Clustering

New data(predictions)

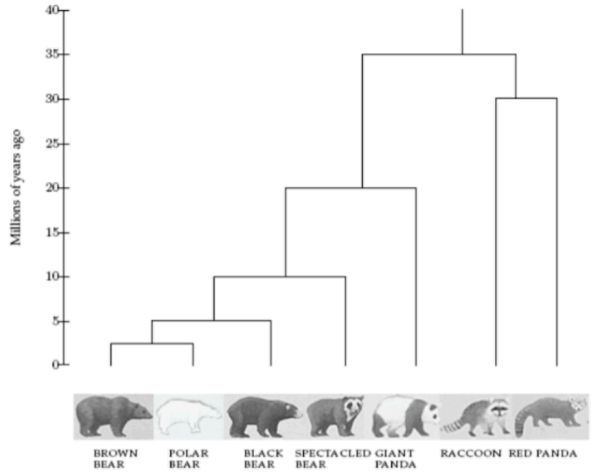
Goal: C = {c1,c2,,...ck}

Such that within cluster the similarity is minimized

2 main types of clustering

- Flat/Partitional
  - K-means
  - Gaussian mixture methods
- Hierarchical
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and neighbor joining

Hierarchical clustering example picture:




K-Means

Initialization step: choose k means (cluster centers) randomly from the data

It is an expectation-maximization (EM) algorithm

- E-step: assign each datapoint to the closest mean
- M-step: recomputing the means as the cluster average

Need to consider what the K-means algorithm minimizes

-within cluster similarity (minimize the distances between the points and the cluster mean)

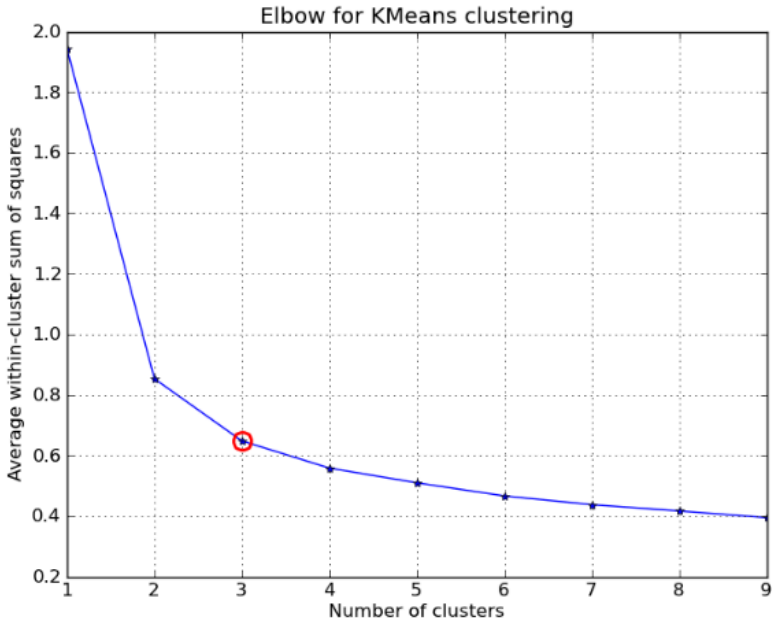
AND what is maximized

Stopping criteria

-no change in cluster membership

-max # of iterations exceeded
-configuration/pattern you've seen before

Elbow plots (how to choose k)




Handout 23

Gaussian Mixed models (GMMs)
- does not allow points to belong to multiple clusters

Does not account for different cluster sizes
Not generative (cannot create new data point(s))

Discriminative vs generative algorithms
Discriminative
Logistic regression, k means
Generative
Naive bayes, gaussian mixture methods