Machine Learning for Large-Scale Data Analysis and Decision Making
80-629-17A

Unsupervised Learning
— Week #7
Today

- Unsupervised learning
- Definition and properties
- A few clustering models
Experience (E)

- What data does f experience?
- (Focus on algorithms that experience whole datasets)

1. Unsupervised. Examples alone
   \[ \{x_i\}_{i=0}^n \]

2. Supervised. Examples come with labels
   \[ \{(x_i, y_i)\}_{i=0}^n \]
1. Unsupervised

- Experience examples alone
  \[ \{x_i\}_{i=0}^{n} \]

- Learn “useful properties of the structure of the data”

- E.g., clustering, density modeling \((p(x))\), PCA, FA.
Different tasks

- Finding patterns
- Clustering
- Dimensionality reduction
- Density modelling
- ...
Clustering

\[ f : X \rightarrow \{1, \ldots, K\} \]

• Task: Assign each point to one of K clusters
  • Cluster: a set of similar points (a group)
  • Alternatively: Divide the space into K regions. Assign points to a cluster based on the region they lie in
  • Similar to classification.
Supervised

Unsupervised
Clustering

- Desideratum: group similar points

- Similarity is often defined as being close according to some similarity or distance function

- E.g., in Euclidean space

\[
\text{distance}(x_i, x_j) = \| x_i - x_j \|^2_2
\]
K-means clustering

- A particular clustering model (and accompanying algorithm)
- For a particular number of clusters (K), find cluster centers \( \mu_k \) that minimize the within cluster distance

\[
\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
\]

\[
r = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
\vdots & \vdots \\
0 & 1
\end{bmatrix}_{N \times K}
\]
K-means clustering

- A particular clustering model (and accompanying algorithm)

- For a particular number of clusters (K), find cluster centers \( \mu_k \) that minimize the within cluster distance

\[
\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
\]

- Algorithm to minimize the objective:
  - Initialize the cluster centers
  - Until convergence:
    1. Update \( r \)
    2. Update cluster centers \( \mu_k \) \( \forall k \)
K-means clustering

- A particular clustering model (and accompanying algorithm)
- For a particular number of clusters ($K$), find cluster centers that minimize the within cluster distance

$$\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2$$

- Algorithm to minimize the objective:
  - Initialize the cluster centers
  - Until convergence:
    1. Update $r$
    2. Update cluster centers $\mu_k \ \forall k$
K-means clustering

- A particular clustering model (and accompanying algorithm)
- For a particular number of clusters (K), find cluster centers \( \mu_k \) that minimize the within cluster distance

\[
\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
\]

- Algorithm to minimize the objective:
  - Initialize the cluster centers
  - Until convergence:
    1. Update \( r \)
    2. Update cluster centers \( \mu_k \) \( \forall k \)
K-means clustering

- A particular clustering model (and accompanying algorithm)
- For a particular number of clusters (K), find cluster centers \( \mu_k \) that minimize the within cluster distance

Objective: 

\[
\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
\]

- Algorithm to minimize the objective:
  - Initialize the cluster centers
  - Until convergence:
    1. Update \( r \)
    2. Update cluster centers \( \mu_k \) for all \( k \)

Laurent Charlin — 80-629
Initial cluster centers

Step 1

Step 2

Step 1

Step 2

Step 1

Final solution

[Figure 9.1 from PRML]
K-means on images

- Each pixel is a datum

\[ x_i = (r_i, g_i, b_i) \]

[Live link: http://104.196.67.140:1080/notebooks/kmeans_test.ipynb]
A few failure cases

- Only works with squared Euclidean distance
- Not robust to outliers
- What about non-continuous data?
- Tends to result in relatively uniform cluster sizes
- Hard assignments

K-medoids Objective:

$$K\text{-medoids Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} d(x_i, \mu_k)$$
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
<th>Image 5</th>
<th>Image 6</th>
<th>Image 7</th>
<th>Image 8</th>
</tr>
</thead>
</table>
Unsupervised learning

- In supervised learning you have:
  1. Clear metric (e.g., mean squared error, accuracy, etc.)
  2. Procedures for comparing different models

- Unsupervised learning
  - Unclear what the right metric is.
  - How to compare different models?
  - Data dimensions can have a big impact on solution
  - You should carefully select the input
A probabilistic approach to k-means clustering

Objective := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
A probabilistic approach to k-means clustering

Objective := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
A probabilistic approach to k-means clustering

Objective := \[
\sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2
\]
A probabilistic approach to k-means clustering

Objective := \[ \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2 \]

where:
- \( r_{ik} \) is the indicator of \( x_i \) to \( k \)
- \( \mu_k \) is the center of cluster \( k \)
- \( x_i \) is the data point

Example:
- \( r = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \vdots \\ 0 & 1 \end{bmatrix}_{N \times K} \)
- \( r_i = [0, 1]_{1 \times K} \)

Diagram:
- Data points in \( x_1 \) and \( x_2 \)
- Clusters represented by squares

