week1

Rajesh Kalakoti

2023-08-03

- Packages
  - devtools
  - tidyverse
    - sub-sub-item 1

Clustering

Given a clustering $C = \{C_1, C_2, \ldots, C_k\}$, we need some scoring function that evaluates its quality or goodness. This sum of squared errors scoring function is defined as:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i:C(i)=k} \|x_i - \bar{x}_k\|^2$$

The goal is to find the clustering that minimizes:

$$C^* = \arg\min_{C}\{W(c)\}$$

K-means employs a greedy iterative approach to find a clustering that minimizes loss function.

Algorithm 13.1: K-means Algorithm

K-means (D, k,): 1. Initialize $t = 0$. Randomly initialize k centroids: $\mu_{t1}, \mu_{t2}, \ldots, \mu_{tk} \in \mathbb{R}^d$. 3. **repeat** 4. $t \leftarrow t + 1$. // **Cluster Assignment Step** 5. **foreach** $x_j \in D$ **do** 6. $j^* \leftarrow \arg\min_i \|x_j - \mu_{ti}\|^2$. // **Assign** x_j **to closest centroid** 7. $C_{j^*} \leftarrow C_{j^*} \cup \{x_j\}$. // **Centroid Update Step** 8. **foreach** $i = 1$ **to** k **do** 9. $\mu_{ti} \leftarrow \frac{1}{|C_i|}\sum_{x_j \in C_i} x_j$. **until** $\sum_{i=1}^{k} \|\mu_{ti} - \mu_{t-1i}\|^2 \leq \varepsilon$.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
?entropy
```

```
## No documentation for 'entropy' in specified packages and libraries:
## you could try '??entropy'
```

Algorithm 1: K-means Algorithm

Data: D, k, ε
Result: Result y

K-means(D, k, ε)
$t \leftarrow 0$;
Randomly initialize k centroids: $\mu_1^t, \mu_2^t, ..., \mu_n^t \in \mathbb{R}^d$;
repeat
    $t \leftarrow t + 1$;
    /* Cluster assignment step */
    for $x_j \in D$ **do**
        $j^* \leftarrow \text{argmin}_i\{||x_j - \mu_i^t||^2\}$;
        /* assign x_j to closest centroid */
        $C_{j^*} \leftarrow C_{j^*} \cup \{x_j\}$;
    end
until *termination condition*;
Perform additional steps;
for $i = 1$ *to* n **do**
    Update $y \leftarrow y \times x$;
end
return y;