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)
  - Unsupervised. Examples alone \( \{X_i\}_{i=0}^{n} \)
  - Supervised. Examples come with labels \( \{(X_i, y_i)\}_{i=0}^{n} \)
1. Unsupervised

- Experience examples alone
- 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

\[
\text{Objective} := \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2^2
\]

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

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:
    - Update $r$
    - 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:
    - Update $r$
    - 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:
    • Update \( r \)
    • 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

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

- Algorithm to minimize the objective:
  - Initialize the cluster centers
  - Until convergence:
    - Update $r$
    - Update cluster centers $\mu_k \ \forall k$
Figure 9.1 from PRML
K-means on images

Each pixel is a datum

\[ x_i = (r_i, g_i, b_i) \]
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: \( \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} d(x_i, \mu_k) \)
Unsupervised learning

- In supervised learning you have:
  - Clear metric (e.g., mean squared error, accuracy, etc.)
  - 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 := \[ \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2^2 \]
A probabilistic approach to k-means clustering

Objective := \( \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2^2 \)

\[ N \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2^2 \]

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

the i'th line \( r_i = [0 \ 1]_{1 \times K} \)
A probabilistic approach to k-means clustering

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

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

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

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

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

Objective:\[
\sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - \mu_k \|_2^2
\]

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

the $i$'th line $r_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}_{1 \times K}$

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

$= \prod_{k=1}^{K} \pi_{rk}$

Data indicator center

$x_i = [0.2 \ 0.4]$

$\mu_k = [0.3 \ 0.1]$
A probabilistic approach to k-means clustering

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

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

the i'th line \( r_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}_{1 \times K} \)

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

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

\( x_i = [0.2, 0.4] \)

\( \mu_k = [0.3, 0.1] \)
A probabilistic approach to k-means clustering

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

\[ \Pr(r_{ik}) = \text{Categorical}(\pi) \]
A probabilistic approach to k-means clustering

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

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

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

$$= \sum_k \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \quad \theta := \{\mu_k, \Sigma_k\}, \forall k$$
A probabilistic approach to k-means clustering

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

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

\[
P(x_i \mid \theta) = \sum_r P(r)P(x_i \mid r, \theta)
= \sum_k \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)
\]

\[
\theta := \{\mu_k, \Sigma_k\}, \forall k
\]

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

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

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

\[ P(x_i \mid \theta) = \sum_r P(r) P(x_i \mid r, \theta) = \sum_k \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \quad \theta := \{\mu_k, \Sigma_k\}, \forall k \]

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

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

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

\[ P(x_i \mid \theta) = \sum_r P(r) P(x_i \mid r, \theta) \]
\[ = \sum_k \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \quad \theta := \{\mu_k, \Sigma_k\}, \forall k \]

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

2) **Mixture of Gaussians (GMMs)**
   convex combination of Gaussian distributions

0 ≤ \( \pi \) ≤ 1
\[ \sum_k \pi_k = 1 \]
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 separate steps (similar as the non-probabilistic version)
- This leads to the Expectation-Maximization (EM) algorithm
Initial cluster centers

Final solution

Figure 9.8 from PRML
Initial cluster centers

(a) $L = 1$

(b) $L = 2$

(c) $L = 5$

(d) $L = 20$

Final solution

(k-means)

[Figure 9.8 from PRML]
GMMs

Similar

GMM better

Similar

GMM better

Similar

Similar
Comparing K-means to GMMs

- GMMs learns covariance matrix
- Per cluster variance
- Covariance terms
- GMMs has many more parameters
- Covariance matrix (MxM)
Autoencoders

- A neural network architecture for unsupervised learning
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)
  - Note that all our examples were in 2D
- Can be used in downstream tasks