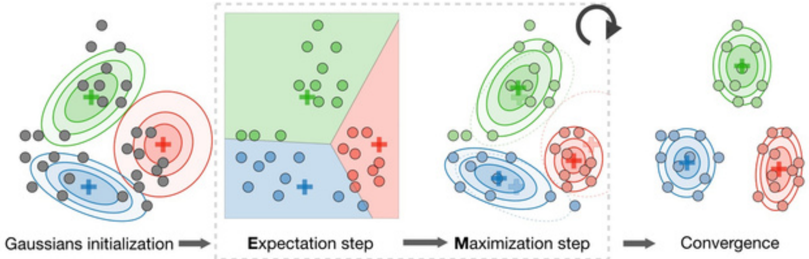
VIP Cheatsheet: Unsupervised Learning



Introduction to Unsupervised Learning

k-means clustering

r **Motivation** – The goal of unsupervised learning is to find hidden patterns in unlabeled data

{x(1),...,x(m)}. Wenotec(i)theclusterofdatapointiandμjthecenterofcluster

r **Jensen's inequality** – Let f be a convex function and X a random variable. We have the r **Algorithm** – After randomly initializing the cluster centroids

1

followinginequality: algorithmrepeatsthefollowingstepuntilconvergence:

E[f(X)]>f(E[X]) m

1

Expectation-Maximizationc(i)=argmin||x(i)−μ||2and==1jμj

j

1

r **Latent variables** – Latent variables are hidden/unobserved variables that make estimation

i=1

problems difficult, and are often denoted z. Here are the most common settings where there are latent variables:

Setting Latentvariablez x

|z **Comments**

MixtureofkGaussians Multinomial(φ)

N(μ,Σ)μ∈Rnjj,φ∈Rkj

Factor analysis

N(0,I) N(μ+Λz,ψ) μ∈Rnj

r **Algorithm** – The Expectation-Maximization (EM) algorithm gives an efficient method at estimating the parameter θ through maximum likelihood estimation by repeatedly constructing a lower-bound on the likelihood (E-step) and optimizing that lower bound (M-step) as follows:

•E-step:EvaluatetheposteriorprobabilityQ(z(i))thateachdatapointx(i)icamefrom a particular cluster z(i) as follows:

Q (z(i)) = P(z(i)

|x(i)i ;θ)

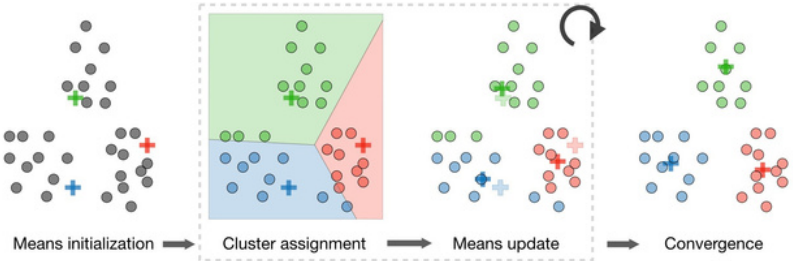
•M-step:UsetheposteriorprobabilitiesQ(z(i)i)asclusterspecificweightsondatapoints x(i) to separately re-estimate each cluster model as follows:

∑()(i)(i)=argmax((i))logP(x,z;θ)θQzdz(i)ii(i)Qz(i)θz^i()i

j.

μ ,μ2,...,μk ∈Rⁿ,the k -means

$$\frac{\sum c(\{\}i=j\ x^{(i)}}{\sum m\,c(i)\}\ \{\ =j}$$



r **Distortion function** – In order to see if the algorithm converges, we look at the distortion function defined as follows:

∑mJ(c,μ)=||x(i)−μ||2c(i)i=1

Hierarchical clustering

r **Algorithm** – It is a clustering algorithm with an agglomerative hierarchical approach that build nested clusters in a successive manner.

r **Types** – There are different sorts of hierarchical clustering algorithms that aims at optimizing different objective functions, which is summed up in the table below:

Ward linkage	Averagelinkage	Complete linkage
Minimize within cluster distance	Minimize average distance between cluster pairs	Minimize maximum distance of between cluster pairs

