Data Science Toolbox Question Sheet

03.2 Clustering

Daniel Lawson

Block 3

Short questions

1. Explain what is mean by **algorithmic**, **distance-based**, and **model-based** clustering, and why they might each be preferred.
2. What is meant by **hierarchical clustering**? Define divisive clustering. Define agglomerative clustering.
3. What does the distance metric $d(\vec{x}, \vec{y}) = [(\vec{x} - \vec{y})C^{-1}(\vec{x} - \vec{y})^T]^{1/2}$ do? What is C? Through specification of C, describe it in relation to the standard Euclidean distance.
4. A measure of distance d satisfies symmetry, non-negativity, is zero if the elements are the same, and satisfies the triangle inequality. Is it a metric?
5. Instead it satisfies $d(x, z) \leq max(d(x, y), d(y, z)$. What type of metric is it, metric, divergence or ultrametric?
6. A clustering procedure iteratively merges clusters a,b based on the minimum inter-cluster distances: $d_{a,b} = \min_{i \in a, j \in b} d_{i,j}$. What type of clustering is this?
7. Give a high-level explanation of the DBSCAN algorithm. How is able to approximate local density? How can it be used to perform outlier detection?
8. Give a high level explanation of K-means clustering. In what sense is it random?
9. What assumptions does K-means make about the clusters? What alternative approach might make fewer assumptions?
10. Given the K-means algorithm, describe its computational complexity in terms of a fixed number of iterations. How might the number of datapoints affect convergence?
11. What issues are there in using BIC to choose between models in Gaussian Mixture modelling? How do these differ from the more general model choice problem?