Machine Learning in Communications
Lecture 5b: Introduction to Unsupervised Learning

Harpreet S. Dhillon

Wireless@VT, Bradley Department of Electrical & Computer Engineering
Virginia Tech, Blacksburg, VA

https://www.dhillon.ece.vt.edu
hdhillon@vt.edu

JTG/IEEE Information Theory Society Summer School
IIT Kanpur

Lecture Objectives

- Introduce the basics of unsupervised learning required for our case study in Lecture 6.
- We will focus on the k-means algorithm and its interpretation.
- Time permitting, we will also introduce the Gaussian mixture model.
- Density estimation and other related topics will be covered in Lecture 7.

Unsupervised Learning

- Unlike supervised learning, we are now given an "unlabelled dataset" $\mathbf{X}$ with no corresponding supervising outputs or labels.
- The learning problem is not as "concretely" defined as the supervised learning case.
- Examples of unsupervised learning:
  - Clustering
  - Dimensionality reduction
  - Density estimation (useful for outlier detection)
- Reference: Kevin Murphy's MLPP.

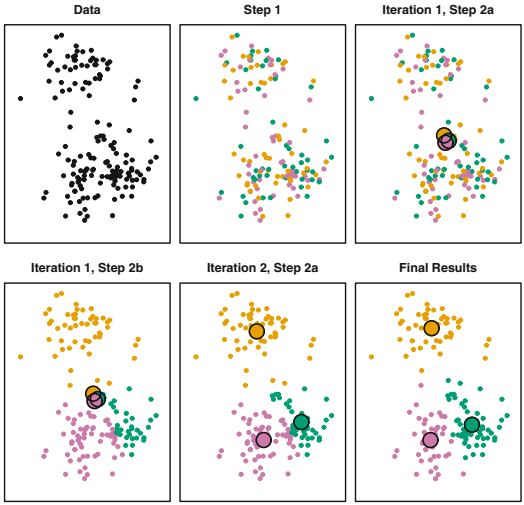
k-means Clustering: Introduction

- As the name suggests, the general idea is to "cluster" similar samples.
- A measure of distance (or similarity) is required.
  - Easy to visualize in the Euclidean domain. We will use squared distance for our discussion but other distance metrics can of course be used as well.
  - May need a little more care in non-Euclidean spaces (Lecture 6).
- We will assume that k (number of clusters) is given.
- Initially, we will consider "hard" assignment, i.e., each point is assigned to one cluster. We will generalize this when we introduce mixture models.

k-means Clustering: Procedure

- Procedure:
  - For a given k, initialize the "centers" of k clusters.
  - Then assign each point of the dataset to its closest cluster center.
  - Define the new center of a given cluster as the centroid of the points attached to it.
  - Repeat the procedure until there is no change in the assignment.
- As we will see shortly, k-means can be interpreted as an approximate solution to a meaningful objective function.

k-means Clustering: Example



[ISL Fig. 10.6] Example of k-means clustering.

k-means Clustering: Parameters of Interest

- Given: k
- Dataset: $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$
- Parameters to be estimated:
  - Cluster centers: $\{\boldsymbol{\mu}_j\}_{i=1}^{k}$, where each $\boldsymbol{\mu}_j \in \mathbb{R}^d$.
  - Assignment of points to clusters: $\{c_i\}_{i=1}^{n}$, where each $c_i \in \{1, 2, \ldots k\}$.
  - Equivalent assignment parameter: $\{a_{ij}\}$, where $1 \leq i \leq n$ and $1 \leq j \leq k$. Here $a_{ij} = 1$ if i^{th} point is assigned to the j^{th} cluster center and 0 otherwise.
- Assignment step: $c_i = \arg\min_j \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2$.
- Recentering step: $\boldsymbol{\mu}_j = \frac{\sum_i a_{ij} \boldsymbol{x}_i}{\sum_i a_{ij}}$

k-means Clustering: Interpretation

- It can be argued that k means clustering algorithm (approximately) minimizes the below objective function:

