CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 15: Clustering

Lecturer: Abir De *Scribe: Group 2*

15.1 Introduction to Clustering

1. Clustering is an unsupervised classification technique which involves grouping of values in a given data set close to the representative value.
2. A typical problem involving clustering might be a case where we are given a set of points in R or N and we need to identify clusters
3. Active learning, on the other hand, is supervised classification technique.
4. For instance, if we have a set given in which the tags are not known, then we make use of clustering to find those tags. The label for the representative value of a cluster would be the label for the entire cluster.

Clustering is unsupervised still it helps in supervision. This is done by ensuring that it reduces the number of labels we need to find.

Let us assume a data set (x,y) and let $f_\theta(x)$ represent the map from x to y with parameters θ:

$$(x, y) \rightarrow f_\theta(x) \xrightarrow{train} f_{\hat{\theta}}(x)$$

For example, we might need to search for a query among a large number of results. Let x_c represent the corpus data and x_q represent the query. In such a case, it is not a good idea to search over the entire corpus data. Alternatively, we can cluster x_c to find the best mean representative to x_q. Thus, it helps in restricting the run search space. Let us consider a function f_θ, which says how relevant a corpus is to the query, $f_\theta(x_q,x_c) = +1$ implies c will be returned to query q

$f_\theta(x_q,x_c) = -1$ implies c will not be returned to query q

where c $\in$ e. So the process of clustering e into k cluster involves:

1. Cluster e into k groups
2. Provide cluster IDs
3. Map the query to cluster ID i
4. Compute $f_{\hat{\theta}}(x_q, x_c)$ for all c in cluster i
5. Return relevant and non-relevant corpus items

Now we introduce the notion of a similarity function with cosine similarity as an example. Let $h(x_q)$ and $h(x_c)$ be the suitable vector representation to each data point x_q and x_c respectively. Define the cosine similarity between the two as: cos(h(x_q),h(x_c)) := $\frac{h(x_q)}{\|h(x_q)\|}.\frac{h(x_c)}{\|h(x_c)\|}$

This gives a measure of the degree of closeness between the two vectors $h(x_c)$ and $h(x_q)$. Now consider the following which can make more rigorous the "similarity" aspect of the cosine similarity s:
Let u $\sim$ N(0,I) be a random vector. For each vector $h(x_c)$ pick a representative vector as: Rep(h(x_c)) = sgn[$\frac{u}{||u||}\odot$h(x_c)]. These representative vectors can be said to divide this space of dimension d into 2^d clusters. We can find suitable bounds:

If s(h(x_q), h(x_c)) > R then P(Rep(h(x_q)) = Rep(h(x_c))) $\geq \epsilon$
If s(h(x_q), h(x_c)) < R/a (where a>1) then P(Rep(h(x_q)) = Rep(h(x_c))) $\leq \epsilon_1$
where $\epsilon >> \epsilon_1$.
This indicates that s indeed behaves like a "similarity function".

15.2 Similarity and Difference Metrics

Given 2 vectors, x_1 and x_2, we have looked at the cosine similarity function

$$s = cos(x_1, x_2) = \frac{x_1}{\|x_1\|}.\frac{x_2}{\|x_2\|} \tag{15.1}$$

We can thus define $v_1 = \frac{x_1}{\|x_1\|}$ and $v_2 = \frac{x_2}{\|x_2\|}$ so that

$$s = v_1.v_2 \quad and \quad \|v_1\| = \|v_2\| = 1 \tag{15.2}$$

We can further define an intuitive distance function, Δ for v_1 and v_2 only in terms of s in the following way

$$\Delta = \sqrt{2 - 2s} = \sqrt{(v_1 - v_2)^2} = \sqrt{\left(\frac{x_1}{\|x_1\|} - \frac{x_2}{\|x_2\|}\right)^2} \tag{15.3}$$

But what if we do not know the norm of both the vectors? How do we proceed then?

Let us take an example. We would like to convert $\frac{x_1}{\|x_1\|}.x_2$ to a cosine similarity under the condition that $\|x_2\| < B$. Therefore, we need to find v_1 and v_2 such that

$$v_1.v_2 = \frac{x_1}{\|x_1\|}.x_2 \quad and \quad \|v_1\| = \|v_2\| = 1 \tag{15.4}$$

We can do this by increasing the dimensionality by 1 and defining v_1 and v_2 in the following manner

$$v_1 = \begin{bmatrix} \frac{x_1}{\|x_1\|} \\ 0 \end{bmatrix} \quad and \quad v_2 = \begin{bmatrix} x_2 \\ \sqrt{B^2 - x_2^2} \end{bmatrix} \tag{15.5}$$

We can easily verify that

$$v_1.v_2 = \frac{x_1}{\|x_1\|}.x_2 \quad and \quad \|v_1\| = \|v_2\| = 1 \tag{15.6}$$

We can extend this idea to convert $x_1.x_2$ to a cosine similarity under the condition that $\|x_1\| < B_1$ and $\|x_2\| < B_2$ by defining v_1 and v_2 in the following manner

$$v_1 = \begin{bmatrix} x_1 \\ 0 \\ \sqrt{B_1^2 - x_1^2} \end{bmatrix} \quad and \quad v_2 = \begin{bmatrix} x_2 \\ \sqrt{B_2^2 - x_2^2} \\ 0 \end{bmatrix} \tag{15.7}$$

Now, let us look at Difference metrics for Sets.

Given 2 sets, $S_1 = \{x_1, x_2, ..., x_n\}$ and $S_2 = \{x'_1, x'_2, ..., x'_n\}$, we define the following distance function:

$$\Delta(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{15.8}$$

Verify whether this function is indeed a distance metric.

