CSC 411: Lecture 12: Clustering

Richard Zemel, Raquel Urtasun and Sanja Fidler

University of Toronto

Today

- Unsupervised learning
- Clustering
  - k-means
  - Soft k-means

Motivating Examples

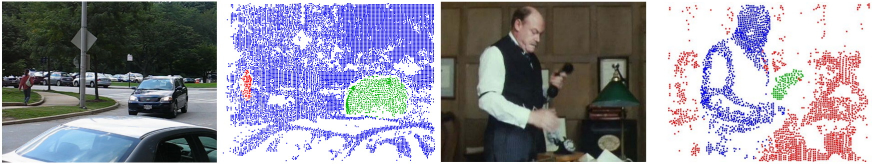
- Determine groups of people in image above

Motivating Examples

- Determine groups of people in image above
  - based on clothing styles
  - gender, age, etc

Motivating Examples

- Determine groups of people in image above
  - based on clothing styles
  - gender, age, etc



- Determine moving objects in videos

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:
  - Reduce dimensionality

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:
  - Reduce dimensionality
  - Find clusters

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:
  - Reduce dimensionality
  - Find clusters
  - Model data density

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:
  - Reduce dimensionality
  - Find clusters
  - Model data density
  - Find hidden causes

Unsupervised Learning

- Supervised learning algorithms have a clear goal: produce desired outputs for given inputs.
  - You are given $\{(x^{(i)}, t^{(i)})\}$ during training (inputs and targets)
- Goal of unsupervised learning algorithms (no explicit feedback whether outputs of system are correct) less clear.
  - You are given the inputs $\{x^{(i)}\}$ during training, labels are unknown.
- Tasks to consider:
  - Reduce dimensionality
  - Find clusters
  - Model data density
  - Find hidden causes
- Key utility
  - Compress data
  - Detect outliers
  - Facilitate other learning

Major Types

- Primary problems, approaches in unsupervised learning fall into three classes:

Major Types

- Primary problems, approaches in unsupervised learning fall into three classes:

  1. Dimensionality reduction: represent each input case using a small number of variables (e.g., principal components analysis, factor analysis, independent components analysis)

Major Types

- Primary problems, approaches in unsupervised learning fall into three classes:

  1. Dimensionality reduction: represent each input case using a small number of variables (e.g., principal components analysis, factor analysis, independent components analysis)
  2. Clustering: represent each input case using a prototype example (e.g., k-means, mixture models)

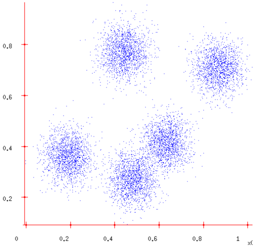
Major Types

- Primary problems, approaches in unsupervised learning fall into three classes:

  1. Dimensionality reduction: represent each input case using a small number of variables (e.g., principal components analysis, factor analysis, independent components analysis)
  2. Clustering: represent each input case using a prototype example (e.g., k-means, mixture models)
  3. Density estimation: estimating the probability distribution over the data space

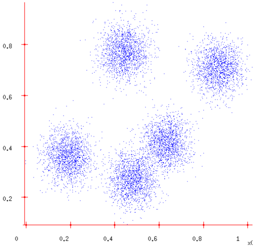
Clustering

- Grouping N examples into K clusters one of canonical problems in unsupervised learning



Clustering

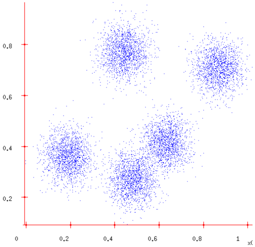
- Grouping N examples into K clusters one of canonical problems in unsupervised learning



- Motivation: prediction; lossy compression; outlier detection

Clustering

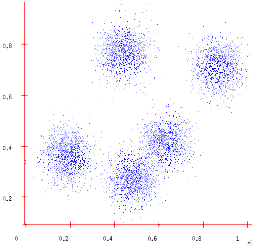
- Grouping N examples into K clusters one of canonical problems in unsupervised learning



- Motivation: prediction; lossy compression; outlier detection
- We assume that the data was generated from a number of different classes. The aim is to cluster data from the same class together.
  - How many classes?
  - Why not put each datapoint into a separate class?

