Today

- Unsupervised learning
- Definition and properties
- A few clustering models
- Other instantiations of unsupervised learning
Unsupervised Learning
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 $f : X \rightarrow \{1, 2, \ldots, K\}$ (K clusters)
- Dimensionality reduction $f : X^p \rightarrow X^k$, $k << p$
- Density modelling $f : X \rightarrow [0, 1]$
- ...

\[ \begin{array}{c}
\text{Laurent Charlin — 80-629} \\
\text{6}
\end{array} \]
Why unsupervised learning?

1. Understand properties of the data
2. Learn useful representations
3. Use the results in a downstream application
   - Opportunity: There are lots of unlabelled data
k-means clustering
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.
Clustering

- Desideratum: group similar points together

- Similarity is often defined as being close according to some similarity or (inverse) 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)
  - There are K clusters. Each point belongs to a cluster. Clusters have centers: $\mu$
  - Objective: Find cluster centers $\mu_k$ that minimize the within cluster distance
K-means clustering

- A particular clustering model (and accompanying algorithm)

- There are $K$ clusters. Each point belongs to a cluster. Clusters have centers: $\mu$

- Objective: Find cluster centers $\mu_k$ that minimize the within cluster distance

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

$$r = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}_{N \times 2}$$
K-means clustering

- A particular clustering model (and accompanying algorithm)
- There are \( K \) clusters. Each point belongs to a cluster. Clusters have centers: \( \mu \)
- Objective: 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 responsibilities: \( r \)
    2. Update cluster centers: \( \mu_k \) \( \forall k \)
K-means clustering

- A particular clustering model (and accompanying algorithm)
  - There are $K$ clusters. Each point belongs to a cluster. Clusters have centers: $\mu$
  - Objective: 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 responsibilities: $r$
    2. Update cluster centers: $\mu_k \ \forall k$
K-means clustering

- A particular clustering model (and accompanying algorithm)
  - There are $K$ clusters. Each point belongs to a cluster. Clusters have centers: $\mu$
  - Objective: 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 responsibilities: $r$
    2. Update cluster centers: $\mu_k \ \forall k$
K-means clustering

• A particular clustering model (and accompanying algorithm)

• There are $K$ clusters. Each point belongs to a cluster. Clusters have centers: $\mu$

• Objective: 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:

  1. Initialize the cluster centers
  2. Until convergence:

    1. Update responsibilities: $r$
    2. Update cluster centers: $\mu_k \forall k$
Algorithm

- Initialize the cluster centers
- Until convergence:
  1. Update responsibilities
  2. Update cluster centers

[Figure 9.1 from PRML]
K-means on images

Each pixel is a datum

\[ x_i = (r_i, g_i, b_i) \]
K-means on images

Each pixel is a datum

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

- We cluster pixels by color
- I.e., pixels of similar color will be belong to the same cluster
- We re-draw the image by replacing the color of each pixel by the color of its cluster
A few failure cases of k-means

- Only works with squared Euclidean distance
- What about non-continuous data?
  - K-medoids Objective: \[ \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} d(x_i, \mu_k) \]
- Not robust to outliers
- Tends to result in relatively uniform cluster sizes
- Hard assignments

[http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html]
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, domain dependant
  • How to compare different models?
  • Data dimensions can have a big impact on solution
  • You should carefully select the input
Gaussian Mixture for Probabilistic Clustering
Motivation

Clustering found using K-means
A probabilistic approach to k-means clustering

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

K-means Clustering

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

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

K-means Clustering

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

Soft K-means Clustering
A probabilistic approach to k-means clustering

K-means Clustering

Soft K-means Clustering

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

- Responsibilities are continuous [0, 1]
- Each cluster has a responsibility: \( \pi_k \)
- Each cluster models data using a Gaussian: \( \mathcal{N}(x_i | \mu_k, \Sigma_k) \)
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i | \mu_k, \Sigma_k)$
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)$$
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)$$

where

- $0 \leq \pi \leq 1$
- $\sum_k \pi_k = 1$
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)$$

