Unsupervised Machine Learning with Python

Section 11.1: Concluding Remarks and Thank You

Unsupervised Machine Learning

- Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a dataset with no pre-existing labels and with a minimum of human supervision.

Algorithms Covered in this Course

Course covered two broad categories of Unsupervised ML Algorithms:

Clustering:

- Hierarchical, DBSCAN, K Means, Gaussian Mixture Model

Dimension Reduction:

- Principal Component Analysis

Summary of Algorithms Presented

Algorithm	Advantages	Disadvantages
Hierarchical	-Identifies clusters at all levels	-Not feasible for large number of data points M, as number of operations is $O(M^3)\ as\ M \to \infty$
DBSCAN	-Can identify clusters of arbitrary shape -Elbow approach applied to nearest neighbours to choose parameter ε	-Does not do well for clusters of varying density -Not unique -May be slow for large number of data points as number of operations can be as large as $O(M^2)\ as\ M \to \infty$
K Means	-Number of operations is $O(M)\ as\ M \to \infty$ -Elbow method to pick number of clusters -Can be used for problems with 1000s of features	-Not unique, as clusters depend on choice of initial means -Does not do well with non-convex or elongated clusters
Gaussian MM	-Number of operations is $O(M)\ as\ M \to \infty$ -Can handle elongated clusters -Elbow method to pick number of clusters -Can use variants if unique covariance matrices not suitable	-Not unique, as clusters depend on choice of initial means, covariances, weights -Does not do well with non-convex clusters -Difficulties for large number of dimensions d, as work scales like d^3, or if determinants of Covariance matrices close to 0, as there may numerical overflow/underflow issues
PCA	-Straightforward approach making use of SVD for dimension reduction -Can significantly improve timings of clustering algorithms, without sacrificing performance	-May be issues performing SVD with large number of data points or large number of dimensions

Software for Unsupervised Learning

Title	Notes (See UnsupervisedML_Resources.pdf Chapter 11) for links
scikit-learn	Machine Learning framework for Python for supervised and unsupervised learning Codes available for Hierarchical Clustering (called Agglomerative Clustering), DBSCAN, K Means, Gaussian Mixture Model, and PCA
fastcluster	Python codes for hierarchical clustering
kneed	Python code for find the “elbow” in the elbow method. This code refers to the “elbow” point as the “knee”
gmr	Python codes for Gaussian Mixture Model
klusterpy	Python codes for K Means, K Mediods, Hierarchical Clustering, DBSCAN

Recommendations:

(1) Try K Means first

- Can handle large number of dimensions and data points
- Reasonably fast

(2) Perform dimension reduction

- Can speed up calculations without sacrificing quality of results

(3) Use a software package

- These packages have been tested and optimized

Thank You

- Congratulations for making to the end of the course!
- Thank you for taking this course
- I hope that it has been a worthwhile experience and that it has increased your interest in machine learning
- Best wishes for your future learning and endeavours!