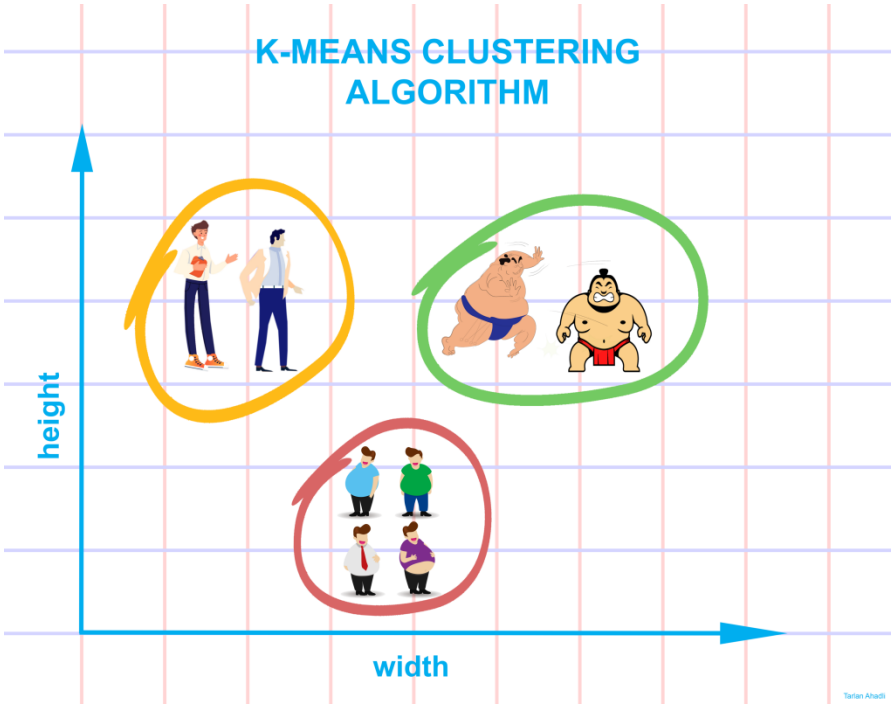
Introduction to K-Means Clustering Algorithm



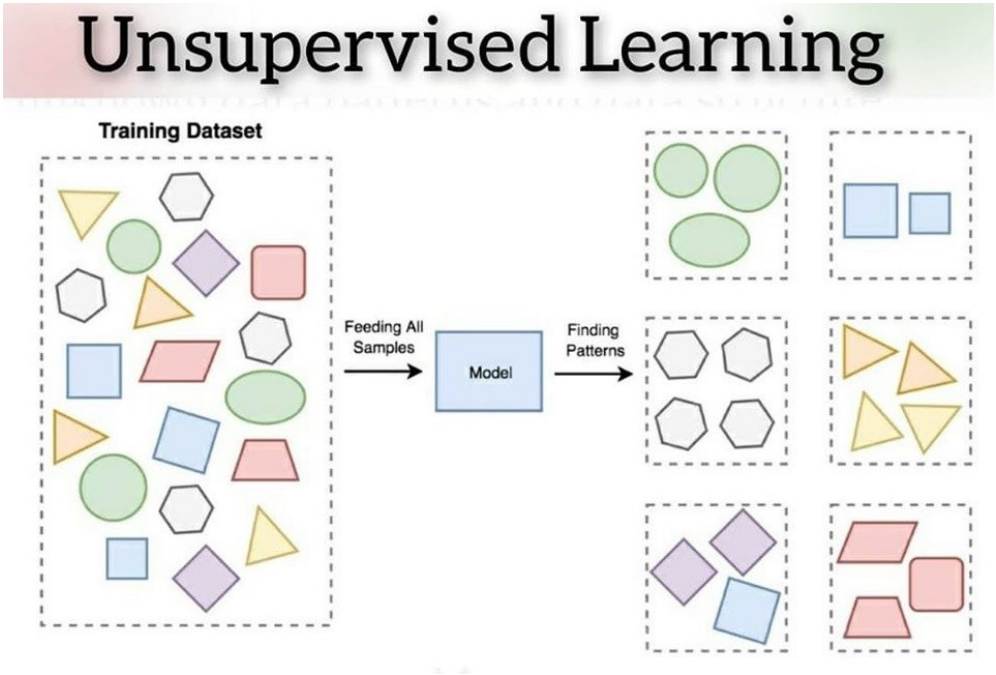
k Mean Clustering algorithm သည် ရိုးရှင်းပြီး ပေါ်ပြုလာဖြစ်သည့် unsupervised machine learning algorithm တစ်မျိုး ဖြစ်သည်။ Clustering လုပ်ခြင်းသည် unsupervised learning အမျိုးအစားတွင် ပါဝင်သည်။ K-Means clustering algorithm ကို အသုံးပြု၍ ဒေတာများကို အလိုအလျောက် အမျိုးအစားခွဲခြား (automatically organized the data) ပေးနိုင်သည်။ k Mean algorithm ကို သုံး၍ clustering လုပ်ခြင်းကြောင့် k Mean Clustering algorithm ဟုလည်းကောင်း partitioning algorithm ဟူ၍လည်း ခေါ်ဆို လေ့ရှိသည်။

k Mean Clustering algorithm ကို လေ့လာမည်သူ တစ်ယောက် အနေဖြင့် clustering အဓိပ္ပာယ်၊ k အဓိပ္ပာယ် နှင့် သက်ဆိုင်သည့် ဝေါဟာရများကို ရှင်းလင်းစွာ နားလည်သဘောပေါက်ထားသင့်သည်။