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} P(z_i = k) P(x_i \mid \mu_k, \Sigma_k)$$
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)$$

$$0 \leq \pi \leq 1$$
$$\sum_k \pi_k = 1$$

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} P(z_i = k) P(x_i \mid \mu_k, \Sigma_k)$$

**Posterior:** $P(z_i = k \mid x_i) \propto P(z_i = k) P(x_i \mid \mu_k, \Sigma_k)$
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$
P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)
$$

$$
\quad 0 \leq \pi \leq 1
$$

$$
\sum_{k} \pi_k = 1
$$

$$
P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} P(z_i = k)P(x_i \mid \mu_k, \Sigma_k)
$$

Posterior: $P(z_i = k \mid x_i) \propto P(z_i = k)P(x_i \mid \mu_k, \Sigma_k)$

- Model is known as a Gaussian Mixture Model
Soft K-means Clustering

- Each cluster gives a probability to each datum: $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

- We can write the complete model as:

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} \pi_k P(x_i \mid \mu_k, \Sigma_k)$$

  - Model is known as a Gaussian Mixture Model
  - Advantages over k-means:
    - Soft Clustering
    - Clusters don’t have to be spherical

$$P(x_i \mid \text{parameters}) = \sum_{k=1}^{K} P(z_i = k)P(x_i \mid \mu_k, \Sigma_k)$$

Posterior: $P(z_i = k \mid x_i) \propto P(z_i = k)P(x_i \mid \mu_k, \Sigma_k)$
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
Laurent Charlin — 80-629

Initial cluster centers

(a) $L = 1$

(b) $L = 2$

(c) $L = 5$

(d) $L = 20$

Final solution

(k-means Final solution)

[Figure 9.8 from PRML]
K-means vs GMMs

Similar

Similar

GMM better

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)
Unsupervised Learning
Beyond Clustering
Different tasks

- Finding patterns
  - Clustering $f : X \to \{1, 2, \ldots, K\}$ (K clusters)
  - Dimensionality reduction $f : X^p \to X^k$, $k << p$
  - Density modelling $f : X \to [0, 1]$
- ...
Different tasks

- Finding patterns

- Clustering \( f : X \rightarrow \{1, 2, \ldots, K\} \) (K clusters)

- Dimensionality reduction \( f : X^p \rightarrow X^k, k << p \)

- Density modelling \( f : X \rightarrow [0, 1] \)

- ...
Autoencoders

- A type of neural network for dimensionality reduction
- Non-linear PCA
Autoencoders

- A type of neural network for dimensionality reduction
  - Non-linear PCA
  - Intuition: let’s learn to copy the data

\[ x = f(x) \]
Autoencoders

• A type of neural network for dimensionality reduction

• Non-linear PCA

• Intuition: let’s learn to copy the data

\[ x = f(x) \]

• We force a “bottleneck”

\[ z = f_1(x) \]
\[ x = f_2(z) \]
\[ \dim(z) < \dim(x) \]
Autoencoders

- A neural network architecture for unsupervised learning
Autoencoders

- A neural network architecture for unsupervised learning
Autoencoders

- A neural network architecture for unsupervised learning
Autoencoders

- A neural network architecture for unsupervised learning

Objective:
How well the network predicts X?

\[
\text{Loss} := \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \\
= \sum_{i=1}^{N} (x_i - f_2(f_1(x)))^2
\]
Autoencoders

- A neural network architecture for unsupervised learning

**Objective:**
How well the network predicts $X$?

$$\text{Loss} := \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

$$= \sum_{i=1}^{N} (x_i - f_2(f_1(x)))^2$$
Unsupervised learning as supervised learning

• Examples:
  • Auto-encoders “predict” their inputs
  • Language models “predict” the next word
Unsupervised learning as supervised learning

- Examples:
  - Auto-encoders “predict” their inputs
  - Language models “predict” the next word

- Create a target from the x’s (or a subset)
  - Find a task to ensure that you learn something useful

- Self-supervised learning
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
  - Note that all our examples were in 2D

- Can be used in downstream tasks