CS 2750: Machine Learning
Clustering

Prof. Adriana Kovashka
University of Pittsburgh
January 25, 2016
What is clustering?

• Grouping items that “belong together” (i.e. have similar features)

• Unsupervised: we only use the features X, not the labels Y

• This is useful because we may not have any labels but we can still detect patterns

• If goal is classification, we can later ask a human to label each group (cluster)
Unsupervised visual discovery

• We don’t know what the objects in red boxes are, but we know they tend to occur in similar context
• If features = the context, objects in red will cluster together
• Then ask human for a label on one example from the cluster, and keep learning new object categories iteratively

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Why do we cluster?

• Summarizing data
  – Look at large amounts of data
  – Represent a large continuous vector with the cluster number

• Counting
  – Computing feature histograms

• Prediction
  – Images in the same cluster may have the same labels

• Segmentation
  – Separate the image into different regions

Slide credit: J. Hays, D. Hoiem
Image segmentation via clustering

• Separate image into coherent “objects”

Source: L. Lazebnik
Image segmentation via clustering

• Separate image into coherent “objects”
• Group together similar-looking pixels for efficiency of further processing

“superpixels”


Source: L. Lazebnik
Today

• Clustering: motivation and applications

• Algorithms

  – K-means (iterate between finding centers and assigning points)
  
  – Mean-shift (find modes in the data)
  
  – Hierarchical clustering (start with all points in separate clusters and merge)
  
  – Normalized cuts (split nodes in a graph based on similarity)
• These intensities define the three groups.
• We could label every pixel in the image according to which of these primary intensities it is.
  • i.e., segment the image based on the intensity feature.
• What if the image isn’t quite so simple?
• Now how to determine the three main intensities that define our groups?
• We need to *cluster*. 

Source: K. Grauman
• Goal: choose three “centers” as the representative intensities, and label every pixel according to which of these centers it is nearest to.

• Best cluster centers are those that minimize SSD between all points and their nearest cluster center $c_i$:

$$
\sum_{\text{clusters } i} \sum_{\text{points } p \text{ in cluster } i} \|p - c_i\|^2
$$

Source: K. Grauman
Clustering

• With this objective, it is a “chicken and egg” problem:
  – If we knew the cluster centers, we could allocate points to groups by assigning each to its closest center.
  – If we knew the group memberships, we could get the centers by computing the mean per group.

Source: K. Grauman
K-means clustering

- Basic idea: randomly initialize the $k$ cluster centers, and iterate between the two steps we just saw.

1. *Randomly* initialize the cluster centers, $c_1, \ldots, c_K$
2. *Given cluster centers*, determine points in each cluster
   - For each point $p$, find the closest $c_i$. Put $p$ into cluster $i$
3. *Given points in each cluster*, solve for $c_i$
   - Set $c_i$ to be the mean of points in cluster $i$
4. If $c_i$ have changed, repeat Step 2

Properties
- Will always converge to *some* solution
- Can be a “local minimum”
  - does not always find the global minimum of objective function:
    \[
    \sum_{\text{clusters } i} \sum_{\text{points } p \text{ in cluster } i} \|p - c_i\|^2
    \]
K-means

1. Ask user how many clusters they’d like. *(e.g. k=5)*
K-means

1. Ask user how many clusters they’d like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
K-means

1. Ask user how many clusters they’d like. \((\text{e.g. } k=5)\)

2. Randomly guess \(k\) cluster Center locations

3. Each datapoint finds out which Center it’s closest to. (Thus each Center “owns” a set of datapoints)
K-means

1. Ask user how many clusters they’d like. *(e.g. $k=5$)*

2. Randomly guess $k$ cluster Center locations

3. Each datapoint finds out which Center it’s closest to.

4. Each Center finds the centroid of the points it owns
K-means

1. Ask user how many clusters they’d like.\textit{(e.g. }k=5\text{)}

2. Randomly guess \( k \) cluster Center locations

3. Each datapoint finds out which Center it’s closest to.

4. Each Center finds the centroid of the points it owns...

5. ...and jumps there

6. ...Repeat until terminated!
K-means converges to a local minimum

Figure from Wikipedia
K-means clustering

- Java demo
  http://home.dei.polimi.it/matteucc/Clustering/tutoria l_html/AppletKM.html