Clustering assessment metrics

In an unsupervised learning setting, it is often hard to assess the performance of a model since we don’t have the ground truth labels as was the case in the supervised learning setting.

r **Silhouette coefficient** – By noting a and b the mean distance between a sample and all other points in the same class, and between a sample and all other points in the next nearest cluster, the silhouette coefficient s for a single sample is defined as follows:

=b−as

max (a,b)

r **Calinski-Harabaz index** – By noting k the number of clusters, B k and Wk the between and within-clustering dispersion matrices respectively defined as

∑k∑mB=n(μ−μ)(μ−μ)T,W=(x(i)μ)(x(i)μ)Tkc(i)c(i)c(i)k−c(i)

−c(i)j=1i=1

the Calinski-Harabaz index s(k) indicates how well a clustering model defines its clusters, such that the higher the score, the more dense and well separated the clusters are. It is defined as follows:

()=Tr(Bk)N−ksk

×Tr(Wk)k

−1

Principal component analysis

It is a dimension reduction technique that finds the variance maximizing directions onto which to project the data.

r **Eigenvalue, eigenvector** – Given a matrix A ∈ Rn×n, λ is said to be an eigenvalue of A if there exists a vector z ∈ Rn\{0}, called eigenvector, such that we have:

Az = λz

r **Spectral theorem** – Let A ∈ Rn×n. If A is symmetric, then A is diagonalizable by a real orthogonal matrix U ∈ Rn×n. By noting Λ = diag(λ1,...,λn), we have:

∃Λ diagonal, A = UΛUT

Remark: the eigenvector associated with the largest eigenvalue is called principal eigenvector of matrix A.

r **Algorithm** – The Principal Component Analysis (PCA) procedure is a dimension reduction technique that projects the data on k dimensions by maximizing the variance of the data as follows:

• Step 1: Normalize the data to have a mean of 0 and standard deviation of 1.

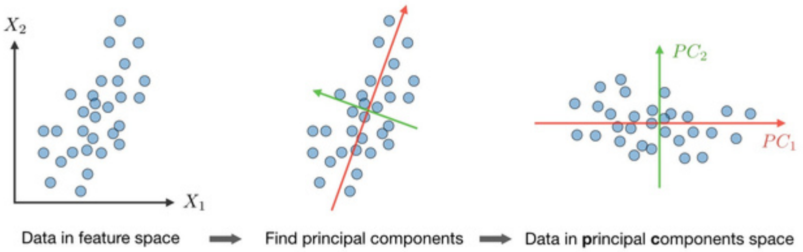
(i)

∑m∑m(x−i)jμjxj←where=1(i)and2=1((i)μ)2jxjσjxj−μjσjmmi=1i=1

∑m•Step2:ComputeΣ=1(i)(i)Txx∈Rn×n,whichissymmetricwithrealeigenvalues.mi=1

•Step3:Computeu1,...,u Rnk ∈ thekorthogonalprincipaleigenvectorsofΣ,i.e.the orthogonal eigenvectors of the k largest eigenvalues.

• Step 4: Project the data on span R(u1,...,uk). This procedure maximizes the variance among all k-dimensional spaces.



Independent component analysis

It is a technique meant to find the underlying generating sources.

r **Assumptions** – We assume that our data x has been generated by the n-dimensional source vector s = (s1,...,sn), where si are independent random variables, via a mixing and non-singular matrix A as follows:

x=As

The goal is to find the unmixing matrix W = A−1 by an update rule.

r **Bell and Sejnowski ICA algorithm** – This algorithm finds the unmixing matrix W by following the steps below:

• Write the probability of x = As = W−1s as:

n

p(x)= p(wTs ∏

ix)·

|W|

i=1

• Write the log likelihood given our training data {x(i),i ∈ [[1,m]]} and by noting g the sigmoid function as: ∑(∑(

m n

))l(W)=logg'(wTx(i)j)+log|W|i=1j=1

Therefore, the stochastic gradient ascent learning rule is such that for each training example x(i), we update W as follows:

1

□ −2g(wTx(i)1) □ □

1

−2g(wT(i))W

←−+2xWα.x(i)T+(WT)−1. □□□□□ □ □

.

1

−2g(wTx(i)n)□ □