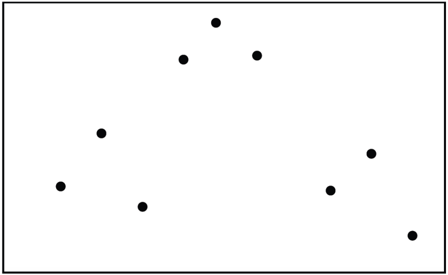
Clustering

- Grouping N examples into K clusters one of canonical problems in unsupervised learning



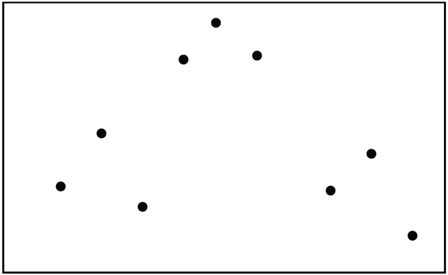
- Motivation: prediction; lossy compression; outlier detection
- We assume that the data was generated from a number of different classes. The aim is to cluster data from the same class together.
  - How many classes?
  - Why not put each datapoint into a separate class?
- What is the objective function that is optimized by sensible clustering?

Clustering



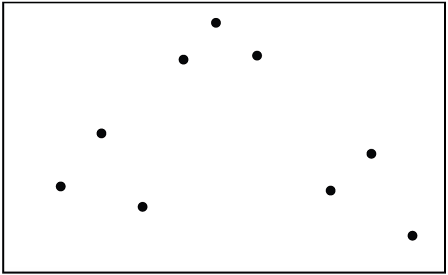
- Assume the data $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ lives in a Euclidean space, $\mathbf{x}^{(n)} \in \mathbb{R}^d$.

Clustering



- Assume the data $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ lives in a Euclidean space, $\mathbf{x}^{(n)} \in \mathbb{R}^d$.
- Assume the data belongs to K classes (patterns)

Clustering



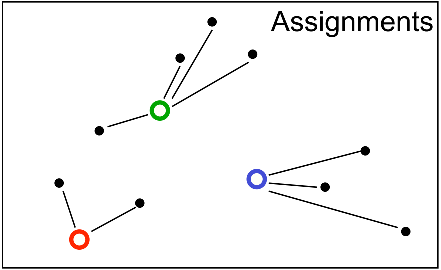
- Assume the data $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ lives in a Euclidean space, $\mathbf{x}^{(n)} \in \mathbb{R}^d$.
- Assume the data belongs to K classes (patterns)
- How can we identify those classes (data points that belong to each class)?

K-means

- Initialization: randomly initialize cluster centers

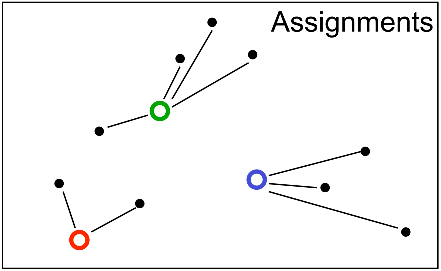
K-means

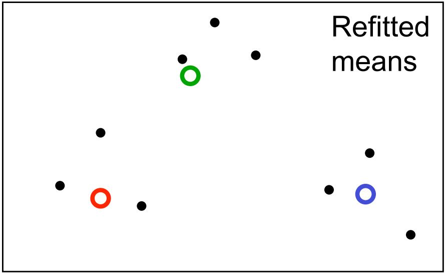
- Initialization: randomly initialize cluster centers
- The algorithm iteratively alternates between two steps:
  - Assignment step: Assign each data point to the closest cluster



K-means

- Initialization: randomly initialize cluster centers
- The algorithm iteratively alternates between two steps:
  - Assignment step: Assign each data point to the closest cluster
  - Refitting step: Move each cluster center to the center of gravity of the data assigned to it





