349:Machine Learning

Fall 2024

k-Means Clustering

Reasons for k-Means Clustering

- Unsupervised ML task:

  Labels are ***expensive!***

- Common use cases

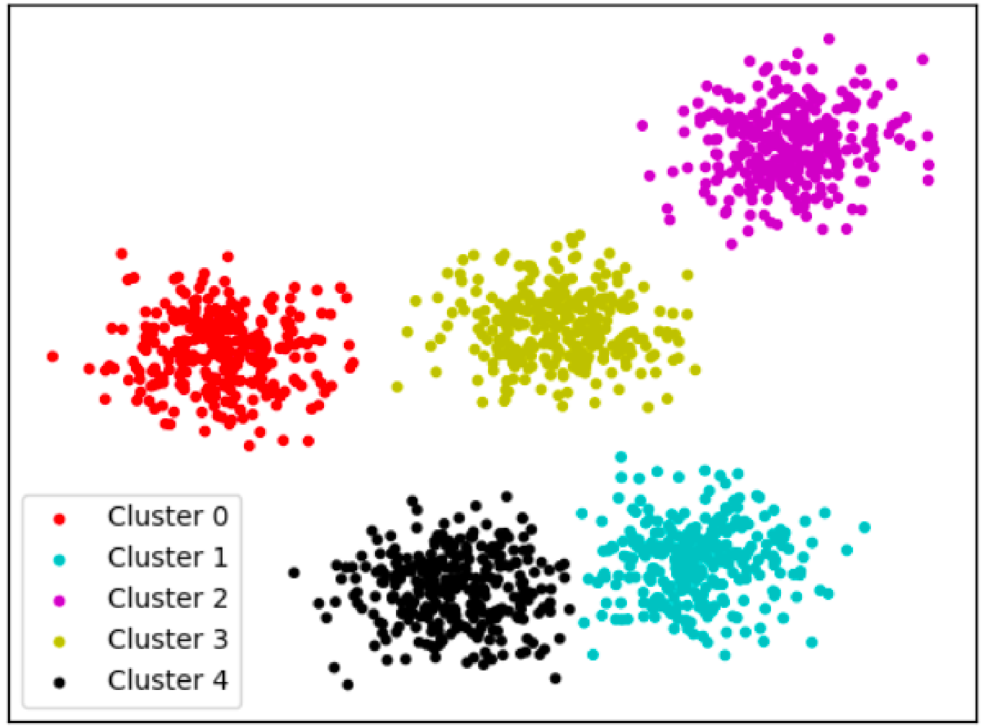
  Social network analysis

  Anomaly detection

  Media recommendation

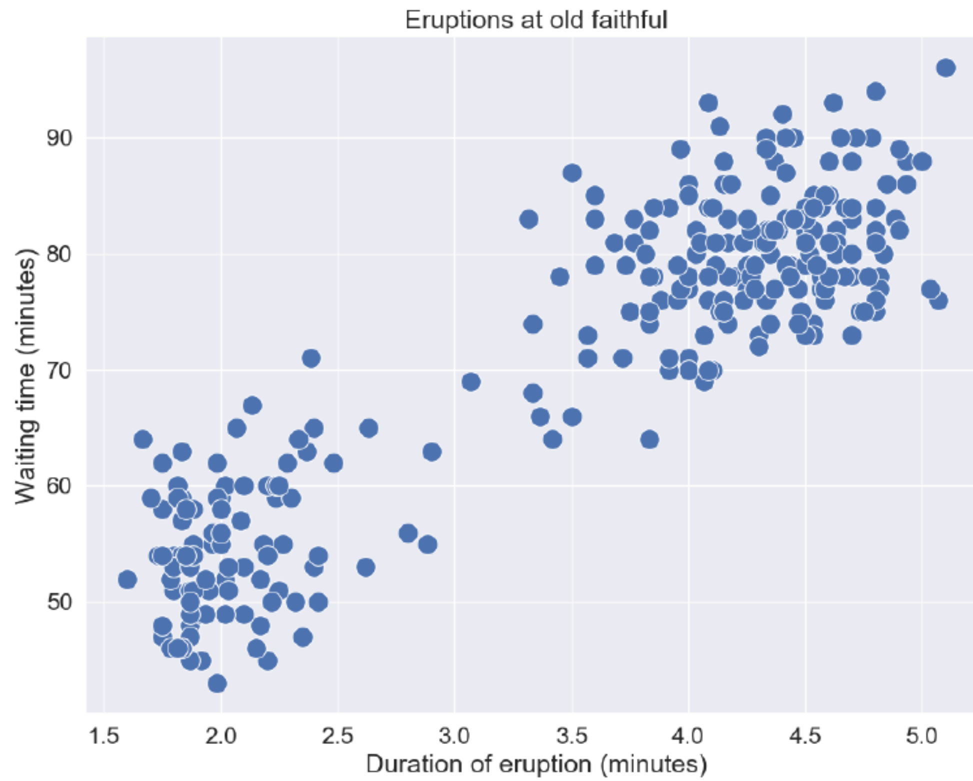
  Speaker recognition

  Many more …



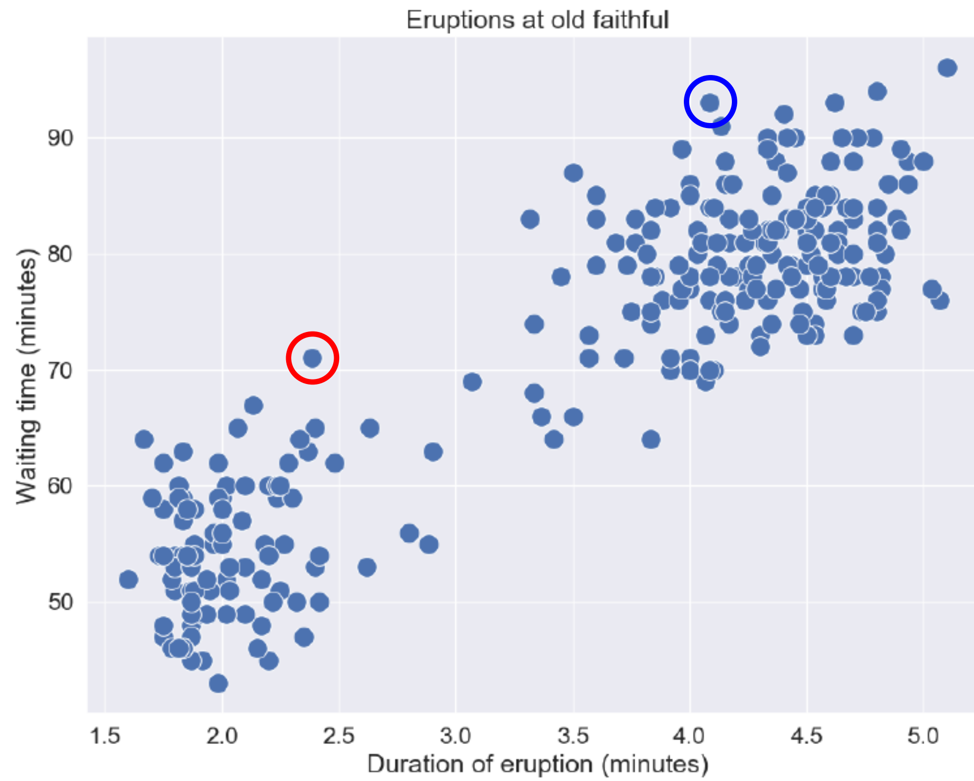
Old Faithful Example

- Raw data from Old Faithful
  - Captures some phenomena
  - How could you possibly use this?



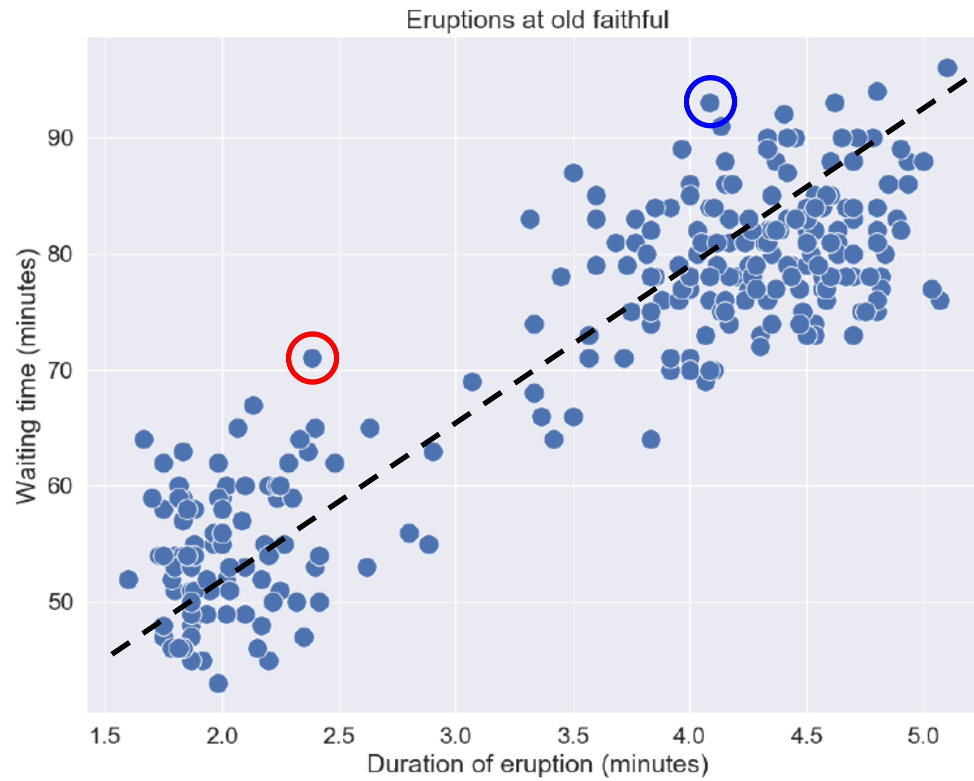
Old Faithful Example

- Raw data from Old Faithful
    - Captures some phenomena
    - How could you possibly use this?
    - What is the underlying relationship?



Old Faithful Example

- Raw data from Old Faithful
  - Captures some phenomena
  - How could you possibly use this?
  - What is the underlying relationship?



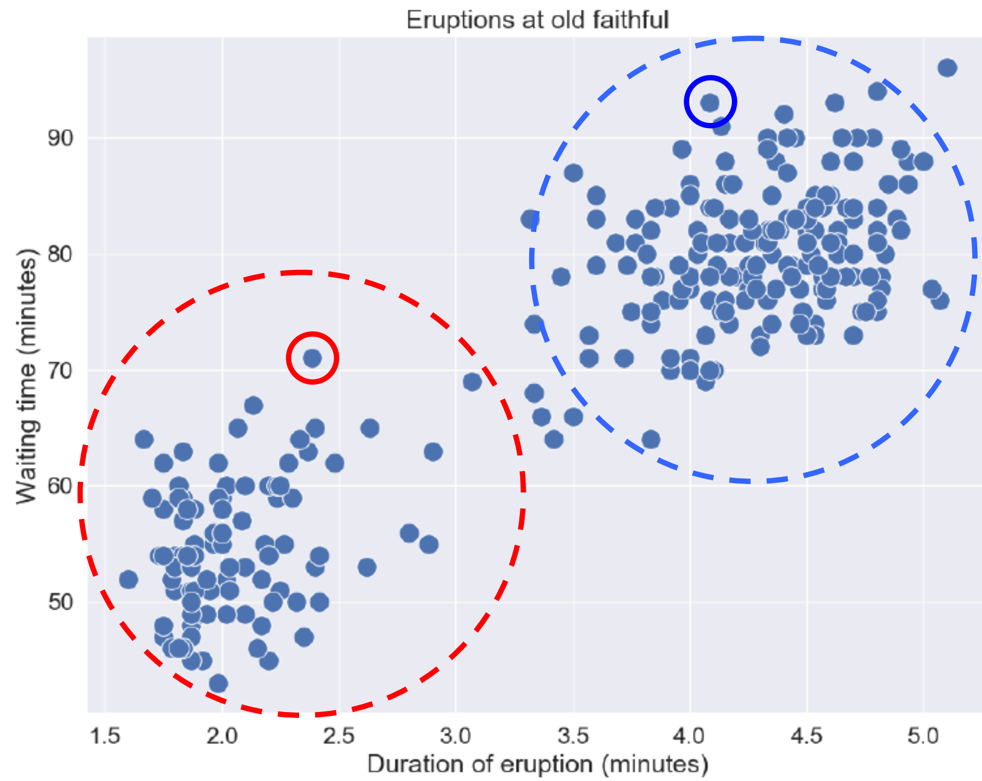
Old Faithful Example

- Raw data from Old Faithful

    Captures some phenomena

    How could you possibly use this?

    What is the underlying relationship?



Old Faithful Example

- Raw data from Old Faithful

  Captures some phenomena

  How could you possibly use this?

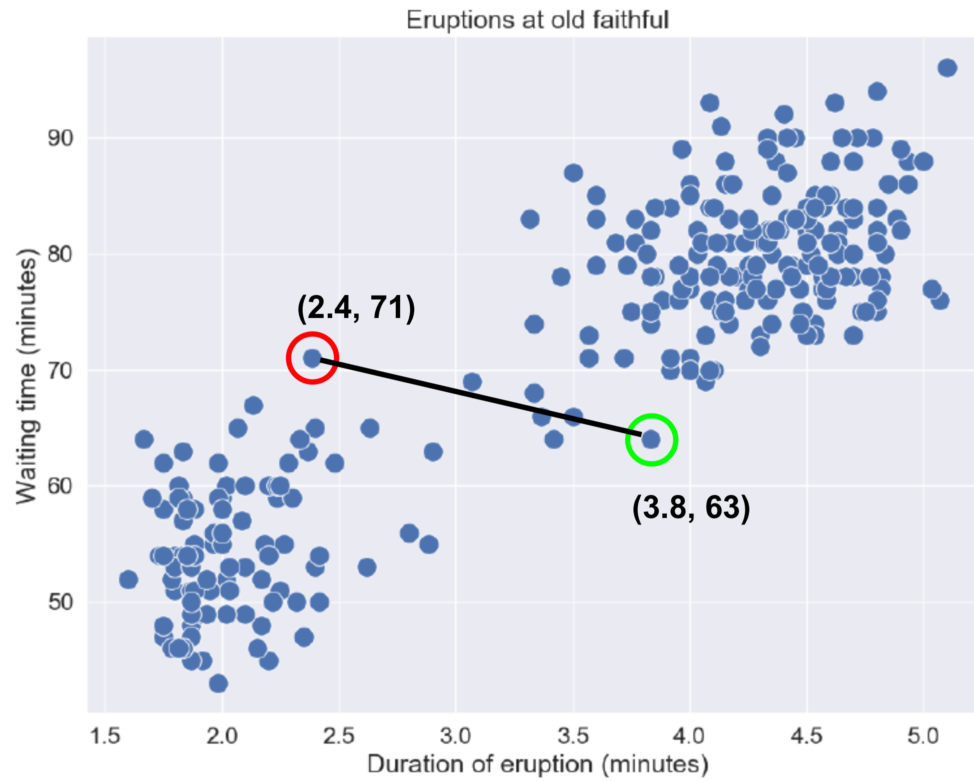
  What is the underlying relationship?

- Distance Metric can be used to measure pair-wise relations

$$d(p_i, p_j) = \sqrt{\sum_A (p_{i,a} - p_{j,a})^2}$$

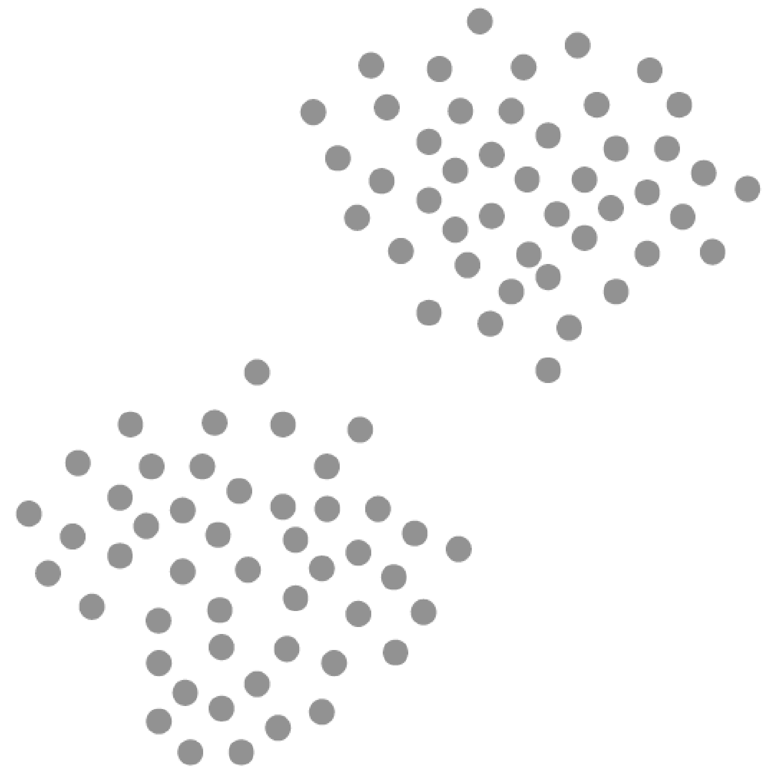
$$= \sqrt{(71-63)^2 + (2.4-3.8)^2}$$

$$= 8.13$$



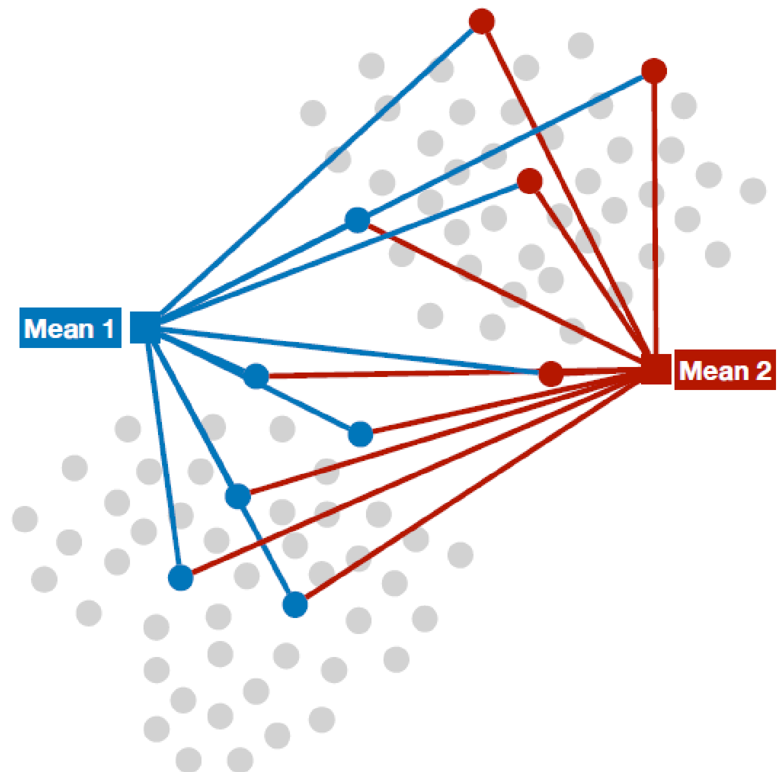
k-Means Algorithm

- We start with unlabeled data



k-Means Algorithm

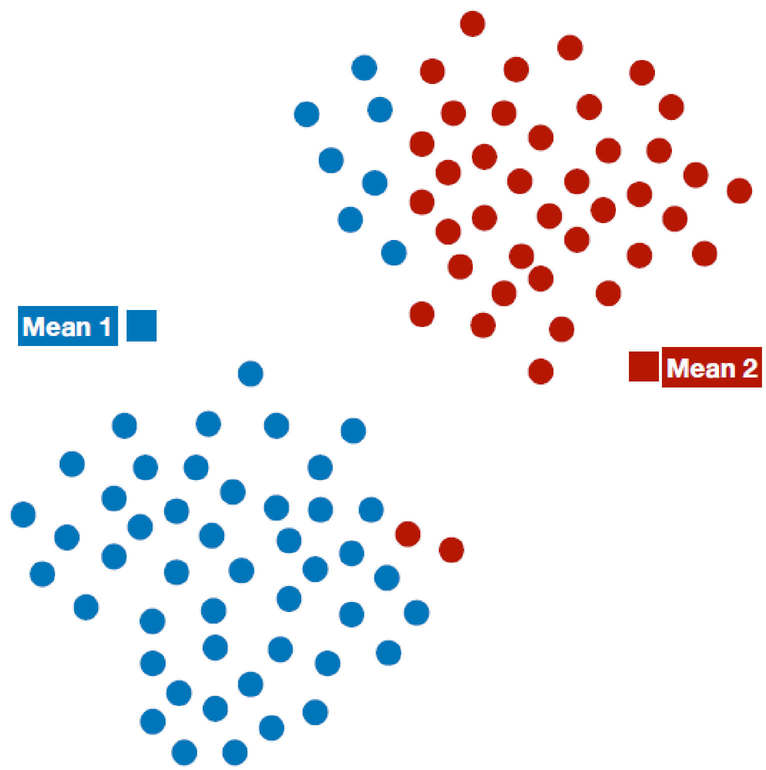
- We start with unlabeled data
    1. Randomly select means
    2. Calculate distance to means for every data point
    3. Assign class labels based upon shortest distance
    4. Update means and repeat



k-Means Algorithm

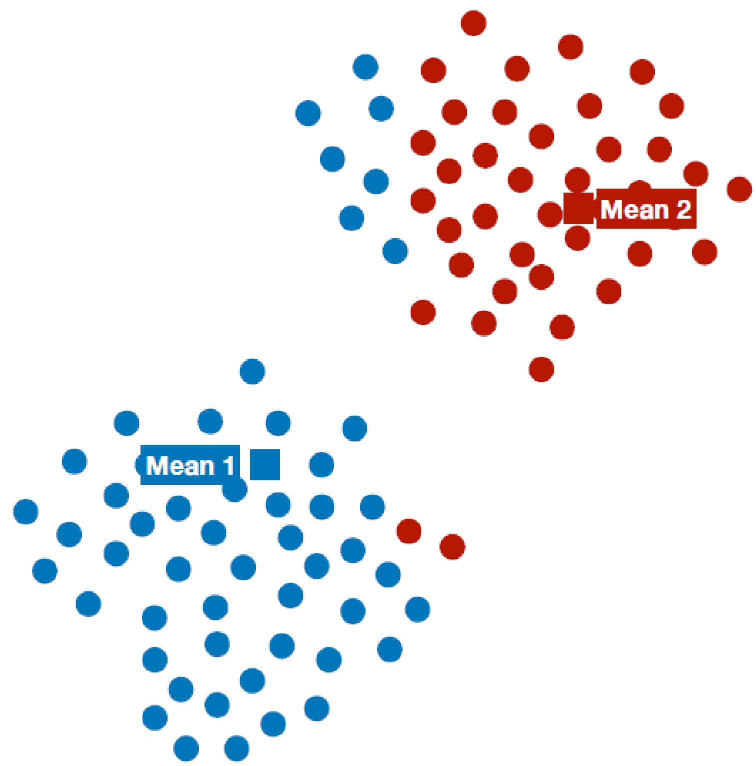
- We start with unlabeled data

  1. Randomly select means
  2. Calculate distance to means for every data point
  3. Assign class labels based upon shortest distance
  4. Update means and repeat



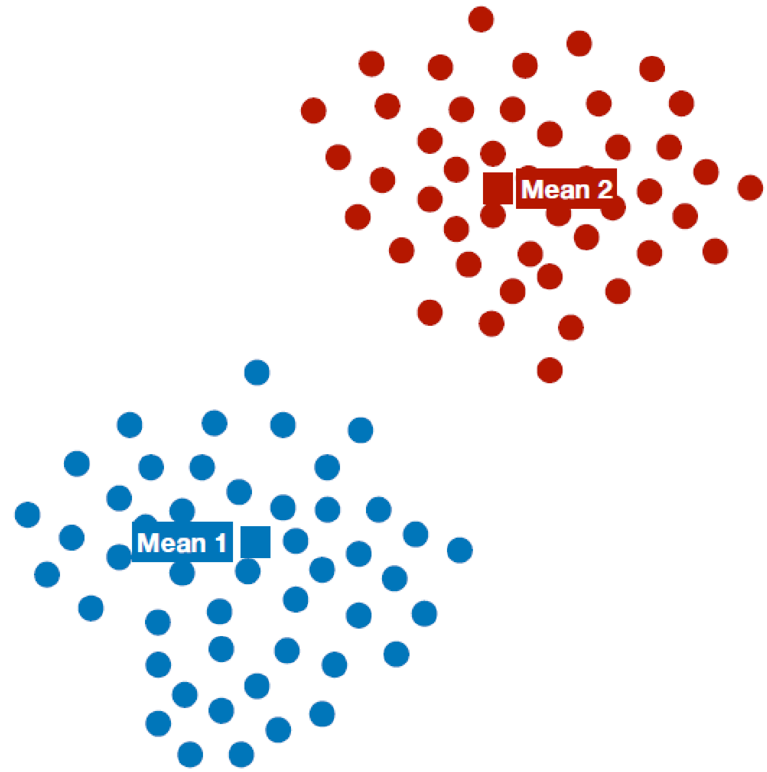
k-Means Algorithm

- We start with unlabeled data
    1. Randomly select means
    2. Calculate distance to means for every data point
    3. Assign class labels based upon shortest distance
    4. Update means and repeat



k-Means Algorithm

- We start with unlabeled data
  1. Randomly select means
  2. Calculate distance to means for every data point
  3. Assign class labels based upon shortest distance
  4. Update means and repeat until ***convergence***



k-Means Algorithm

- Assign class labels based upon distance to the means

$$c_i \equiv \underset{m}{\arg\min} \left\| x_i - \mu_m \right\|_2$$

- Update the means:

$$\mu_{m,a} \equiv \frac{\sum_i \mathbf{1}_{c \in m} \cdot x_{i,a}}{\sum_i \mathbf{1}_{c \in m}}$$

where:

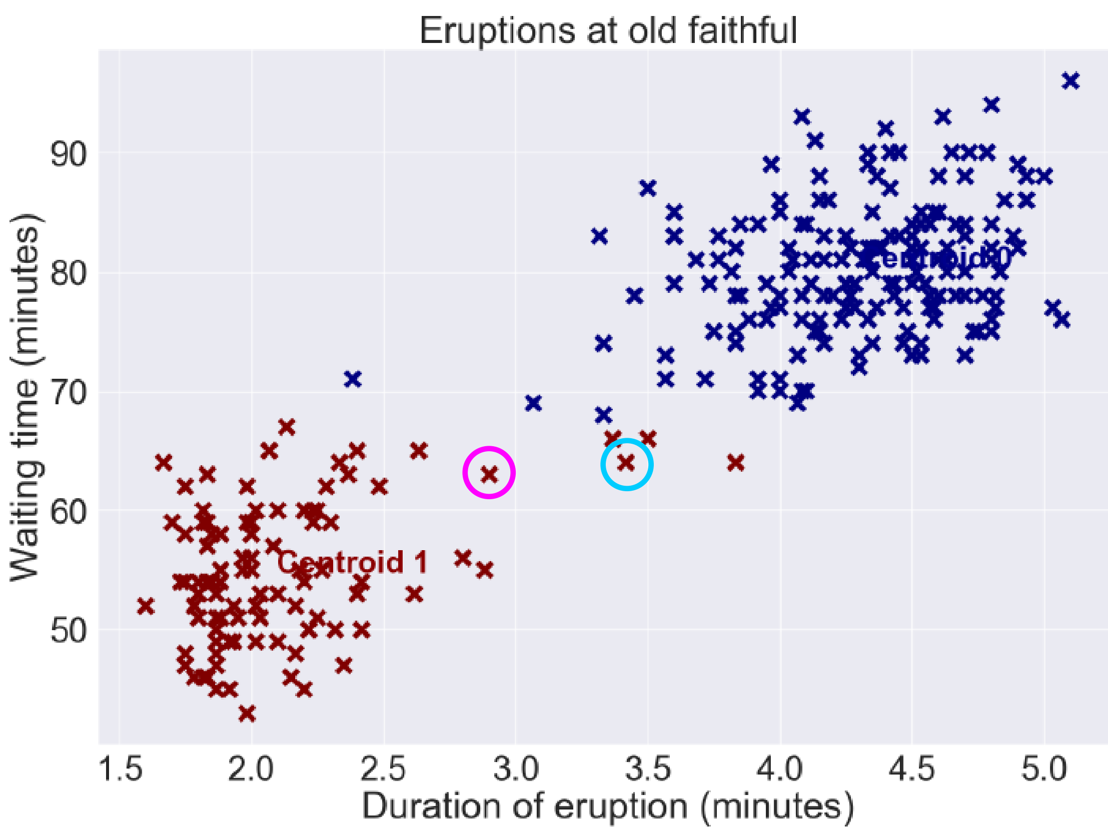
a is the attribute, m is the class label, and

$\mu = (a_1, a_{2,} \ldots a_A)$ is a vector in the attribute-space

Back to Old Faithful Example

- The k-Means algorithm classifies Old Faithful data into two clusters:

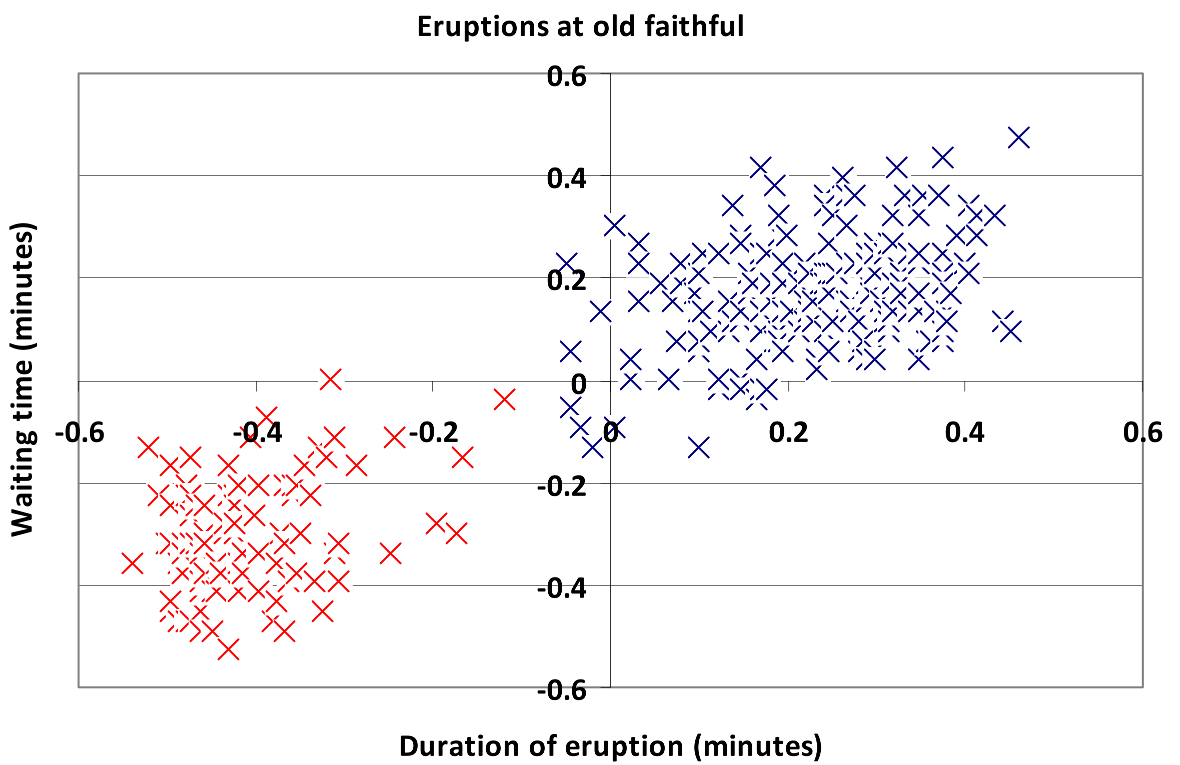
  But is this correct?



Old Faithful Example -- Revisited

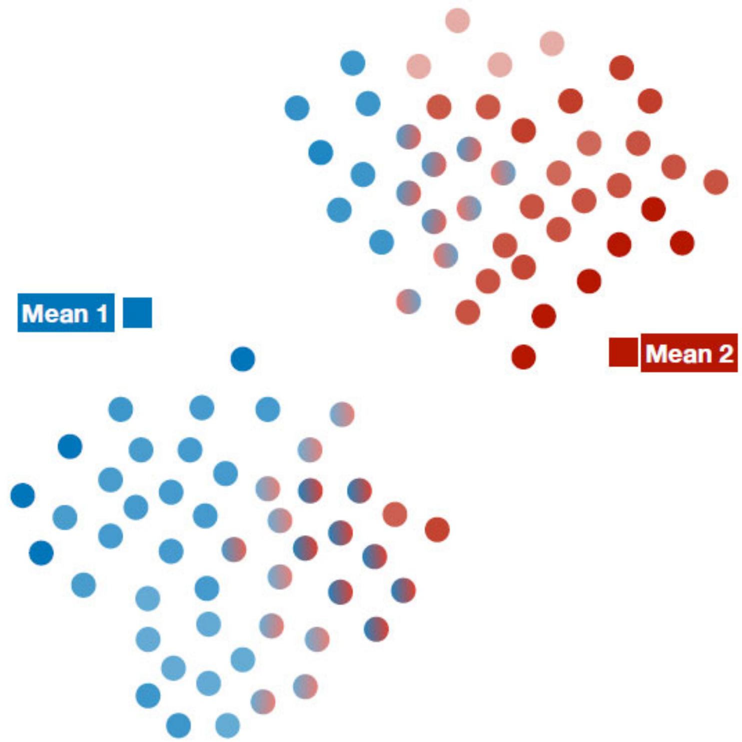
- Normalizing data allows but attributes to be considered

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$



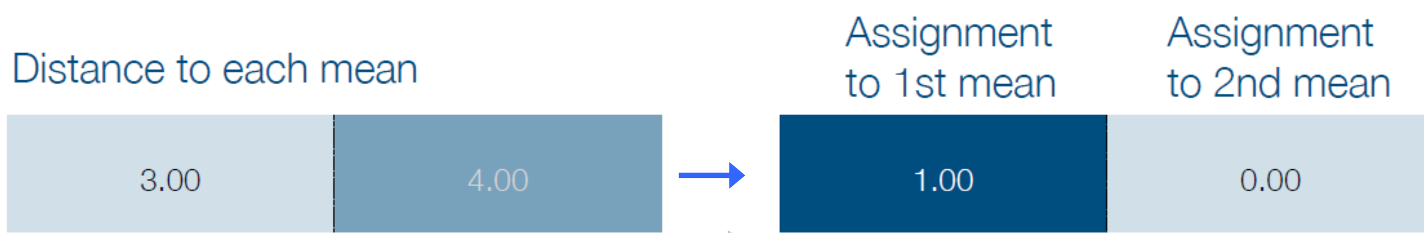
Soft k-Means Algorithm

- Soft decision is based on how much closer an observation is to one means versus others.

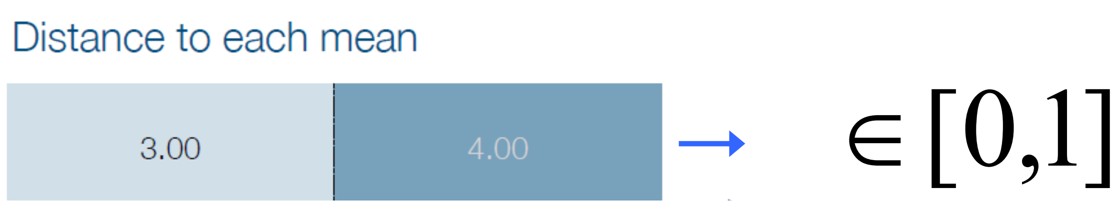


Soft k-Means Algorithm

- A hard-margin classifier assigns an ***atomic*** class label



- A soft-margin classifier will assign a probability



Soft k-Means Algorithm

- Probability is assigned to each class using the softmax function

$$P(x_i \in m) = \text{softmax}(z_{im}) = \frac{e^{z_{im}}}{\sum_m e^{z_{im}}}$$

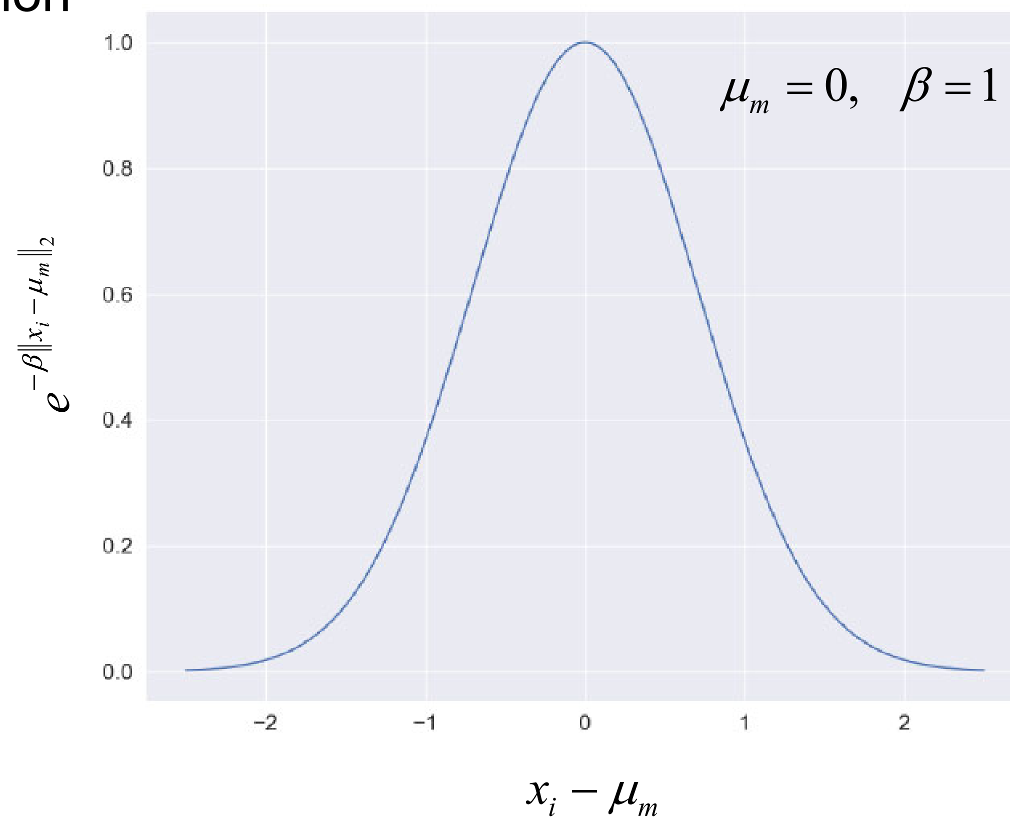
where:

$$z_{im} = -\beta \left\| x_i - \mu_m \right\|_2$$

and β controls the ***sharpness*** of the distribution

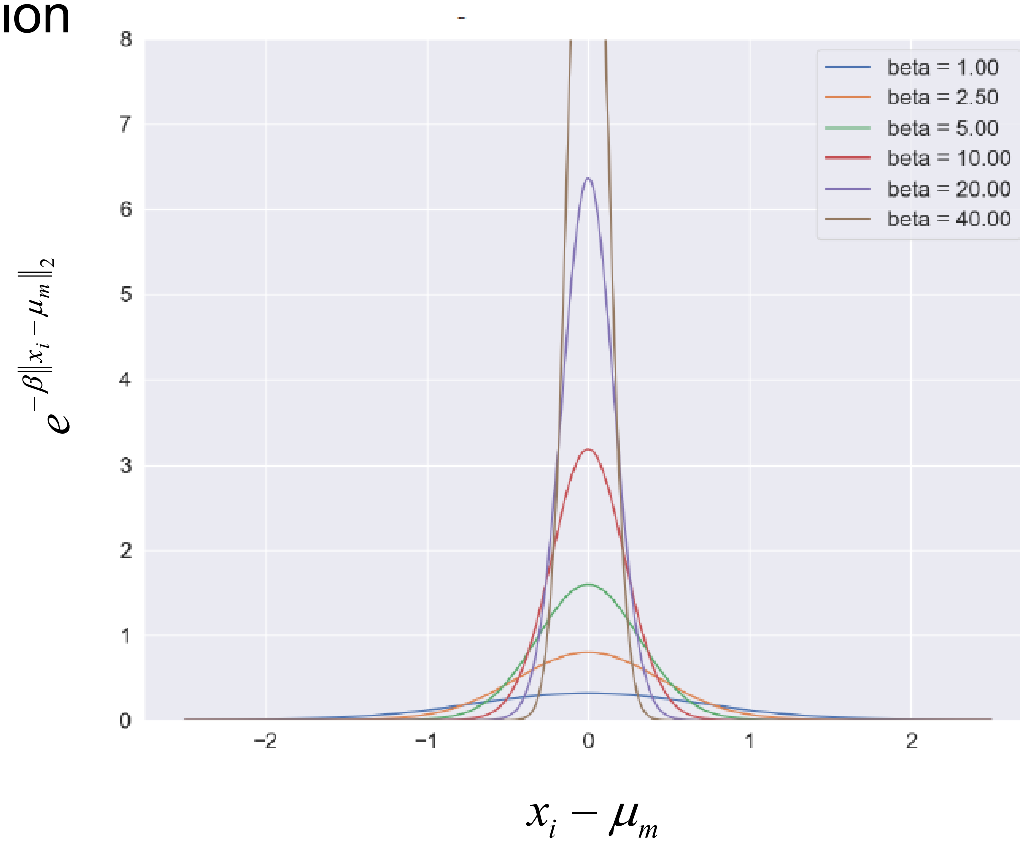
Soft k-Means Algorithm

- Resulting probability distribution is a Gaussian function of the distance between an observation and a mean



Soft k-Means Algorithm

- Resulting probability distribution is a Gaussian function of the distance between an observation and a mean
- β is sometimes called the ***temperature***



Soft k-Means Algorithm

- Applying this to the previous example with $\beta = 1$

Distance to each mean

3.00	4.00

$$P(x_i \in m_1) = \frac{e^{-3}}{e^{-3} + e^{-4}}$$

Distance to each mean

3.00	4.00

Assignment to 1st mean	Assignment to 2nd mean
0.731	0.269

and for $\beta = 2$

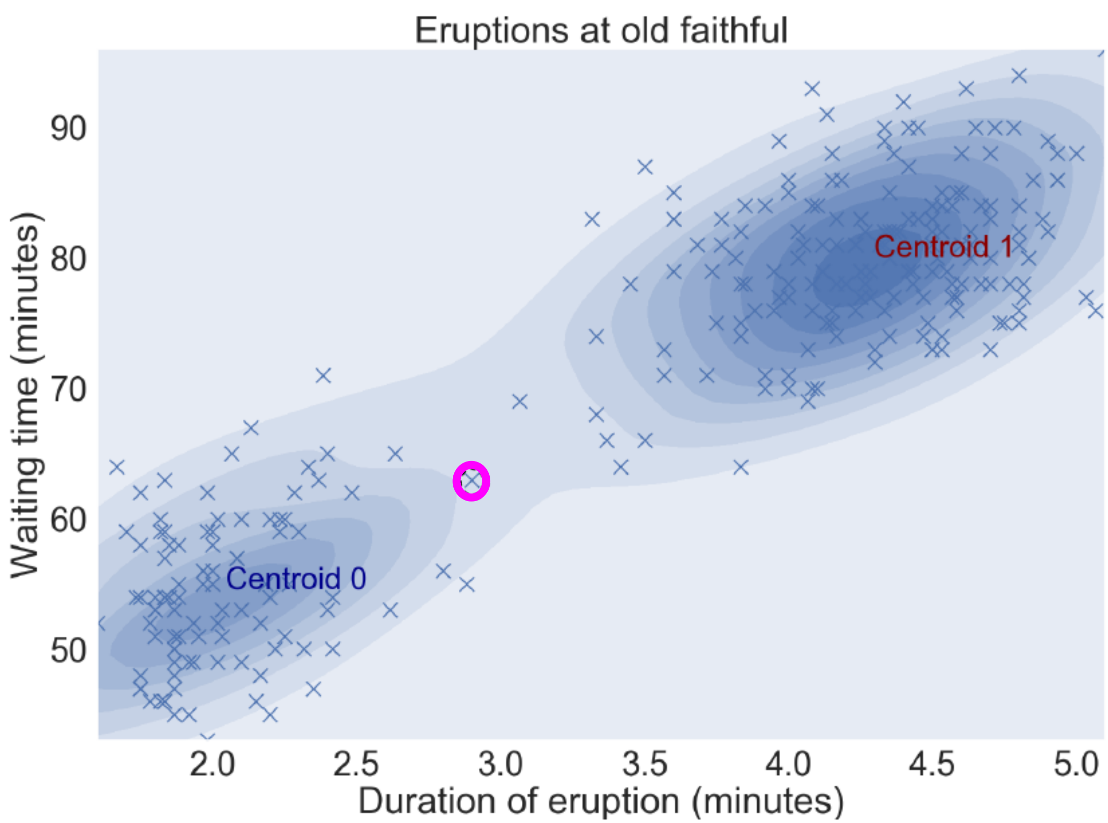
Distance to each mean

3.00	4.00

Assignment to 1st mean	Assignment to 2nd mean
0.881	0.119

Soft k-Means Algorithm

- Applying Soft k-Means to Old Faithful resolves ambiguity



Take Aways

- We learned how to classify unlabeled data k-Means clustering
- We formulated a Soft k-Means to handle uncertainty (and express confidence)

Precision vs Recall

Classifiers are often evaluated with an eye towards their being search engines. (e.g. labeling documents as either relevant or not to a search query). In this case people often use the following measures:

$$precision \qquad p = \frac{tp}{tp + fp}$$

$$recall \qquad r = \frac{tp}{tp + fn}$$

$$F - measure \qquad F = 2\frac{p \cdot r}{p + r}$$

True Classification

Machine's Classification	True	False
True	**True positive (tp)**	**False positive (fp)**
False	**False negative (fn)**	**True negative (tn)**

Confusion Matrix

Lets us see which things the classifier is mixing up. Helps direct improvement.

Correct Classification

Machine's Classification

	Dog	**Coyote**	**Cactus**	**Road Runner**
Dog	8	5	0	2
Coyote	2	5	0	2
Cactus	0	0	8	2
Road Runner	0	0	2	4