Clustering အဓိပ္ပာယ်

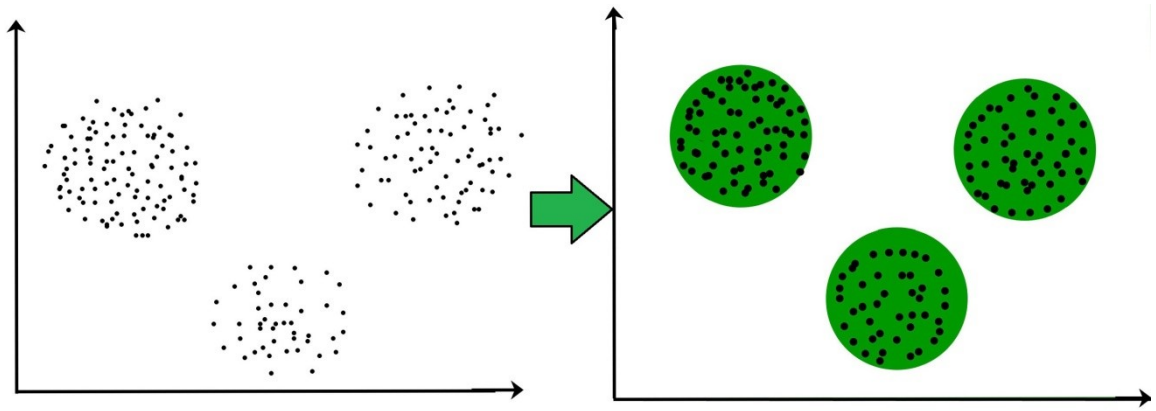
Clustering လုပ်ခြင်းဆိုသည်မှာ အမျိုးမျိုးသော အရာများ(object)ကို တူရာတူရာ စုဝေးစေခြင်း သို့မဟုတ် တူရာ(similarities) အရာများ(object)ကို တစ်စု တစ်ဝေးတည်း ဖြစ်အောင် အစုငယ်များ ဖွဲ့ပေးခြင်း ဖြစ်သည်။ (Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data. A cluster refers to a collection of data points aggregated together because of certain similarities.)

ဥပမာ- တေးဂီတကြိုက်သူများထဲမှ ဂန္ဓဝင်သီချင်း ကြိုက်သူများ အစုငယ်(cluster) ၊ ရော့သီချင်း ကြိုက်သူများ အစုငယ်(cluster) ၊ ပေါ့သီချင်း ကြိုက်သူများအစုငယ်(cluster) များဖွဲ့ခြင်းကို clustering လုပ်သည် ဟုခေါ်သည်။ စတိုးဆိုင်များတွင် တူရာ ပစ္စည်းများ တစ်နေရာတွင် တစ်စုတစ်ဝေး နေရာချခြင်းသည် clustering လုပ်ခြင်း ဖြစ်သည်။



ရုပ်ရှင်ကားများကို စစ်ကား၊ အက်ရှင်ကား၊ ဒရာမာကား စသည်ဖြင့် clustering လုပ်ထာခြင်းကြောင့် စစ်ကား ကြိုက်သူများကို စစ်ကားများ ချပေးနိုင်သည်။ ထိုအတူ ဒရာမာကား ကြိုက်သူများကို ဒရာမာကားများ ချပေးနိုင်သည်။

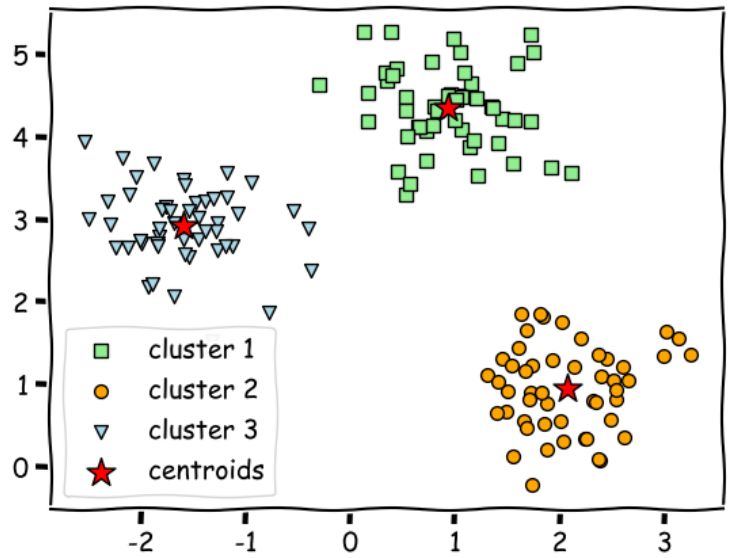
Clustering လုပ်ခြင်းသည် unsupervised learning အမျိုးအစားတွင် ပါဝင်သည်။ Input ဒေတာများတွင် label မပါသောကြောင့် unsupervised algorithm ကို သုံးရခြင်းဖြစ်သည်။ တစ်နည်းအားဖြင့် clustering မလုပ်ခင် မည်သည့်ဒေတာသည် မည်သည့်အမျိုအစားဖြစ်သည်ကို မသိသောကြောင့် unsupervised algorithm ကို သုံးရခြင်း ဖြစ်သည်။ Unsupervised algorithm တွင် training ဒေတာများသည် မည်သည့်အမျိုးအစားများ ဖြစ်သည်ကို မသိနိုင်သည့် ဒေတာများ ဖြစ်ကြသည်။