- Matlab demo
  http://www.cs.pitt.edu/~kovashka/cs1699/kmeans demo.m
Time Complexity

• Let $n =$ number of instances, $m =$ dimensionality of the vectors, $k =$ number of clusters
• Assume computing distance between two instances is $O(m)$
• Reassigning clusters:
  – $O(kn)$ distance computations, or $O(knm)$
• Computing centroids:
  – Each instance vector gets added once to a centroid: $O(nm)$
• Assume these two steps are each done once for a fixed number of iterations $l$: $O(lknm)$
  – Linear in all relevant factors

Adapted from Ray Mooney
K-means Variations

• K-means:

\[ \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \]

\[ J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \mu_k \|^2 \]

• K-medoids:

\[ \tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(x_n, \mu_k) \]
Distance Metrics

• Euclidian distance ($L_2$ norm):
  \[ L_2(\vec{x}, \vec{y}) = \sum_{i=1}^{m} (x_i - y_i)^2 \]

• $L_1$ norm:
  \[ L_1(\vec{x}, \vec{y}) = \sum_{i=1}^{m} |x_i - y_i| \]

• Cosine Similarity (transform to a distance by subtracting from 1):
  \[ 1 - \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||} \]
Segmentation as clustering

Depending on what we choose as the *feature space*, we can group pixels in different ways.

Grouping pixels based on *intensity* similarity

Feature space: intensity value (1-d)

Source: K. Grauman
quantization of the feature space; segmentation label map

Source: K. Grauman
Segmentation as clustering

Depending on what we choose as the feature space, we can group pixels in different ways.

Grouping pixels based on color similarity

Feature space: color value (3-d)
K-means: pros and cons

**Pros**
- Simple, fast to compute
- Converges to local minimum of within-cluster squared error

**Cons/issues**
- Setting k?
  - One way: silhouette coefficient
- Sensitive to initial centers
  - Use heuristics or output of another method
- Sensitive to outliers
- Detects spherical clusters

Adapted from K. Grauman
Today

• Clustering: motivation and applications
• Algorithms
  – K-means (iterate between finding centers and assigning points)
  – Mean-shift (find modes in the data)
  – Hierarchical clustering (start with all points in separate clusters and merge)
  – Normalized cuts (split nodes in a graph based on similarity)
Mean shift algorithm

- The mean shift algorithm seeks *modes* or local maxima of density in the feature space

*Image*  
*Feature space*  
(L* u* v* color values)

Source: K. Grauman
Kernel density estimation

Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. It is particularly useful when the underlying distribution is unknown or complex.

In the diagram, the blue points represent the data points (1-D), the red line represents the estimated density, and the blue curve inside the box represents the kernel. The kernel is a smoothing function that is applied to the data points to estimate the density.

Source: D. Hoiem
Mean shift

Search window
Center of mass
Mean Shift vector
Mean shift
Mean shift

Search window
Center of mass

Mean Shift vector
Mean shift

Search window
Center of mass
Mean Shift vector
Mean shift
Mean shift

Search window
Center of mass

Mean Shift vector

Slide by Y. Ukrainitz & B. Sarel
Mean shift

Search window

Center of mass

Slide by Y. Ukrainitz & B. Sarel
Points in same cluster converge
Mean shift clustering

- Cluster: all data points in the attraction basin of a mode
- Attraction basin: the region for which all trajectories lead to the same mode
Simple Mean Shift procedure:
- Compute mean shift vector
- Translate the Kernel window by $m(x)$
Mean shift clustering/segmentation

