# Lecture 12 Notes

CS-419M
Spring 2021

## 1 Unsupervised Learning

Previous lectures were dealing with supervised learning , where you are given inputs $\mathbf{X}$ and desired outputs $\mathbf{y}$ and the goal is to approximate a good mapping function from $\mathbf{X}$ to $\mathbf{y}$ . But in unsupervised learning the expected output or labels are not provided .The goal is generally to find some hidden patterns in the data. More formally given N observations $(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$ having joint density $Pr(X)$ , the goal is to directly infer the properties of this probability density without the help any supervision ( True labels ).

### 1.1 Clustering

Clustering is the problem of identifying groups, or clusters of data points in a multidimensional space.

**Problem Setting** :
Given a set of features $x_i$ , where $i \in \{1, 2, \ldots, N\}$ , how to cluster these points ?

Before approaching the problem we will require two things . First , a notion of distance ,which we will represent here by euclidean distance . Second , the number of clusters K , which we will assume to be given with the problem .Given these our goal is to assign each data-point to a cluster $c \in \{1, 2, \ldots, K\}$.

Let's define $p_{ic}$, $i \in \{1, 2, \ldots, N\}$ and $c \in \{1, 2, \ldots, K\}$ the probability that data-point $x_i$ is assigned to cluster c. So $\sum_{c \in \{1,2..,K\}} p_{ic} = 1$. Also let $\mu_c$ denote the centroid of cluster c. i.e mean of the data-points in cluster c.

Given these we can pose the above problem as an optimization problem .Precisely ,

$$\min_{\mu_c, p_{ic}} \sum_{i=1}^{N} \sum_{c=1}^{K} p_{ic} \|x_i - \mu_c\|^2$$

such that

$$\sum_{c \in \{1,2..,K\}} p_{ic} = 1$$

This optimization problem can be solved through a iterative procedure , where each iteration involves two successive steps corresponding to successive optimizations with respect to the $p_{ic}$ and the $\mu_c$. First we chose some initial value for $\mu_c$ and minimize with respect to $p_{ic}$ keeping $\mu_c$ fixed. Then fixing the $p_{ic}$ values we minimize with respect to $\mu_c$.This steps are continued until no significant changes happen to $\mu_c$ or for some predefined iterations .

**First step** (minimizing with respect to $p_{ic}$ keeping $\mu_c$ fixed) :

$$\min_{p_{ic}} \sum_{i=1}^{N} \sum_{c=1}^{K} p_{ic} \|x_i - \mu_c\|^2 \text{ subject to } \sum_{c \in \{1,2..,K\}} p_{ic} = 1$$
$$= \sum_{i=1}^{N} \min_{p_{ic}} \sum_{c=1}^{K} p_{ic} \|x_i - \mu_c\|^2 \text{ subject to } \sum_{c \in \{1,2..,K\}} p_{ic} = 1$$

Solving this will give us

$$p_{ic} = \begin{cases} 1 & \text{if } c = argmin_j \|x_i - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

Essentially this is just same as assigning each data-point to the cluster which has it's current centroid closest to the data-point.One should note that even though we had assumed $p_{ic}$ to be a probabilistic variable taking values from $[0, 1]$ it will actually be assigned a value from $\{0, 1\}$.

**Second step** (minimizing with respect to $\mu_c$ keeping $p_{ic}$ fixed) :

$$\min_{\mu_c} \sum_{i=1}^{N} \sum_{c=1}^{K} p_{ic} \|x_i - \mu_c\|^2$$

Since $p_{ic}$ are constant , The function is just a quadratic function of $\mu_c$. So we can actually solve this in the usual way of taking first derivative and then assigning the critical point to zero.After solving we will get

$$\mu_c = \frac{\sum_{i=1}^{N} p_{ic} x_i}{\sum_{i=1}^{N} p_{ic}}$$

Intuitively it is just taking the mean of the data-points assigned to each cluster. That's why also it is called K-means clustering .

Reference : 'Pattern recognition and machine learning' by Christopher M. Bishop