ဒေတာများတွင် independent variable များသာ ပါဝင်ပြီး target သို့မဟုတ် dependent variable များ မပါရှိလျှင် unsupervised algorithm ကို အသုံးပြုသည်။ Clustering လုပ်ရန်အတွက် ဒေတာများသည့်

independent variables များသာ ပါဝင်ပြီး target သို့မဟုတ် dependent variable များ မပါရှိသည့် ဒေတာများ ဖြစ်ကြသည်။

ဒေတာများ တစ်ခုနှင့်တစ်ခု တူညီသည် သို့မဟုတ် ကွဲပြားသည်ကို ဆုံးဖြတ်ရန် ထိုဒေတာများ၏ တန်ဖိုးများကို နှိုင်းယှဉ် ရသည်။ သို့မဟုတ် ထိုဒေတာများ၏ coordinate များကို နှိုင်းယှဉ် ရသည်။ Vector များကို နှိုင်းယှဉ် ရသည်။ သန်းပေါင်းသောတရားအရ အကွာအဝေးအနည်းဆုံးဖြစ်လျှင် တူညီသည်ဟု သတ်မှတ်သည်။



K-means clustering algorithm အသုံးပြုပုံ

စာများကို အမျိုးအစား ခွဲခြင်း(Document Clustering)

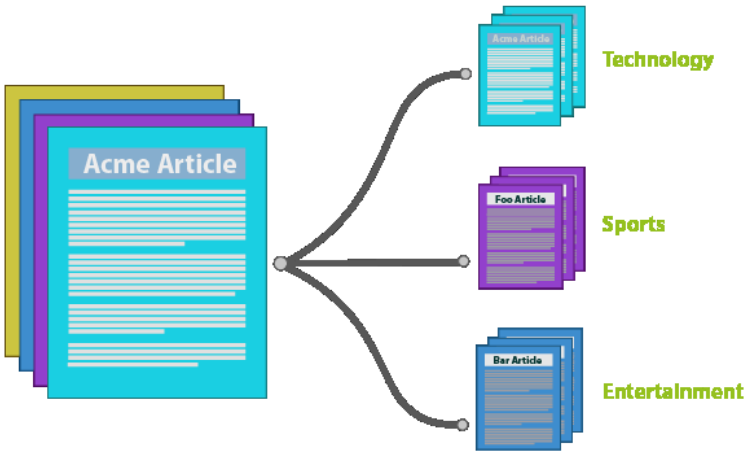
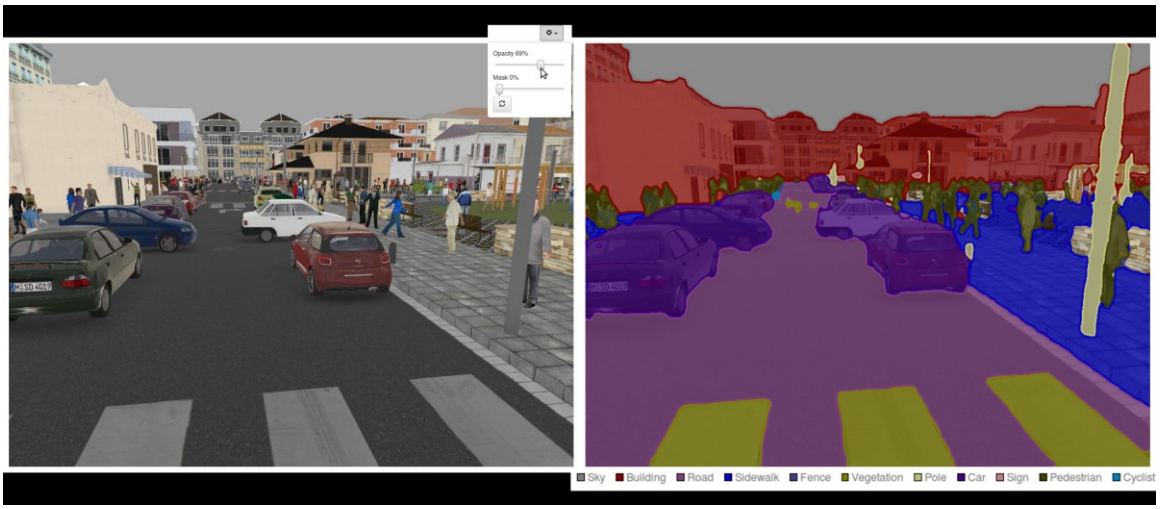


Image Segmentation

Image တစ်ခုအတွင် တူသည့်(similar) pixel များကို အစုလိုက်ခွဲခြားခြင်း သည် အသုံးများသည့် Image segmentation ဖြစ်သည်။



K-means clustering algorithm ကို စီပွားရေး၊ လူမှုရေးနယ်ပယ်များတွင် အသုံးပြုကြသည့် နေရာများမှာ

Behavioral segmentation:

- Segment by purchase history
- Segment by activities on application, website, or platform
- Define personas based on interests
- Create profiles based on activity monitoring

Inventory categorization:

- Group inventory by sales activity
- Group inventory by manufacturing metrics

Sorting sensor measurements:

- Detect activity types in motion sensors
- Group images
- Separate audio
- Identify groups in health monitoring

Detecting bots or anomalies:

- Separate valid activity groups from bots
- Group valid activity to clean up outlier detection

Properties of Clusters

(၁) Cluster တစ်ခုအဖြစ် သတ်မှတ်ရန်အတွက် cluster တစ်ခုအတွင်းရှိ ဒေတာများအားလုံးသည် အချင်းချင်း တစ်ခုနှင့်တစ်ခု တူညီကြရမည်။ (All the data points in a cluster should be similar to each other.)