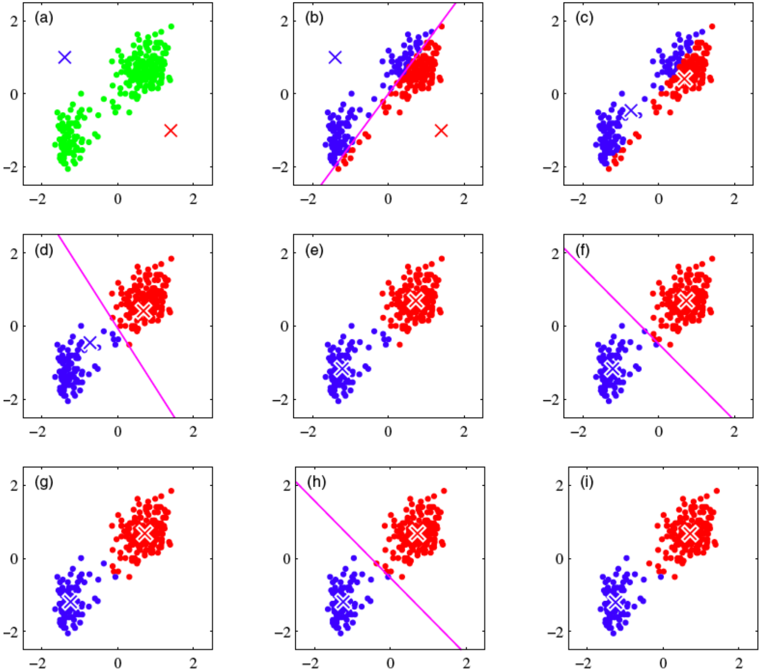


Figure from Bishop

Simple demo: http://syskall.com/kmeans.js/

K-means Objective

What is actually being optimized?

K-means Objective

What is actually being optimized?

K-means Objective:
Find cluster centers $\mathbf{m}$ and assignments $\mathbf{r}$ to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}$ to their assigned cluster centers

$$\min_{\{\mathbf{m}\},\{\mathbf{r}\}} J(\{\mathbf{m}\},\{\mathbf{r}\}) = \min_{\{\mathbf{m}\},\{\mathbf{r}\}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

$$\text{s.t.} \sum_k r_k^{(n)} = 1, \forall n, \quad \text{where} \quad r_k^{(n)} \in \{0,1\}, \forall k, n$$

where $r_k^{(n)} = 1$ means that $\mathbf{x}^{(n)}$ is assigned to cluster k (with center $\mathbf{m}_k$)

K-means Objective

What is actually being optimized?

K-means Objective:
Find cluster centers **m** and assignments **r** to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}$ to their assigned cluster centers

$$\min_{\{\mathbf{m}\},\{\mathbf{r}\}} J(\{\mathbf{m}\},\{\mathbf{r}\}) = \min_{\{\mathbf{m}\},\{\mathbf{r}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

$$\text{s.t.} \sum_k r_k^{(n)} = 1, \forall n, \quad \text{where} \quad r_k^{(n)} \in \{0,1\}, \forall k, n$$

where $r_k^{(n)} = 1$ means that $\mathbf{x}^{(n)}$ is assigned to cluster k (with center $\mathbf{m}_k$)

- Optimization method is a form of coordinate descent ("block coordinate descent")

K-means Objective

What is actually being optimized?

K-means Objective:
Find cluster centers $\mathbf{m}$ and assignments $\mathbf{r}$ to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}$ to their assigned cluster centers

$$\min_{\{\mathbf{m}\},\{\mathbf{r}\}} J(\{\mathbf{m}\},\{\mathbf{r}\}) = \min_{\{\mathbf{m}\},\{\mathbf{r}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

$$\text{s.t.} \sum_k r_k^{(n)} = 1, \forall n, \quad \text{where} \quad r_k^{(n)} \in \{0,1\}, \forall k, n$$

where $r_k^{(n)} = 1$ means that $\mathbf{x}^{(n)}$ is assigned to cluster k (with center $\mathbf{m}_k$)

- Optimization method is a form of coordinate descent ("block coordinate descent")
  - Fix centers, optimize assignments (choose cluster whose mean is closest)

K-means Objective

What is actually being optimized?

K-means Objective:
Find cluster centers $\mathbf{m}$ and assignments $\mathbf{r}$ to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}$ to their assigned cluster centers

$$\min_{\{\mathbf{m}\},\{\mathbf{r}\}} J(\{\mathbf{m}\},\{\mathbf{r}\}) = \min_{\{\mathbf{m}\},\{\mathbf{r}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

$$\text{s.t.} \sum_k r_k^{(n)} = 1, \forall n, \quad \text{where} \quad r_k^{(n)} \in \{0,1\}, \forall k, n$$

where $r_k^{(n)} = 1$ means that $\mathbf{x}^{(n)}$ is assigned to cluster k (with center $\mathbf{m}_k$)

- Optimization method is a form of coordinate descent ("block coordinate descent")
  - Fix centers, optimize assignments (choose cluster whose mean is closest)
  - Fix assignments, optimize means (average of assigned datapoints)

The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values

The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values
- Repeat until convergence (until assignments do not change):

The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values
- Repeat until convergence (until assignments do not change):
  - Assignment: Each data point $\mathbf{x}^{(n)}$ assigned to nearest mean

$$\hat{k}^n = arg\min_k d(\mathbf{m}_k, \mathbf{x}^{(n)})$$

(with, for example, L2 norm: $\hat{k}^n = arg\min_k ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$)