Laurent Charlin – 80-629
A probabilistic approach to k-means clustering

Objective: $\sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|^2$

$r = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \vdots \\ 0 & 1 \end{bmatrix}_{N \times K}$

$r_{i:} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}_{1 \times K}$

$P(r_{i:}) = \text{Categorical}(\pi)$

$= \prod_{k=1}^{K} \pi_{ik}^{r_{ik}}$
A probabilistic approach to k-means clustering

Objective := \( \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \left\| x_i - \mu_k \right\|^2 \)

\( r = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \vdots \\ 0 & 1 \end{bmatrix}_{N \times K} \)

the i'th line: \( r_i = [0 \ 1]_{1 \times K} \)

\( x_i = [0.2 \ 0.4] \)

\( \mu_k = [0.3 \ 0.1] \)

\( P(r_{ik}) = \text{Categorical}(\pi) \)

\( = \prod_{k=1}^{K} \pi_{ik}^{r_{ik}} \)
A probabilistic approach to k-means clustering

Objective := \[ \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2 \]

indicator center

\[ r = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \vdots \\ 0 & 1 \end{bmatrix}_{N \times K} \]

data

the i'th line

\[ r_i = \begin{bmatrix} 0 & 1 \end{bmatrix}_{1 \times K} \]

\[ P(r_i) = \text{Categorical}(\pi) \quad P(x_i | r_{ik} = 1) = \mathcal{N}(x_i | \mu_k, \Sigma_k) \]

\[ = \prod_{k=1}^{K} \pi_k^{r_{ik}} \]
A probabilistic approach
to k-means clustering

\[
P(r_i) = \text{Categorical}(\pi)
\]

\[
P(x_i | r_{ik} = 1) = \mathcal{N}(x_i | \mu_k, \Sigma_k)
\]

\[
P(x_i) = \sum_r P(r)P(x_i | r)
\]

\[
= \sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)
\]
A probabilistic approach to k-means clustering

\[ P(r_i) = \text{Categorical}(\pi) \]

\[ P(x_i \mid r_{ik} = 1) = \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \]

\[ P(x_i) = \sum_r P(r)P(x_i \mid r) \]

\[ = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \]

\[ 0 \leq \pi \leq 1 \]

\[ \sum_k \pi_k = 1 \]
A probabilistic approach to k-means clustering

\[ P(r_i) = \text{Categorical}(\pi) \]
\[ P(x_i | r_{ik} = 1) = \mathcal{N}(x_i | \mu_k, \Sigma_k) \]

\[ P(x_i) = \sum_r P(r)P(x_i | r) \]
\[ = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \]

1) **Soft clustering**
- each cluster \( k \) explains \( \pi_k \) of each datapoint
- The posterior over \( z \) (\( P(z | x) \)) makes it more explicit

\[ 0 \leq \pi \leq 1 \]
\[ \sum_k \pi_k = 1 \]
A probabilistic approach to k-means clustering

$$P(r_i) = \text{Categorical}(\pi)$$

$$P(x_i \mid r_{ik} = 1) = \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

$$P(x_i) = \sum_r P(r)P(x_i \mid r)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

1) **Soft clustering**
   - each cluster \( k \) explains \( \pi_k \) of each datapoint
   - The posterior over \( z \) \( P(z \mid x_i) \) makes it more explicit

2) **Mixture of Gaussians (GMMs)**
   - convex combination of Gaussian distributions
A probabilistic approach to k-means clustering

- Estimation of the parameters: Maximum likelihood estimate
  - Could do it jointly (“a la” neural network)
  - Often in two steps (just like for the non-probabilistic version)
  - This leads to the EM algorithm
Initial cluster centers

(a) $L = 1$

(b) $L = 2$

(c) $L = 5$

(d) $L = 20$

Final solution

[Figure 9.8 from PRML]
Initial cluster centers

Final solution

k-means Final solution

[Figure 9.8 from PRML]
K-means

Similar

Similar

GMM better

GMMs
K-means

GMMs

Similar

Similar

GMM better

GMM better
K-means  GMMs

Similar

Similar

GMM better

GMM better

Similar
Comparing K-means to GMMs

- GMMs learns covariance matrix
  - Per cluster variance
  - Covariance terms
- GMMs has many more parameters
  - Covariance matrix (MxM)
Some thoughts on probabilistic approaches

- Many “equivalences” from known models/algorithms to probabilistic formulations
  - K-means -> Mixture of Gaussians
  - Linear Regression -> MLE in Gaussian model
  - Logistic Regression -> MLE in Bernoulli model

- Allow you to think of models in a common framework
  - Build a joint distribution from a marginal and a conditional
    \[ P(x, z) = P(x | z)P(z) \]

- Learning procedures
  - Well motivated
  - Can be shared across problems
Unsupervised learning takeaways

• Most data (in the world) is unlabeled

• Useful tasks: clustering, density estimation, dimensionality reduction

• K-means and Gaussian mixtures (GMMs)

• Performance is harder to judge (measure)

• All our examples were in 2D

• Can be used in downstream tasks