(၂) Cluster မတူညီသည့် ဒေတာများသည် တစ်ခုနှင့် တစ်ခု အတတ်နိုင်ဆုံးလုံးဝ ကွဲပြားခြားနား နေရမည်။(The data points from different clusters should be as different as possible.)

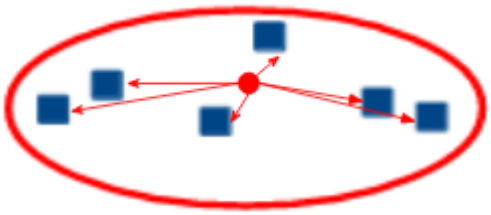
တကယ်လက်တွေ့အခြေအနေတွင် ဒေတာများသည် အလွန်ရှုပ်ထွေးသောကြောင့် ဒေတာများကို clustering လုပ်ပြီး အစုငယ်များကွဲပြားသွားခြင်းသည် အဓိကအချက်မဟုတ်ပေ။ ထိုကွဲပြားသွားသည့် အစုငယ်များသည် အဓိပ္ပာယ်ရှိပြီး အသုံးချရန် နားလည်နိုင်သည့် အစုငယ်များ ဖြစ်သင့်သည်။ (The primary aim of clustering is not just to make clusters, but to make good and meaningful ones.)

Inertia ဆိုတာဘာလဲ

Inertia ဆိုသည်မှာ centroid နှင့် အစုငယ်(cluster)အကြား အကွာအဝေး အားလုံးကို စုပေါင်းထားသည့် အကွာအဝေး ဖြစ်သည်။ (inertia is the sum of distances of all the points within a cluster from the centroid of that cluster.)။ ထို့ကြောင့် အစုငယ်(cluster)တစ်ခုချင်းစီအတွက် Inertia တန်ဖိုး ရှိသည်။

အစုငယ်(cluster)တစ်ခုချင်း၏ centroid ၏ နေရာ(location)သည် ပြောင်းလဲ နေသောကြောင့် Inertia တန်ဖိုးလည်း ပြောင်းလဲနေသည်။ ထို့ကြောင့် Centroid နေရာ(location)ကို နောက်ဆုံးအဆင့် သတ်မှတ်ပြီးမှ inertia တန်ဖိုးကို တွက်ရသည်။

အစုငယ်(cluster)အတွင်းရှိ centroid နှင့် အစုငယ်(cluster)အကြား အကွာအဝေးကို intracluster distance ဟု ခေါ်ဆိုသည်။ Inertia တန်ဖိုးသည် intracluster distance များ အားလုံးစုစုပေါင်း အကွာအဝေး ဖြစ်သည်။ (inertia gives us the sum of intracluster distances)



Intra cluster distance

Inertia value နည်းလေလေ တည်ဆောက်လိုက်သည့် cluster များ ပို၍ စုစည်း ကျစ်လစ်လေ ဖြစ်သည်။ (the lesser the inertia value, the better our clusters are.)။ စုစည်း ကျစ်လစ်သည့်(compact) cluster ဖြစ်ဘို့ intracluster distance တွေ အနည်းဆုံး(တစ်ခုနှင့် တစ်ခု အနီးစပ်ဆုံး) ဖြစ်ဘို့ လိုအပ်သည်။

Cluster ၏ centroid နှင့် အဖွဲ့ ဝင်ဒေတာများ အကြား အကွာအဝေး(distance) နည်းလျှင် cluster သည် စုစည်း ကျစ်လစ်သည်။ တစ်နည်းအားဖြင့် ထို cluster များအတွင်းရှိ ဒေတာများ(အဖွဲ့ဝင်များ)သည် တစ်ခုနှင့် တစ်ခု နီးကပ်စွာ တည်ရှိကြသည်။ (if the distance between the centroid of a cluster and the points in that cluster is small, it means that the points are closer to each other.)

အထက်တွင် ဖော်ပြခဲ့သည့် Cluster တစ်ခုအဖြစ် သတ်မှတ်ရန်အတွက် cluster တစ်ခုအတွင်းရှိ ဒေတာများအားလုံးသည် အချင်းချင်း တစ်ခုနှင့်တစ်ခု တူညီကြရမည်ဆိုသည့် အချက်ကို Inertia တန်ဖိုးက ဆုံးဖြတ်ပေးသည်။ Inertia တန်ဖိုးသည် Cluster တစ်ခု အတွင်းအတွက်သာ ဖြစ်သည်။ တခြား Cluster များအတွက် မသက်ဆိုင်သည့် အရာသာ ဖြစ်သည်။

Dunn Index

Cluster တစ်ခု အတွင်း ဒေတာများကောင်းစွာ စုဝေးနေခြင်းသည် intracluster distance နည်းခြင်း ဖြစ်သည်။ Cluster များ တစ်ခုနှင့် သီးခြားကင်းလွတ်စွာ ရှိနေခြင်းသည် inter-cluster distance များခြင်း ဖြစ်သည်။

Dunn index သည် cluster တစ်ခု အတွင်း ဒေတာများကောင်းစွာ စုဝေးနေခြင်းအချက် အပြင် တခြားသော cluster များတည်ရှိရာနေရာကိုပါ ထည့်သွင်းတွက်ချက်ထားခြင်း ဖြစ်သည်။

မတူညီသည့် cluster နှစ်ခု၏ centroid အကြား အကွာဝေးကို inter-cluster distance ဟုခေါ်သည်။ (The distance between the centroids of two different clusters is known as inter-cluster distance.)