- Compute features for each point (color, texture, etc)
- Initialize windows at individual feature points
- Perform mean shift for each window until convergence
- Merge windows that end up near the same “peak” or mode

Source: D. Hoiem
Mean shift segmentation results

http://www.caip.rutgers.edu/~comanici/MSPAMI/msPamiResults.html
Mean shift segmentation results
Mean shift

• **Pros:**
  – Does not assume shape on clusters
  – Robust to outliers

• **Cons/issues:**
  – Need to choose window size
  – Does not scale well with dimension of feature space
  – Expensive: O(I n^2)
Mean-shift reading

• Nicely written mean-shift explanation (with math)
  http://saravananthirumuruganathan.wordpress.com/2010/04/01/introduction-to-mean-shift-algorithm/
  • Includes .m code for mean-shift clustering

• Mean-shift paper by Comaniciu and Meer
  http://www.caip.rutgers.edu/~comanici/Papers/MsRobustApproach.pdf

• Adaptive mean shift in higher dimensions

Source: K. Grauman
Today

• Clustering: motivation and applications
• Algorithms
  – K-means (iterate between finding centers and assigning points)
  – Mean-shift (find modes in the data)
  – Hierarchical clustering (start with all points in separate clusters and merge)
  – Normalized cuts (split nodes in a graph based on similarity)
Hierarchical Agglomerative Clustering (HAC)

• Assumes a *similarity function* for determining the similarity of two instances.
• Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster.
• The history of merging forms a binary tree or hierarchy.
HAC Algorithm

Start with all instances in their own cluster.
Until there is only one cluster:
   Among the current clusters, determine the two
   clusters, $c_i$ and $c_j$, that are most similar.
   Replace $c_i$ and $c_j$ with a single cluster $c_i \cup c_j$
Agglomerative clustering

1. Say “Every point is its own cluster”
Agglomerative clustering

1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters
Agglomerative clustering

1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
Agglomerative clustering

1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
4. Repeat
Agglomerative clustering

1. Say “Every point is its own cluster”
2. Find “most similar” pair of clusters
3. Merge it into a parent cluster
4. Repeat

Copyright © 2001, 2004, Andrew W. Moore
Agglomerative clustering

How many clusters?

- Clustering creates a dendrogram (a tree)
- To get final clusters, pick a threshold
  - max number of clusters or
  - max distance within clusters (y axis)

Adapted from J. Hays
Cluster Similarity

- How to compute similarity of two clusters each possibly containing multiple instances?

  - **Single Link**: Similarity of *two most similar* members.
    \[
    sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)
    \]

  - **Complete Link**: Similarity of *two least similar* members.
    \[
    sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)
    \]

  - **Group Average**: Average similarity between members.

Adapted from Ray Mooney
Today

- Clustering: motivation and applications
- Algorithms
  - K-means (iterate between finding centers and assigning points)
  - Mean-shift (find modes in the data)
  - Hierarchical clustering (start with all points in separate clusters and merge)
  - Normalized cuts (split nodes in a graph based on similarity)
Images as graphs

Fully-connected graph

- node (vertex) for every pixel
- link between every pair of pixels, \( p, q \)
- affinity weight \( w_{pq} \) for each link (edge)
  - \( w_{pq} \) measures similarity
    - similarity is inversely proportional to difference (in color and position…)

Source: Steve Seitz
Segmentation by Graph Cuts

Break Graph into Segments

- Want to delete links that cross **between** segments
- Easiest to break links that have low similarity (low weight)
  - similar pixels should be in the same segments
  - dissimilar pixels should be in different segments

Source: Steve Seitz
Concluding thoughts

• Lots of ways to do clustering
• How to evaluate performance?
  – Purity

\[\text{purity}(\Omega, \mathcal{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|\]

where \( \Omega = \{\omega_1, \omega_2, \ldots, \omega_K\} \) is the set of clusters and \( \mathcal{C} = \{c_1, c_2, \ldots, c_J\} \) is the set of classes

– Might depend on application