Note that for a distance function, Δ to be a distance metric, it must satisfy the following 3 conditions:

$$\Delta(S, S) = 0 \tag{15.9}$$

$$\Delta(S_1, S_2) \geq 0 \quad and \quad \Delta(S_1, S_2) = 0 \implies S_1 = S_2 \tag{15.10}$$

$$\Delta(S_1, S_3) + \Delta(S_2, S_3) \geq \Delta(S_1, S_2) \tag{15.11}$$

Hint: The distance function given above is indeed a distance metric (called the Jaccard Distance). For proving that it satisfies the 3 conditions, one can look up the Steinhaus Transform.

15.3 K means Clustering

Suppose we are given a dataset $D = \{x_1, x_2, .., x_n\}$ (here suppose the elements are vectors). Our aim is to split this data into K classes/clusters $C_1, C_2, ..., C_K$ such that points in each cluster are "closest" to their own cluster. We can formally say that we want to solve the following optimisation problem. Lets call it **Problem A**:

$$\begin{aligned} &\min_{C_1,..,C_K} \sum_{k=1}^{K} \sum_{i,j \in C_k} \|x_i - x_j\|^2 \\ &\text{with: } D = \bigcup_{i=1}^{K} C_i \ , \ C_i \cap C_j = \emptyset \ \forall \ i \neq j \end{aligned}$$

However, it is more convenient for us to solve the following problem. Lets call it **Problem B**:

$$\begin{aligned} &\min_{\{\bar{x}_k\},\{C_k\}} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2 \\ &\text{with: } D = \bigcup_{i=1}^{K} C_i \ , \ C_i \cap C_j = \emptyset \ \forall \ i \neq j \end{aligned}$$

Note that to minimise the objective function, we need to choose $\bar{x}_k$ as the mean of cluster C_k, so we can say: $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$.

We can rewrite the above two problems using a collection of weights p_{ik}, which indicate the cluster each element belongs to.
Here $p_{ik} = 1$ iff $i \in C_k$ and 0 otherwise:

$$\begin{aligned} &\textbf{Problem A: } \min_{\{p_{ik}\}} \sum_{k=1}^{K} \sum_{i,j \in D} p_{ik} \ p_{jk} \ \|x_i - x_j\|^2 \\ &\text{with: } \sum_{k=1}^{K} p_{ik} = 1 \end{aligned}$$

$$\begin{aligned} &\textbf{Problem B: } \min_{\{\bar{x}_k\},\{p_{ik}\}} \sum_{k=1}^{K} \sum_{i \in D} p_{ik} \|x_i - \bar{x}_k\|^2 \\ &\text{with: } \sum_{k=1}^{K} p_{ik} = 1 \end{aligned}$$

This just rewrites the optimisation problem w.r.t. the classes C_k as an optimisation w.r.t. a set of coefficients p_{ik}.

Now let us try and somewhat relate the problems A and B using simple algebra. Consider a set C of size $|C|$ (C can be any cluster C_k), containing elements $x_1, x_2, .., x_{|C|}$ (taken to be vectors). Let $\bar{x} := \frac{1}{|C|} \sum_{i \in C} x_i$ be the mean of the elements. We have:

$$\sum_{i \in C} \|x_i - \bar{x}\|^2 = \sum_{i \in C} \|x_i - \frac{1}{|C|} \sum_{j \in C} x_j\|^2 \tag{15.12}$$

$$= \frac{1}{|C|^2} \sum_{i \in C} \| \sum_{j \in C} (x_i - x_j)\|^2 = \frac{1}{|C|^2} \sum_{i \in C} (\sum_{j \in C} (x_i - x_j)^T \sum_{k \in C} (x_i - x_k)) \tag{15.13}$$

$$= \frac{1}{|C|^2} \sum_{i,j,k \in C} (x_i - x_j)^T (x_i - x_k) \tag{15.14}$$

$$= \frac{1}{|C|^2} \sum_{i,j,k \in C} (x_i - x_j)^T (x_i - x_j) + \frac{1}{|C|^2} \sum_{i,j,k \in C} (x_i - x_j)^T (x_j - x_k) \tag{15.15}$$

$$= \frac{1}{|C|} \sum_{i,j \in C} \|x_i - x_j\|^2 + \frac{1}{|C|} \sum_{i,j \in C} (x_i - x_j)^T (x_j - \bar{x}) \tag{15.16}$$

$$= \frac{1}{|C|} \sum_{i,j \in C} \|x_i - x_j\|^2 + \sum_{j \in C} (\bar{x} - x_j)^T (x_j - \bar{x}) \tag{15.17}$$

$$= \frac{1}{|C|} \sum_{i,j \in C} \|x_i - x_j\|^2 - \sum_{j \in C} \|x_j - \bar{x}\|^2 \tag{15.18}$$

On rearranging the final equivalence, we obtain:

$$\sum_{i \in C} \|x_i - \bar{x}\|^2 = \frac{1}{2|C|} \sum_{i,j \in C} \|x_i - x_j\|^2 \tag{15.19}$$

From this result, we cannot deduce the equivalence of the two problems. This is because each cluster size $|C_k|$ may be different and so due to the $2|C_k|$ factor in the above result, the optimisation functions are not related by a constant factor.

15.4 Group Details and Individual Contribution

The names, roll numbers and contributions of the group members who worked on this scribe are mentioned below:

- Raghav Rander (200040113) - Section 15.1
- Devak Sinha (180070017) - Section 15.2
- Waqar Mirza (200070090) - Section 15.3