Formula of the Dunn index:

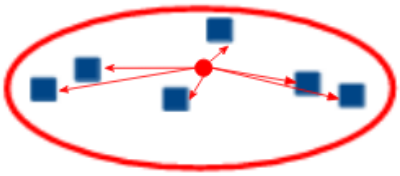
$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

Dunn index ဆိုသည်မှာ minimum of inter-cluster distances နှင့် maximum of intracluster distances တို့၏ အချိုး ဖြစ်သည်။ (Dunn index is the ratio of the minimum of inter-cluster distances and Maximum of intracluster distances.)

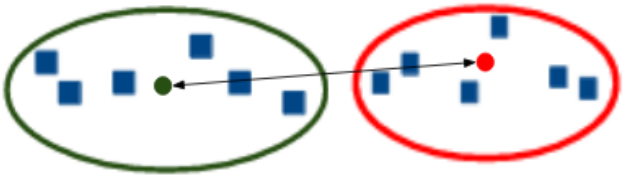
Dunn index များလေ cluster များကို ကောင်းစွာ ခွဲခြားနိုင်လေဖြစ်သည်။ ကောင်းစွာ ကောင်းစွာ ခွဲခြားနိုင်သည်ဟုဆိုရာတွင် cluster တစ်ခု အတွင် ဒေတာများကောင်းစွာ စုဝေးနေပြီး cluster များ တစ်ခုနှင့် သီးခြားကင်းလွတ်စွာ ရှိနေခြင်းကို ဆိုလိုသည်။

Minimum of the inter-cluster distances

Dunn index တန်ဖိုးများရန်(to maximize the Dunn index)အတွက် ပိုင်းဝေတွင် ရှိရမည့် တန်ဖိုးများနိုင်သမျှ များရမည်။ (In order to maximize the value of the Dunn index, the numerator should be maximum.)



Intra cluster distance



Inter cluster distance

Maximum of intra cluster distance

Dunn index တန်ဖိုးများရန်(to maximize the Dunn index)အတွက် ပိုင်းခြေတွင် ရှိရမည့် တန်ဖိုး နည်းနိုင်သမျှ နည်းရမည်။ (the denominator should be minimum to maximize the Dunn index)

Cluster centroid များ အကြား အကွာဝေးအများဆုံးနှင့် cluster အတွင်း centroid နှင့် point များ အကြား အကွာအဝေးအနည်းဆုံး ဖြစ်လျှင် cluster များ သီးခြားကင်းလွတ်ပြီး ကျစ်လစ်နေလိမ့်မည်။ (The maximum distance between the cluster centroids and the points should be minimum which will eventually make sure that the clusters are compact.)

K-means algorithm အလုပ် လုပ်ပုံ

k-means clustering technique ၏ အဓိက အလုပ်မှာ cluster တစ်ခုအတွင်းရှိ point များ နှင့် centroid အကြား အကွာဝေးကို အနည်းဆုံး နေရာသို့ ရောက်အောင် centroid တည်နေရာကို ရွှေ့ပေးရန်ဖြစ်သည်။ (The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.)

K-means algorithm ကို centroid-based algorithm သို့မဟုတ် distance-based algorithm ဟူ၍လည်းခေါ်ဆိုကြသည်။ K-means algorithm အလုပ် လုပ်ပုံကို အောက်တွင် ပုံများဖြင့် ဥပမာတစ်ခု ဖော်ပြထားသည်။

Point (၈)ခု ပါရှိသည့် ဒေတာများကို cluster (၂)ခု ခွဲသည့် ဥပမာတစ်ခုကို စတင်လေ့လာကြရအောင်

ပထမအဆင့်: cluster အရေအတွက်ကို ရွေးချယ်ခြင်း(Choose the number of clusters k)

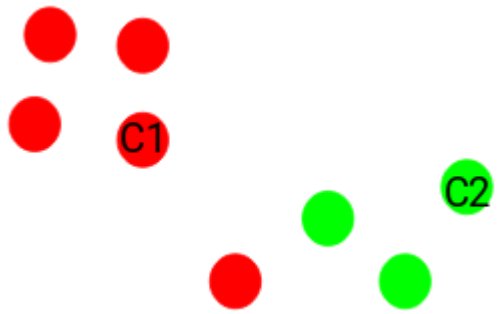
ပထမ cluster အရေအတွက်(number of clusters, k)ကို သတ်မှတ်ရပါမည်။ Cluster အရေအတွက် (number of clusters, k)မည်ကဲ့သို့ ရွေးချယ်ရမည်ကို နောက်ပိုင်းတွင်ဆက်လက် လေ့လာပါမည်။ ယခု ဥပမာအတွက် k = 2 ဖြစ်သည်။ k = 2 ဆိုသည်မှာ cluster (၂)ခု ဖြစ်အောင် ပိုင်းခြားပါမည်။ ထို့ကြောင့် centroid နှစ်ခု ရှိရပါမည်။

