Lecture 9: *"Unsupervised Learning Techniques"*

Book Chapter

Please read the *Chapter 9* **"Unsupervised Learning Techniques"** till page 258 ('Other Clustering Algorithm') and answer the following questions.

Keep in mind: If you answer these questions and write a detailed summary, you won't need to read these chapters again while preparing for the exam

Video Nugget

There is a nice video on K-Means clustering and also on DBSCAN.

Questions

Clustering

1. What is the major benefit of *unsupervised learning*?
2. Describe the tasks *Clustering*, *Anomaly Detection* and *Density Estimation* in one sentence.
3. In which applications is *clustering* used?

K-Means

4. How can K-Means be used for dimensionality reduction?
5. Explain the difference between hard and soft clustering.
6. Describe the K-Means algorithm in your own words.
7. What can happen if we get bad random initial centroids?
8. How can we initialize the centroids better?
9. What is the idea behind K-Means++?
10. How did accelerated K-Means and mini-batch K-Means improve training time?
11. How can we measure the performance of K-Means?
12. Why is minimizing inertia a bad metric if we try to get the best number of clusters?
13. How can we find a good guess for k?
14. How does the *silhouette score* work?
15. How can we interpret the *silhouette diagrams* on page 248?
16. What are the limitations of K-Means?

Applications of K-Means

17. What is the idea behind clustering for image segmentation?
18. How can clustering used for semi-supervised Learning?
19. What is the idea behind label propagation?
20. What does *Active Learning* mean?

DBSCAN

21. Which instances are taken together in clusters using the DBSCAN algorithm?
22. What is a core instance in DBSCAN?
23. With which parameters is the density in DBSCAN defined?
24. List the advantages and disadvantages of the DBSCAN algorithm.

Homework Assignment

Please work on the exercises given in Unsupervised_Learning_task.ipynb

Answers

1. No need for labels
2. Clustering - find groups with similar instances, anomaly detection - what is "normal"-> find outliers or a sequence of outliers, density estimation -> find the probability density function for a random process where data is given (this is used to find outlier, outliers are the regions with low density)
3. customer segmentation, data analysis, dimensionality reduction (if you have k clusters just the distances (=affinities) to the clusters are stored, these are the new dimensions), anomaly detection, semi-supervised learning, search engines for similar images, image segmentation
4. The new dimensions are the distances to each center of each cluster
5. *hard* - find single cluster, *soft* score per cluster Score can be: distance, similarity (Gaussian RBF)
6. choose k, find k-center points (randomly), for each instance evaluate the distance to each cluster center and put the instance to the cluster with the nearest cluster center, compute new cluster centers for each cluster, for each evaluate again the distances to the new centers and put it to the nearest cluster, do this iteration until there are no changes in the clusters
7. we get suboptimal solutions or it takes many iterations
8. Give rough init points run the algorithm multiple times with different inits and keep best solution or kmeans++: choose the centroids with a wide distance from each other, but also with a kind of randomness

9. Find new centroids far away from actual centroids
10. Accelerated K-Means: using triangle inequality, mini-batch K-Means: using mini-batches
11. *inertia* = mean squared distance between each instance and closest centroid
12. Inertia always decreases w/ higher number of k
13. using elbow criteria: Do for each k a clustering and plot the inertia in dependence to k. In the elbow point is the best choice for k or with silhouette coefficient (later in chapter)
14. ...
15. best clusters for k=4, negative values in plots for k=3 and k=5, example for a bad cluster: second for k=6, because it is sharp and has low values
16. need to specify the number of clusters, does not behave well for different densities of the clusters or nonspherical shapes
17. Find objects by same color Find representative images by clustering, train the classifier with them. Next step: Label also the instances near to the representative with the same label, then apply classification for the labeled data set
18. label one instance of a cluster and propagate the label to other instances
19. present an instance to a human and let them label it. the instance is chosen by maximum knowledge gain or other metrics.
20. The instances which build a high density, that means they are lying close together.
21. A core instance has at least min_points samples in its neighborhood.
22. With epsilon and min_samples: epsilon is the radius of a circle around an instances and min_samples gives the number of instances which should be inside this circle to be "dense".
23. pro: simple, finds clusters of any shape, robust to outliers, finds outliers con: no centroid, can't find clusters of different density, O(m^2) for large eps