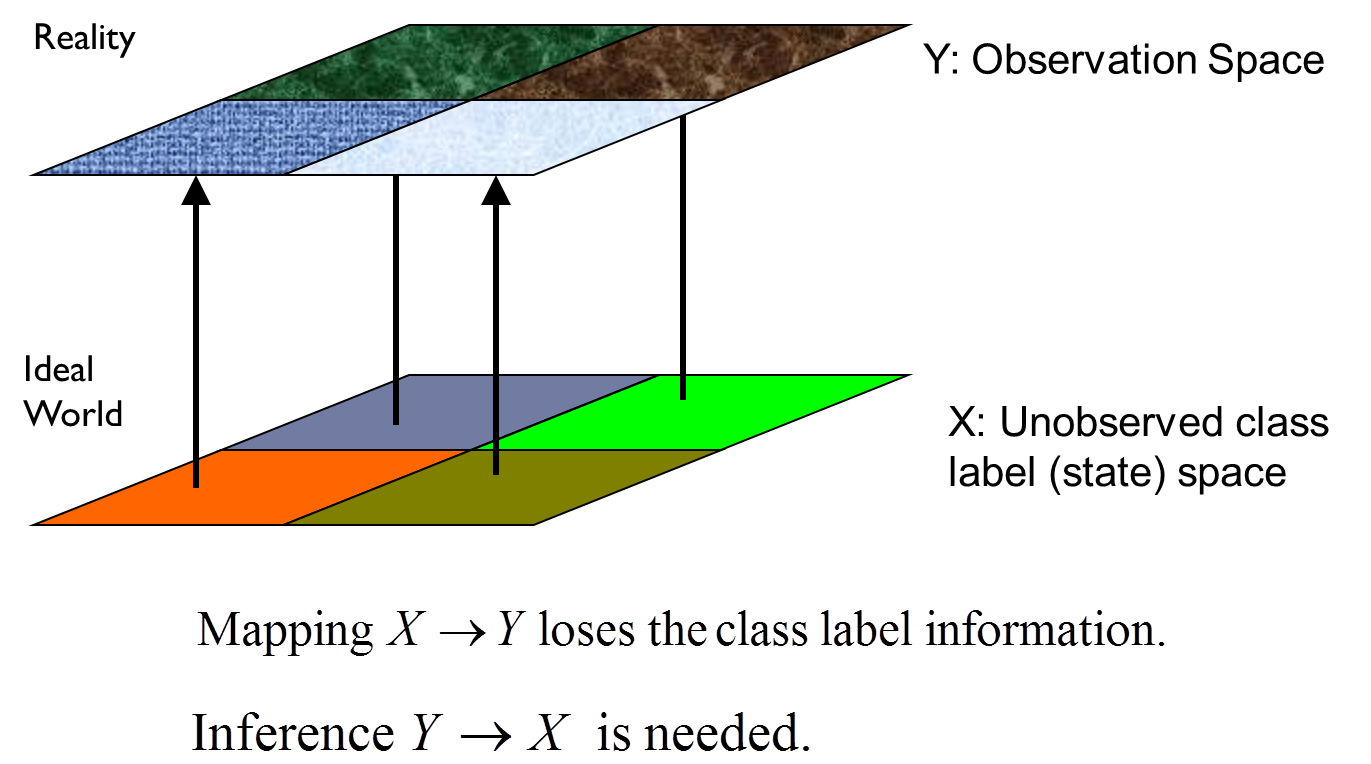
### K-means limitations

* K-mean depends strongly on the initialization. Therefore, it can be trapped into local minimum/maximum.
* K-means does NOT consider the spread of different clusters, structure of each cluster and proportion of different clusters.

## EM

### EM formulation

Before introducing EM, we restate the missing data problem briefly.



In this problem, X is the ideal world (ground-truth data) while Y is the reality world (observed data). The problem is that the class label information is not available in Y and that is the mission of EM algorithm.

Y: Observation Space

X: Unobserved class label (state) space



Reality

Ideal

World

EM solves the missing data problem by using two-steps iteration followed an initialization, as following:

* Initialization: initialize the parameters:



* Step 1 (E-Step): Estimate the missing data in terms of the posterior probability of each data sample:



* Step 2 (M-Step):

# Algorithm and Implementation

# Experimental Results

# Conclusion