ဒုတိယအဆင့်: centroid နေရာသတ်မှတ်ပေးခြင်း

Centroid နှစ်ခု တည်နေရာကို အဆင်ပြေသည့် နေရာတွင် သတ်မှတ်ပေးလိုက်ပါမည်။ (randomly select the centroid for each cluster.)။ အောက်ပုံတွင် ပြထားသည့်အတိုင်း centroid နှစ်ခု တည်နေရာ C1 နှင့် C2 ကို အဆင်ပြေသည့် နေရာတွင် ချထားလိုက်ပါသည်။

တတိယအဆင့်: cluster ၏ centroid နီးသည် point များကို ထို cluster တွင် ပါဝင်သည်ဟု သတ်မှတ်ခြင်း(Assign all the points to the closest cluster centroid)

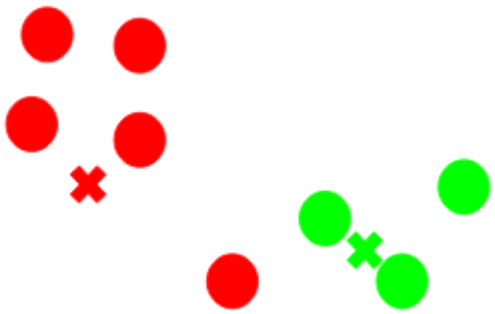
Centroid ကို နေရာချပြီးသည့်အခါ(initialized the centroids) cluster ၏ centroid နီးသည် point များကို ထို cluster တွင် ပါဝင်သည်ဟု သတ်မှတ်(Assign all the points to the closest cluster centroid)လိုက်ပါသည်။



အောက်ပုံတွင် C1 သည် အနီရောင် cluster centroid ၏ centroid ဖြစ်သောကြောင့် ထို C1 နှင့် နီးသည့် point များသည် အနီရောင် cluster တွင် ပါဝင်ကြသည်။ ထိုအတူ C2 သည် အစိမ်းရောင် cluster centroid ၏ centroid ဖြစ်သောကြောင့် ထို C2 နှင့် နီးသည့် point များသည် အစိမ်းရောင် cluster တွင် ပါဝင်ကြသည်။

စတုတ္ထအဆင့်: ဖြစ်ပေါ်လာသည့် cluster ကို အခြေခံ၍ centroid နေရာအသစ်ကို ပြန်တွက်ခြင်း

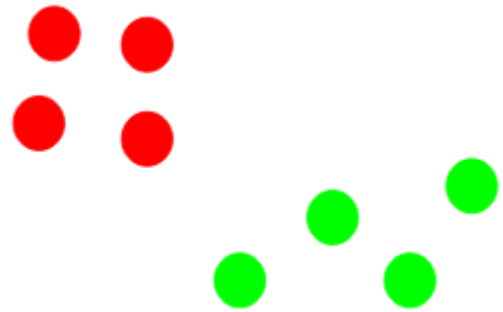
ဖြစ်ပေါ်လာသည့် cluster (newly formed clusters)ကို အခြေခံ၍ cluster တစ်ခုအတွင်းရှိ point များနှင့် centroid အကြား အကွာဝေးကို အနည်းဆုံး နေရာသို့ ရောက်အောင် centroid တည်နေရာကို ရွှေ့ပေးရန် ဖြစ်သည်။ ထို့ကြောင့် centroid ရှိရမည့် တည်နေရာအသစ်ကို ထပ်မံတွက်ရမည်။ ထို့သို့ centroid ရှိရမည့် တည်နေရာအသစ်ကို တွက်ပေးပြီး ရွှေ့သည့်အခါတိုင်းကို iteration တစ်ကြိမ် ပြုလုပ်သည်ဟု သတ်မှတ်သည်။



အစိမ်းရောင် ကြက်ခြေခတ်နှင့် အနီရောင် ကြက်ခြေခတ်တို့သည် အသစ်ဖြစ်ပေါ်လာသည့် centroid များ ဖြစ်ကြသည်။

ပဉ္စမအဆင့် : တတိယအဆင့် နှင့် စတုတ္ထအဆင့်တို့ကို ထပ်မံပြုလုပ်ခြင်း(Repeat steps 3 and 4)

တတိယအဆင့် နှင့် စတုတ္ထအဆင့်တို့ကို ထပ်မံပြုလုပ်ခြင်း ဖြင့် သီးခြားကွဲပြားနေသည့် အနီရောင် အစု နှင့် အစိမ်းရောင်အစု ဖြစ်ပေါ်လာသည်။ ထိုကဲ့သို့ clustering လုပ်ပြီးနောက် centroid များ မလိုအပ်တော့ပါ။



k အဓိပ္ပာယ်

k သည် clustering လုပ်လိုသည့် အရေအတွက် ဖြစ်သည်။ ရုပ်ရှင်ကားများကို အမျိုးအစား မည်မျှ ခွဲခြားမည်ဟုသတ်မှတ်ခြင်းသည် k အရေအတွက်ကို ဆုံးဖြတ်ခြင်း ဖြစ်သည်။ k အရေအတွက်ကို (၅)(k = 5)ဟု ဆုံးဖြတ်လိုက်လျှင် ပေးထားသည့် ဒေတာများကို အစု (၅)စု ခွဲမည်ဟု သတ်မှတ်လိုက်ခြင်း ဖြစ်သည်။

