Lab 08

-Means Clustering

2023-11-17

Table of contents

# 1. K-Means Clustering

-Means Clustering is a widely-used unsupervised machine learning algorithm, ideal for partitioning datasets into distinct, non-overlapping groups or ‘clusters’. We’ve seen it in the context of regional frequency analysis.

## 1.1 Algorithm

* **Inputs:** (number of clusters), $\vb{x} = \{x\_1, x\_2, x\_n\}$ (data points)
* **Outputs:** $\vb{c} = \{\mu\_1, \mu\_2, \ldots, \mu\_n\}$ (cluster assignments), $\vb{\mu} = \{\mu\_1, \mu\_2, \ldots, \mu\_k \}$ (cluster centroids)
* **Steps:**
  1. Randomly initialize cluster centers: $\vb{\mu} = \mu\_1^{(0)}, \mu\_2^{(0)}, \ldots, \mu\_k^{(0)} \in \mathbb{R}^d$
  2. Iterate until convergence:
     1. Assign each observation to the closest (in Euclidean distance) mean:
     2. Recompute each as the mean of all points assigned to it
     3. Terminate when the total change of the cluster centroids satisfies

# 2. Instructions

using CSV  
using DataFrames  
using Plots  
using StatsBase: mean, std

We will work in an external script. Open the file kmeans.jl and edit the functions provided. It’s a Julia file, so you can run line by line and work in the REPL.

To make the functions created available to you here, run the following command:

include("kmeans.jl")

check\_convergence (generic function with 1 method)

## 2.1 Initialize Centroids

First, edit the init\_centroids function. It takes in a matrix indexed by observations and features, and returns a matrix with rows (one for each centroid) and columns (one for each feature) where is the number of features of . The code provided initializes each centroid to a random value.

You can change this to whatever you like – be sure to explain your reasoning. One common approach is to choose random observations from the dataset as your initial centroids. Be sure to make sure that your centroids are distinct!

## 2.2 Euclidean Distance

In order to assign observations to clusters, we need to be able to compute the distance between between an observation and a centroid. We will use the Euclidean distance, which is defined above. This function should take in two generic vectors and return a scalar.

## 2.3 Assign Clusters

There is just one line of code to edit here.

The argmin function may be your friend.

## 2.4 Update Centroids

As you loop through the algorithm, you will need to update the centroids. This function takes in the data matrix , the cluster assignments $\vb{c}$, which is a vector of integers, and the number of clusters . It returns a matrix with rows (one for each centroid) and columns (one for each feature) where is the number of features of .

## 2.5 -means algorithm

This function is provided for you. You do not need to edit it. You simply need to define all the functions it calls.

function kmeans(X::AbstractMatrix, k::Int; τ=1e-5, maxiter=500)  
 n, d = size(X) # get the number of observations and features  
  
 # initialize the cluster centroids (μ)  
 μ = init\_centroids(X, k)  
 μ\_history = [μ]  
  
 is\_converged = false # initialize the flag  
 j = 1 # initialize the counter  
  
 # go through the loop until convergence is reached  
 while !is\_converged  
 cluster\_assignments = assign\_clusters(X, μ)  
 cluster\_centroids = update\_centroids(X, cluster\_assignments, k) # update the centroids  
  
 # add the current centroids to the history  
 push!(μ\_history, cluster\_centroids)  
  
 # check for convergence  
 is\_converged = check\_convergence(μ\_history, τ) # check for convergence  
  
 # if we have run too many iterations, throw an error (avoid infinite loops)  
 if j > maxiter  
 throw("Failed to converge after $j iterations")  
 end  
  
 # increase the counter  
 j += 1  
 end  
  
 cluster\_assignments = assign\_clusters(X, μ)  
  
 return cluster\_assignments, μ\_history  
end

# 3. Analysis

Our input data for this clustering analysis will be stations from the GHCND dataset (original [here](https://www.ncei.noaa.gov/data/global-historical-climatology-network-daily/doc/ghcnd-stations.txt)). We will subset only stations in Texas, and we will cluster on their longitude, latitude, and elevation. Importantly, we’ll need to *rescale our data* because latitude/longitude and elevation are on different scales.

# Define a function to parse each line  
function parse\_line(line)  
 station = strip(line[1:11]) # Station ID  
 latitude = parse(Float64, strip(line[13:20])) # Latitude  
 longitude = parse(Float64, strip(line[22:30])) # Longitude  
 elevation = parse(Float64, strip(line[32:37])) # Elevation  
 state = strip(line[39:40]) # State Abbreviation (if present)  
 name = strip(line[41:end]) # Station Name  
 return (station, latitude, longitude, elevation, state, name)  
end  
  
# Read the file and process each line  
function read\_file(filename)  
 data = []  
 open(filename) do file  
 for line in eachline(file)  
 push!(data, parse\_line(line))  
 end  
 end  
 return DataFrame(data, [:Station, :Latitude, :Longitude, :Elevation, :State, :Name])  
end  
  
# Usage  
filename = "data/ghcnd\_stations.txt"  
stations = read\_file(filename)  
stations = stations[stations[!, "State"] .== "TX", :]  
  
describe(stations)

Now we can run the clustering analysis we’ve implemented

X = Matrix(stations[!, [:Latitude, :Longitude]])  
X\_rescaled = (X .- mean(X; dims=1)) ./ std(X; dims=1)  
cluster\_assignments, μ\_history = kmeans(X\_rescaled, 10) # change the value of k here

([1, 1, 1, 1, 5, 3, 3, 3, 3, 3 … 4, 1, 5, 1, 1, 1, 3, 1, 1, 3], [[0.8890456786906505 0.5269196000275747; -2.3676715590951294 0.09465833236639393; … ; -0.10862184586031075 -0.0704742301983788; -1.4408045377461132 0.5071921260239539], [1.0038428982774308 0.4109569474950425; -2.2203654002913886 -0.007570875737819042; … ; -0.14147454684362115 -0.5464913616135716; -1.2945844485975537 0.46621359091801107], [1.0038428982774308 0.4109569474950425; -2.2203654002913886 -0.007570875737819042; … ; -0.14147454684362115 -0.5464913616135716; -1.2945844485975537 0.46621359091801107]])

# 4. Analysis

Once your code appears to be working:

1. Plot your cluster assignments on a map. Do they look logical? What does / does not make sense to you?
2. Check for consistency by re-running your code and comparing the plots. Do they look the same? Why or why not?
3. Plot the (two-dimensional) cluster centroids as a function of the number of iterations.
4. Try different values of . How does the clustering change? What is the best value of in your opinion? How could you determine this?

If you have extra time, alter other parts of the code (e.g., don’t use the elevation to cluster, or use a different distance metric). How does this change your results?