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i}\|^2.$$

- It is useful to think of this objective function in terms of data compression or quantization.
- For instance, we may have applications where $\boldsymbol{x}_i$ may be represented by its "quantized" value $\boldsymbol{\mu}_{c_i}$. This will be the case in Lecture 6.
  - Therefore, k-means (approximately) minimizes average distortion introduced by the quantization process.

k-means Clustering: Interpretation

- Let's think in terms of coordinate gradient descent (will provide approximate solution for our objective function).
- Fix $\boldsymbol{\mu}_j$, and optimize $\{c_i\}$ or equivalently $\{a_{ij}\}$:

$$\min_{\{c_i\}} \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i}\|^2 = \frac{1}{n} \sum_{i=1}^{n} \min_{c_i} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i}\|^2$$

  This is just the assignment step.
- Now fix $\{a_{ij}\}$ and optimize $\boldsymbol{\mu}_j$:

$$\min_{\{\boldsymbol{\mu}_j\}} \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{1}{n} a_{ij} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2 = \sum_{j=1}^{k} \min_{\boldsymbol{\mu}_j} \sum_{i=1}^{n} \frac{1}{n} a_{ij} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2$$

  This is just the centroid (or recentering) step.

Shortcomings of k-means

- In order to understand the shortcomings of k-means (or the implicit assumptions made about the clusters), it will be useful to interpret it in terms of a specific Gaussian mixture model (GMM), which is our next topic.
- This will naturally lead us to soft assignments, density estimation, and expectation maximization.
- We will just introduce the GMM today and do the rest in Lecture 7.

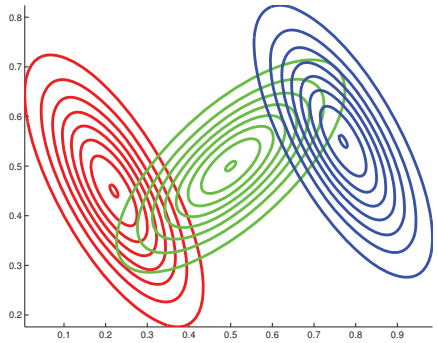
Gaussian Mixture Model

- Let's introduce a latent variable z_i for every data sample. It gives us the probability with which a given point was sampled from a specific Gaussian.
  - $z_i \in \{1, 2, \ldots, k\}$ with $p(z_i = j) = \pi_j$.
- Likelihood: $p(\boldsymbol{x}_i | z_i = j) = p_j(\boldsymbol{x}_i)$, where $p_j = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is the distribution of the j^{th} Gaussian.
- Mixture: It is called a mixture model because:

$$p(\boldsymbol{x}_i | \boldsymbol{\theta}) = \sum_{j=1}^{k} p(\boldsymbol{x}_i, z_j | \boldsymbol{\theta}) = \sum_{j=1}^{k} p(\boldsymbol{x}_i, | \boldsymbol{\theta}, z_j) \pi_j$$
$$= \sum_{j=1}^{k} p_j(\boldsymbol{x}_i | \boldsymbol{\theta}) \pi_j$$

  - Mixture of k base Gaussians.

Gaussian Mixture Model: Illustration



[MLPP Figure 11.3] A mixture of three Gaussians in 2D.
Keep in mind that this is our "model" and is not how the data was actually generated.

Clustering using Mixture Models

General idea: Fit the mixture model to the data and compute the posterior probability that the point belongs to cluster j (also called *responsibility*):

$$r_{ij} = p(z_i = j|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{p(z_i = j|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = j, \boldsymbol{\theta})}{\sum_{j=1}^{k} p(z_i = j|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = j, \boldsymbol{\theta})}$$

- This is soft clustering.
- Hard clustering can be done by:

$$z_i^* = \arg\max_j r_{ij}.$$

Summary

- In this short lecture, we introduced the idea of unsupervised learning.
- We looked at the k-means and one of its interpretations.
- We also introduced GMM. We will build on this discussion in Lecture 7.