ဥပမာ-သင် စက်ဖြင့် အဝတ်လျှော်အခါ လျှော်ရမည်အဝတ်များကို ကြည့်၍ အဖြူရောင် အဝတ်များ၊ အရောင်မကျွတ်သည့် အဝတ်များအစုနှင့် အရောင်ကျွတ်နိုင်သည့် ကျန်အဝတ်များအစုအဖြစ် နှစ်စု(k=2) ခွဲခြား လိုက်ခြင်းသည် clustering လုပ်လိုက်ခြင်း ဖြစ်သည်။

k အရေအတွက်ကို ဆုံးဖြတ်ခြင်း

k အရေအတွက်ကို ဆုံးဖြတ်ခြင်းသည် ရိုးရှင်းလွယ်ကူသည့် ကိစ္စ တစ်ခု မဟုတ်ပါ။ အောက်တွင် ဖော်ပြ ထားသည့် ပုံတွင် k အရေအတွက် မည်မျှဖြစ်မည်ကို အလွယ်တကူ ဆုံးဖြတ်ရန် ခက်ခဲသည်။

Centroid ဆိုသည်မှာ အစု(cluster)များ၏ အလယ်ဗဟိုနေရာ ဖြစ်သည်။ လက်ဆုတ်လက်ကိုင် ပြနိုင်သည့်နေရာ ဖြစ်နိုင်သလို imaginary location လည်း ဖြစ်နိုင်သည်။ (A centroid is the imaginary or real location representing the center of the cluster.)

Centroid အရေအတွက်(number of centroids)နှင့် K အရေအတွက် တူညီသည်။ အစုဝင်များသည် centroid ဗဟိုပြု၍ centroid နှင့် နီးနိုင်သမျှ အနီးဆုံးနေရာတွင် တည်ရှိနေအောင် အစု(cluster)များကို ဖွဲ့ရမည်။ *'means'* အဓိပ္ပာယ်မှာ cluster အတွင်းရှိ ဒေတာများနှင့် centroid အကြားရှိ အကွာအဝေးကို ပျှမ်းမျှယူတားသည့် တန်ဖိုးဖြစ်သည်။

K အရေအတွက် ဘယ်လို ရွေးချယ်မလဲ(How to choose K?)

K Mean ကိုသုံးသည့်အခါ K အရေအတွက် (K number)ကို ဆုံးဖြတ်ပေးရခြင်းသည် အားနည်းချက် ဖြစ်သည်။ ထိုအချက်ကို ဖြေရှင်းရန်အတွက် K အရေအတွက်အမျိုးမျိုး(different numbers of centroids.)ဖြင့် algorithm ၏ performance ကို တွက်ကြည့်ရသည်။

K အရေအတွက် အမျိုးမျိုးဖြင့် K-means clustering algorithm ဖြင့် စမ်းသပ်ပြီး ရသည်အဖြေကို နှိုင်းယှဉ် ကြည့်ရသည်။ ယေဘူယျအားဖြင့် အကောင်းဆုံးအဖြေကို ပေးနိုင်သည့် K အရေအတွက် အတိအကျကို သိနိုင်သည့်နည်း မရှိပေ။ (In general, there is no method for determining exact value of K,)

သို့သော် တိကျသည့် K အရေအတွက်ကို သိနိုင်ရန် အောက်ပါနည်းများကို အသုံးပြုနိုင်သည်။

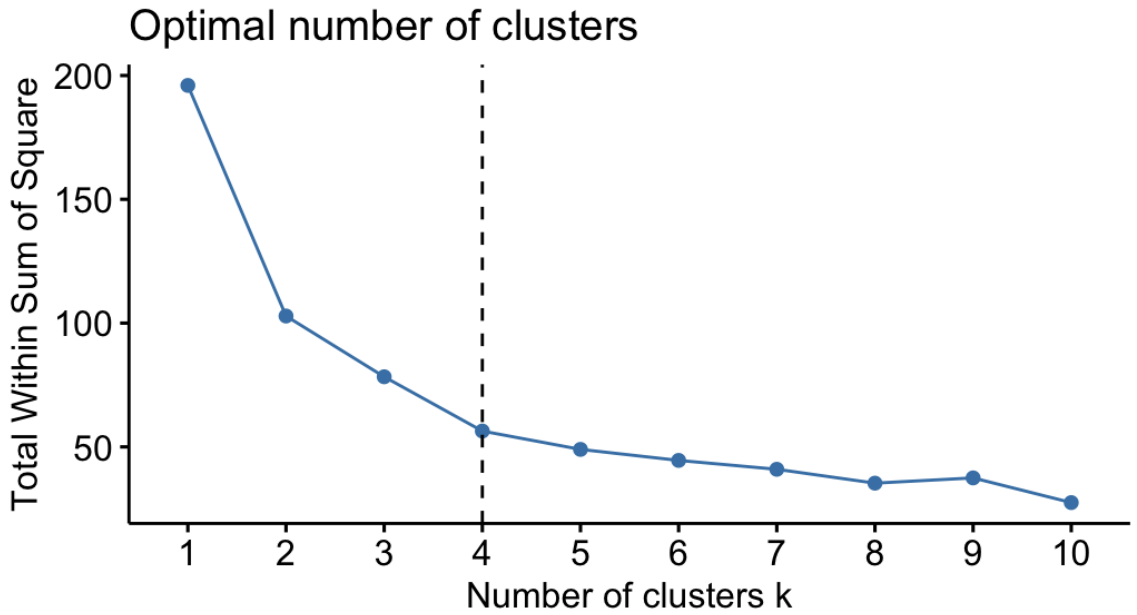
K အရေအတွက် အမျိုးမျိုးဖြင့် data point များ နှင့် သက်ဆိုင်ရာ cluster centroid တို့၏အကြား ပျမ်းမျှ တန်ဖိုး(the mean distance between data points and their cluster centroid)များကို တွက်၍ နှိုင်းယှဉ် ခြင်းဖြင့် အကောင်းဆုံး K အရေအတွက် ကို သိနိုင်သည်။

