Data Clustering: Some Other Aspects (K-means++, Overlapping Clustering, Evaluation)

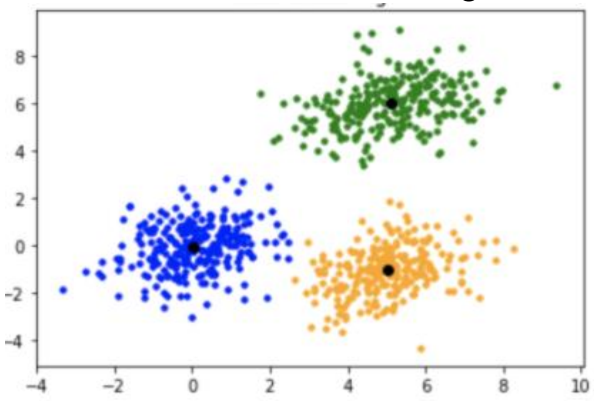
CS771: Introduction to Machine Learning

Piyush Rai

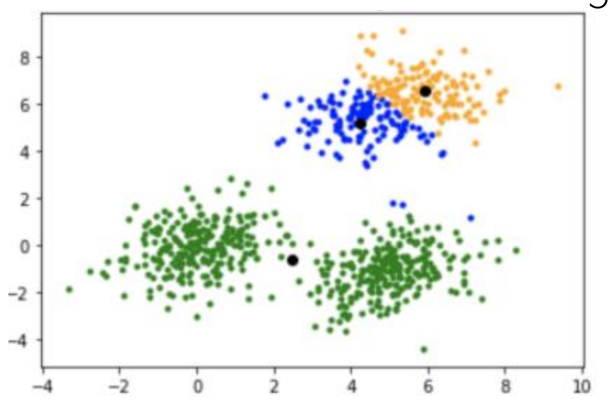
K-means++

- K-means results can be sensitive to initialization

Desired clustering



Poor initialization: bad clustering



- K-means++ (Arthur and Vassilvitskii, 2007) an improvement over K-means
  - Only difference is the way we initialize the cluster centers (rest of it is just K-means)
  - Basic idea: Initialize cluster centers such that they are reasonably far from each other
  - Note: In K-means++, the cluster centers are chosen to be K of the data points themselves

K-means++

- K-means++ works as follows
  - Choose the first cluster mean uniformly randomly to be one of the data points
  - The subsequent $K-1$ cluster means are chosen as follows
    1. For each unselected point $\boldsymbol{x}$, compute its smallest distance $D(\boldsymbol{x})$ from already initialized means
    2. Select the next cluster mean unif. rand. to be one of the unselected points based on probability prop. to $D(\boldsymbol{x})^2$
    3. Repeat 1 and 2 until the $K-1$ cluster means are initialized

Thus farthest points are most likely to be selected as cluster means

  - Now run standard K-means with these initial cluster means
- K-means++ initialization scheme sort of ensures that the initial cluster means are located in different clusters

Overlapping Clustering

Kind of unsupervised version of multi-label classification (just like standard clustering is like unsupervised multi-class classification)

Example: Clustering people based on the interests they may have (a person may have multiple interests; thus may belong to more than one cluster simultaneously)

- Have seen hard clustering and soft clustering
- In hard clustering, z_n is a one-hot vector
- In soft clustering, z_n is a vector of probabilities
- Overlapping Clustering: A point can simultaneously belong to multiple clusters
  - This is different from soft-clustering
  - z_n would be a binary vector, rather than a one hot or probability vector, e.g.,

$$z_n = [1\ 0\ 0\ 1\ 0]$$

K=5 clusters with point x_n belonging (in whole, not in terms of probabilities) to clusters 1 and 4

- In general, more difficult than hard/soft clustering (for N data points and K clusters, the size of the space of possible solutions is not K^N but 2^{NK} - exp in both N and K)
- K-means has extensions* for doing overlapping clustering. There also exist latent variable models for doing overlapping clustering

*An extended version of the k-means method for overlapping clustering (Cleuziou, 2008); Non-exhaustive, Overlapping k-means (Whang et al, 2015)

Evaluating Clustering Algorithms

- Clustering algos are in general harder to evaluate since we rarely know the ground truth clustering (since clustering is unsupervised)

- If ground truth labels not available, use output of clustering for some other task
  - For example, use cluster assignment z_n (hard or soft) as a new feature representation
  - Performance on some task using this new rep. is a measure of goodness of clustering

- If ground truth labels are available, can compare them with clustering based labels
  - Not straightforward to compute accuracy since the label identities may not be the same, e.g.,

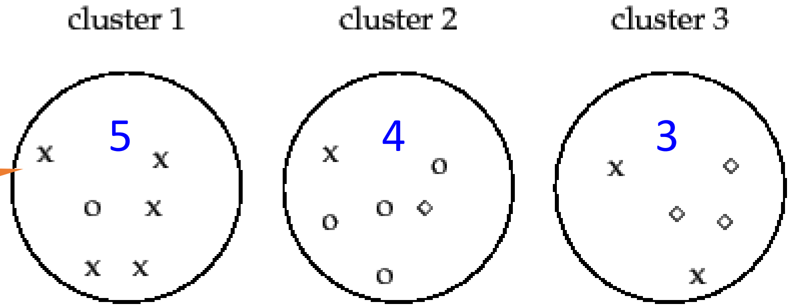
    Ground truth = [1 1 1 0 0 0]   Clustering = [0 0 0 1 1 1]

    (Perfect clustering but zero "accuracy" if we just do a direct match)
  - There are various metrics that take into account the above fact
    - Purity, Rand Index, F-score, Normalized Mutual Information, etc

Evaluating Clustering Algorithms

- Purity: Looks at how many points in each cluster belong to the majority class in that cluster

3 classes (x, o, △, assuming known ground truth labels)



Sum and divide by total number of points

Purity = (5+4+3)/17 ≈ 0.71

Close to 0 for bad clustering, 1 for perfect clustering

Also a bad metric if number of clusters is very large – each cluster will be kind of pure anyway

- Rand Index (RI): Can also look at what fractions of pairs of points with same (resp. different) label are assigned to same (resp. different) cluster

True Positive: No. of pairs with same true label and same cluster

True Negative: No. of pairs with diff true label and diff clusters

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

False Positive: No. of pairs with diff true label and same cluster

False Negative: No. of pairs with same true label and diff cluster

F_β score is also popular

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Precision

Recall

Coming up next

- Latent variable models