The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values
- Repeat until convergence (until assignments do not change):
  - Assignment: Each data point $\mathbf{x}^{(n)}$ assigned to nearest mean

$$\hat{k}^n = arg\min_k d(\mathbf{m}_k, \mathbf{x}^{(n)})$$

(with, for example, L2 norm: $\hat{k}^n = arg\min_k ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$)

and Responsibilities (1 of k encoding)

$$r_k^{(n)} = 1 \longleftrightarrow \hat{k}^{(n)} = k$$

The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values
- Repeat until convergence (until assignments do not change):
  - Assignment: Each data point $\mathbf{x}^{(n)}$ assigned to nearest mean

$$\hat{k}^n = arg \min_k d(\mathbf{m}_k, \mathbf{x}^{(n)})$$

(with, for example, L2 norm: $\hat{k}^n = arg \min_k ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$)

and Responsibilities (1 of k encoding)

$$r_k^{(n)} = 1 \longleftrightarrow \hat{k}^{(n)} = k$$

  - Update: Model parameters, means are adjusted to match sample means of data points they are responsible for:

$$\mathbf{m}_k = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{\sum_n r_k^{(n)}}$$

K-means for Vector Quantization

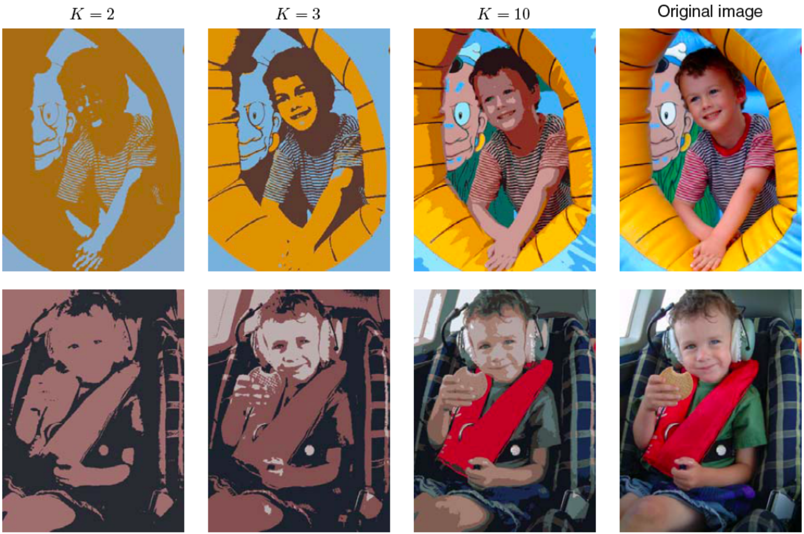
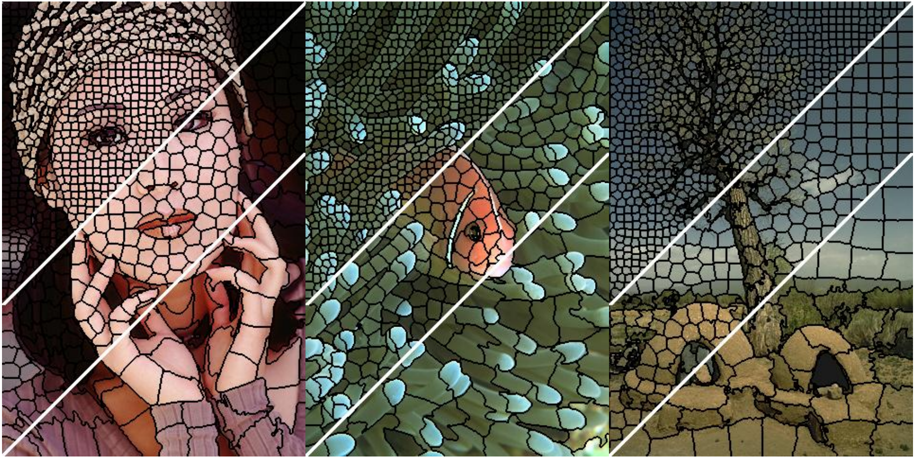


Figure from Bishop

K-means for Image Segmentation



- How would you modify k-means to get super pixels?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?
- How can we choose best distance?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?
- How can we choose best distance?
- How to choose K?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?
- How can we choose best distance?
- How to choose K?
- How can we choose between alternative clusterings?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?
- How can we choose best distance?
- How to choose K?
- How can we choose between alternative clusterings?
- Will it converge?

Questions about K-means

- Why does update set $\mathbf{m}_k$ to mean of assigned points?
- Where does distance d come from?
- What if we used a different distance measure?
- How can we choose best distance?
- How to choose K?
- How can we choose between alternative clusterings?
- Will it converge?