K အရေအတွက်ကို များအောင် တိုးပေးခြင်းဖြင့် data points များ အကြား အကွာအဝေး နည်းနိုင်သော်လည်း K အရေအတွက်ကို များသောကြောင့် မလိုအပ်ပဲအစုများစွာ ဖြစ်ပေါ်လာလိမ့်မည်။ အစွန်းဆုံး အခြေအနေတစ်ခုကို စဉ်းစားကြည့်ရအောင်။ K အရေအတွက်ကို data point အရေအတွက် နှင့် တူညီအောင် ထားလိုက်လျှင် data point တစ်ခုစီတိုင်းသည် centroid ဖြစ်ပြီး data point တစ်ခုစီတိုင်းသည် cluster ဖြစ်သွားလိမ့်မည်။

Cluster centroid အရေအတွက် ပိုများလာလေ objective function ၏ magnitude ပိုနည်းလာလေဖြစ်သည်။ (As the number of cluster centroids increases, the magnitude of the objective function will be less.)

အောက်ပုံတွင် number of clusters K နှင့် value of the objective function ကို ဖော်ပြထားသည်။ အကောင်းဆုံး K အရေအတွက်ကို ရွေးရန်(to choose the optimal K) elbow method ကို အသုံးပြုနိုင်သည်။ elbow method ဆိုသည်မှာ အောက်ပုံတွင် ဂရပ်လိုင်းသည် လက်မောင်း နှင့် တံတောင်ဆစ်ကဲ့သို့ ကွေးဆင်း သွားသကဲ့သို့ ဖြစ်နေသည်။ တံတောင်ဆစ် ကွေးနေရာသည် အကောင်းဆုံး K အရေအတွက်ကို ဖော်ပြပေးသည့် နေရာ ဖြစ်သည်ဟု ဆိုလိုသည်။ ထိုနေရာကို "elbow point" ဟုခေါ်ဆိုသည်။

"Elbow point," မတိုင်ခင်နေရာသည် K အရေအတွက်သည်နှင့်အမျှ claster ဖွဲ့စည်းမှု ပိုကျစ်လစ်လာသည် ။ သို့သော် "elbow point," ကျော်သွားသည့်အခါ K အရေအတွက် ပိုများအောင် လုပ်သော်လည်း claster ဖွဲ့စည်းမှု မဆိုစလောက်သာ ကောင်းလာသောကြောင့် ဖြစ်သည်။



K-Means algorithm ၏ အားသာချက်များ

- တခြားသော algorithm များနှင့် နှိုင်းယှဉ်လျှင် ရိုးရှင်းလွယ်ကူသည်။(Relatively simple to implement.)
- အလွန်ကြီးမားသည့် data set များအတွက် scaling လုပ်နိုင်သောကြောင့် အဆင်ပြေသည်။(Scales to large data sets.)
- Convergence ဖြစ်ရန် သေချာသည်။ (Guarantees convergence.)
- Centroid တည်နေရာကို အဆင်ပြေသည့်နေရာမှ စတင်နိုင်သည်။(Can warm-start the positions of centroids.)
- မြန်သည်။ Efficient ဖြစ်သည်။ (The algorithm is fast and efficient in terms of computational cost.)

K-Means algorithm ၏ အားနည်းချက်များ

- Algorithms မစခင် k (number of clusters)ကို ကြိုတင်သတ်မှတ်ပေးရသည်။
- Algorithms ၏ ရလဒ်သည် cluster centers(centroid)တည်နေရာ စတင် ချထားသည့် နေရာအပေါ်တွင် မူတည်သည်။
- cluster အပြင်ဘက် အဝေးနေရာတွင် ရောက်နေသော များကြောင့် ရလဒ် မကောင်းခြင်း ဖြစ်နိုင်သည်။ (It's sensitive to outliers. Outliers must be removed before clustering, or they may affect the position of the centroid or make a new cluster of their own.)
- အရွယ်အစား အမျိုးမျိုနှင့် သိပ်သည်းခြင်း ကွဲပြားသည့် ဒေတာများ အတွက် သိပ်ကောင်းသည့်နည်း မဟုတ်ပေ။ ထိုကဲ့သိုသော ဒေတာများနှင့် ကြုံတွေ့ပါက generalize လုပ်ရန် လိုသည်။ (Clustering data of varying sizes and density. K-means doesn't perform well with clusters of different sizes, shapes, and density. To cluster such data, you need to generalize k-means.)
- convex မဖြစ်သည့် cluster များအတွက် မသင့်လျော်ပါ။(It's not suitable for finding non-convex clusters)
- local minimum တစ်နေရာရာတွင်သာ ရပ်တန့်နေပြီး global optimum ဖြစ်ရန်အတွက် အာမခံချက်မပေးနိုင်ပါ။ (It's not guaranteed to find a global optimum, so it can get stuck in a local minimum)

Reference

https://medium.com/@tarlanahad/a-friendly-introduction-to-k-means-clustering-algorithm-b31ff7df7ef1

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantage