1 Introduction

Week 7 introduces 2 unsupervised learning methods for clustering of data. The first method being explored in K-means which aims to cluster data into K groups by minimizing a criterion known as inertia. K is a parameter that needs to be chosen as a parameter before execution by the user. Along side that, we will also explore Gaussian mixture models (GMMs) which are a generalisation of K-means to incorporate covariance information Pedregosa et al. [2011]. This model uses a combination of Gaussian distributions to model the data.

2 K-Means Clustering

- Why is it called K-Means?
  In K-Means the term K refers to the number of clusters that need to be identified; and, means refers to the process of averaging of data to find the centroid of each cluster.

- Monothetic and Polythetic Clustering: In a monothetic scheme, cluster membership is based on the presence or absence of a single characteristic. Polythetic schemes use more than one characteristic. For example, classifying people solely on the basis of their gender is a monothetic classification, but if both gender and handedness (left or right handed) are used, the classification is polythetic.

- To read about hard and soft clustering, please refer to this article.

- The objective of K-means as defined in Bishop [2006] Section 9.1 is the minimisation of the cost function J where $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} ||x_n - \mu_k||^2$ such that, $r_{n,k}$ denotes if point $n$ belongs to cluster $k$ and $||x_n - \mu_k||^2$ is the squared error.

- To understand the K-means algorithm, please refer to Wu et al. [2008] Section 2.1. The basic steps can be elucidated as:
1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids or maximum iterations has been reached.

- An improvement on the basic K-means algorithm is to introduce a kernel on top of the data to project it into a high-dimensional space Dhillon et al. [2004]. Although the boundaries will be linear in the high-dimensional space, on projecting back to the lower dimensions, it becomes non-linear.
- To read about the limitations of K-means, please refer to Wu et al. [2008] Section 2.2.
- To get a quick overview of the K-means algorithm, please refer to Barber [2012] Section 20.3.5. [Requires an understanding of Expectation Maximization]

3 Gaussian Mixture Models

- This topic requires an intuition about Maximum Likelihood Estimation. To get a quick refresher, please refer to this article.
- What is Expectation-Maximization?
  Expectation maximization is an iterative process of improving the probability of a model to predict if an observation belongs to a specific distribution in the presence of latent variables.
  - E-Step ⇒ Estimate the missing variables in the dataset
  - M-Step ⇒ Maximize the parameters of the model in the presence of the data

Maximum Likelihood estimate the same probability in the absence of latent variables.
- This can be used good starter video to understand the intuition about Expectation-Maximization (EM).
- To get a deeper understanding of the mathematics behind the general EM algorithm, please refer to Bishop [2006] Section 9.4. Another approach to EM, based on mathematical derivations, is provided in Section 2 of this document.
- Basic Representation of Mixture Models is provided in Figure 1
- An intuitive concept of Gaussian Mixture model is provided in this article
Section 2 and 3 from this document provides an elaborate explanation of Gaussian Mixture models and Expectation Maximization.

A thorough and clear explanation of Gaussian Mixture Models (albeit, slightly lengthy) is also provided in Bishop [2006] Section 9.2.

4 Comparison between K-means and GMM

<table>
<thead>
<tr>
<th>Criterion</th>
<th>K-Means</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Convergence</strong></td>
<td>Faster than GMM</td>
<td>Slower than K-Means</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>Computationally less intensive</td>
<td>Computationally intensive</td>
</tr>
<tr>
<td><strong>Initialization</strong></td>
<td>Random Initialisation</td>
<td>Use K-means to determine the means of the Gaussian</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Single hard assignment to clusters</td>
<td>Probability distribution over the cluster assignment</td>
</tr>
</tbody>
</table>

Table 1: This table provides a comparative analysis of K-Means clustering and Gaussian Mixture Models over 4 criteria.
References