Hard cases – unequal spreads, non-circular spreads, in-between points

Why K-means Converges

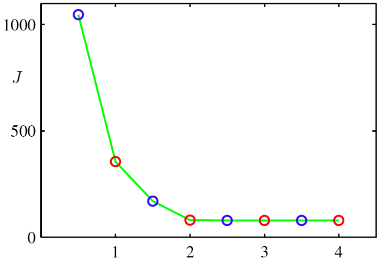
- Whenever an assignment is changed, the sum squared distances J of data points from their assigned cluster centers is reduced.

Why K-means Converges

- Whenever an assignment is changed, the sum squared distances J of data points from their assigned cluster centers is reduced.
- Whenever a cluster center is moved, J is reduced.

Why K-means Converges

- Whenever an assignment is changed, the sum squared distances J of data points from their assigned cluster centers is reduced.
- Whenever a cluster center is moved, J is reduced.
- Test for convergence: If the assignments do not change in the assignment step, we have converged (to at least a local minimum).

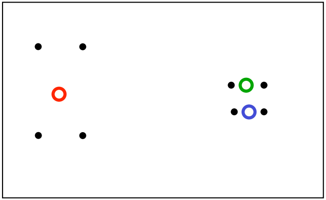


- K-means cost function after each E step (blue) and M step (red). The algorithm has converged after the third M step

Local Minima

- The objective J is non-convex (so coordinate descent on J is not guaranteed to converge to the global minimum)
- There is nothing to prevent k-means getting stuck at local minima.
- We could try many random starting points
- We could try non-local split-and-merge moves:
    - Simultaneously merge two nearby clusters
    - and split a big cluster into two

A bad local optimum



Soft K-means

- Instead of making hard assignments of data points to clusters, we can make soft assignments. One cluster may have a responsibility of .7 for a datapoint and another may have a responsibility of .3.

Soft K-means

- Instead of making hard assignments of data points to clusters, we can make soft assignments. One cluster may have a responsibility of .7 for a datapoint and another may have a responsibility of .3.
  - Allows a cluster to use more information about the data in the refitting step.

Soft K-means

- Instead of making hard assignments of data points to clusters, we can make soft assignments. One cluster may have a responsibility of .7 for a datapoint and another may have a responsibility of .3.
  - Allows a cluster to use more information about the data in the refitting step.
  - What happens to our convergence guarantee?

Soft K-means

- Instead of making hard assignments of data points to clusters, we can make soft assignments. One cluster may have a responsibility of .7 for a datapoint and another may have a responsibility of .3.
  - Allows a cluster to use more information about the data in the refitting step.
  - What happens to our convergence guarantee?
  - How do we decide on the soft assignments?

Soft K-means Algorithm

- Initialization: Set K means $\{\mathbf{m}_k\}$ to random values
- Repeat until convergence (until assignments do not change):
  - Assignment: Each data point n given soft "degree of assignment" to each cluster mean k, based on responsibilities

$$r_k^{(n)} = \frac{\exp[-\beta d(\mathbf{m}_k, \mathbf{x}^{(n)})]}{\sum_j \exp[-\beta d(\mathbf{m}_j, \mathbf{x}^{(n)})]}$$

  - Update: Model parameters, means, are adjusted to match sample means of datapoints they are responsible for:

$$\mathbf{m}_k = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{\sum_n r_k^{(n)}}$$

Questions about Soft K-means

- How to set β?

Questions about Soft K-means

- How to set β?
- What about problems with elongated clusters?

Questions about Soft K-means

- How to set β?
- What about problems with elongated clusters?
- Clusters with unequal weight and width

A Generative View of Clustering

- We need a sensible measure of what it means to cluster the data well.

A Generative View of Clustering

- We need a sensible measure of what it means to cluster the data well.
  - This makes it possible to judge different models.

A Generative View of Clustering

- We need a sensible measure of what it means to cluster the data well.
  - This makes it possible to judge different models.
  - It may make it possible to decide on the number of clusters.

A Generative View of Clustering

- We need a sensible measure of what it means to cluster the data well.
  - This makes it possible to judge different models.
  - It may make it possible to decide on the number of clusters.
- An obvious approach is to imagine that the data was produced by a generative model.

A Generative View of Clustering

- We need a sensible measure of what it means to cluster the data well.
    - This makes it possible to judge different models.
    - It may make it possible to decide on the number of clusters.
- An obvious approach is to imagine that the data was produced by a generative model.
    - Then we can adjust the parameters of the model to maximize the probability that it would produce exactly the data we observed.