Hands-on Machine learning

Chapter 9 Unsupervised learning techniques (summary)

At the beginning we knew how important is unsupervised learning is and how big it is compared to supervised and reinforcement learning, as if unsupervised learning is like a cake then supervised learning is just the icing of the cake and reinforcement learning will be like the cherry on the top of the cake.

Then we knew…

What is Clustering , Anomaly detection and Density estimation.

Clustering	Grouping of similar instances together in one cluster
Anomaly detection (Outliers detection)	Learning to differentiate between normal data and how it looks like and abnormal data
Density estimation	Estimating probability density function (PDF) of random process which is useful in anomaly detection as instances that are in very low density regions are more likely be anomalies.

Is **Clustering** useful ? what is the applications of **Clustering**?
Clustering is used in many things like customer segmentation , data analysis , dimensionality reduction , feature engineering , anomaly detection , images segmentation , search engines

K-means Algorithm

K-means is finding each blob center and assign each instance to the closest blob.

K-means checks the distance between instance and every centroid.

The metrics used for the **k-means** algorithm is **model's inertia** which is the sum of squares of distance between instances and centroids.

it is better to use kmeans++ than random initialization because in random initialization every centroid may start next to each other, but with kmeans++ its guaranteed that each centroid will be initialized far away from other centroids and that decrease the probability of getting sub-optimal solutions.

Algorithm	Accelerated K-means	Mini-batch k-means
Advantages	Avoid unnecessary distance calculations by exploiting triangle of inequality	Instead of taking whole dataset once, it takes mini-batches which make it faster and better if the whole data can't load in memory
Disadvantages	May slowdown it significantly as it depends on the dataset	Generally lower inertia

Another question too,
is how to find the optimal number of clusters in k-means algorithm?

We don't just decrease the number of inertia carelessly,
Because as we increase the number of clusters inertia will decrease as the centroids will be chopped into multiple pieces which doesn't add any value to the model and only take more time to compute,
Instead, We will use elbow method or silhouette score.

Elbow method	Silhouette score
We choose the value of clusters (k) the same as the value of (k) in the arms.	It is the mean silhouette coefficient and can range between (-1 , 1)
Any smaller value the inertia will be huge	1 or ~ 1 means that the instance is great in this cluster
Any bigger value means that the center will be chopped into pieces and making no difference in our clustering	0 or ~ 0 means that the instance is close to another cluster
	(-1) or ~ (-1) means that the instance is maybe assigned to the wrong cluster

It is very useful too to analyze with silhouette diagrams as it takes number of instances in consideration, not only scores.
So, we can balance the number of instances with the number of clusters to make each cluster contains about the same length of instances.

K-means is useful in images segmentation , semi-supervised learning (partially propagation)

K-means limits:
- That we need to run it several times to obtain an optimal solution as centroids starts with random values.

- It behaves bad versus varying sizes and non spherical shapes clusters.

-**Gaussian mixture of models** perform better than k-means in elliptical clusters.

Active learning is an uncertainly sampling model training and lowest estimated probability are given to experts for labeling.

DBScan is good for arbitrary shaped clusters its approach based on local density estimation.
There is HDBScan too which is Hierarchical shaped.

Other clustering algorithms:

Agglomerative Clustering	**It is called Hierarchical Clustering too.**
BIRCH Clustering	**Good for huge datasets and faster than Mini-batch K-means.**
Mean-shift Clustering	**Bad for big datasets and good for circular shaped clusters.**
Affinity Clustering	**Bad for big datasets, Use the voting method to select the representator like K-means.**
Spectral Clustering	**Can capture complex structures clusters , Use similarity matrix between instances , bad if every cluster size differs a lot from other clusters.**

Gaussian mixtures of models algorithm(GMM):
We start by having a dataset (**X**),
And we want to get weights (**Φ**)
and a
ll distribution parameters:
- Distribution **means** ($\mu^{(1)}$ - $\mu^{(k)}$)
- Distribution **covariance** ($\varepsilon^{(1)}$ - $\varepsilon^{(k)}$)
it uses **Expectation-Maximization** algorithm

Expectation	Maximization
It is assigning instances to clusters.	It is updating clusters.

To know how many clusters in **Gaussian mixtures of models** algorithm we use Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC)

Then we knew the difference between **probability** and **likelihood** in statistics

probability Density Function (PDF)	Likelihood
It is a function of x (with θ fixed)	It is a function of θ (with x fixed)
Given a statistical model with some parameters θ	
It is used to describe how plausible a future outcome x is (knowing the parameter values θ)	It is used to describe how plausible a particular set of parameter values θ are, after the outcome x is known