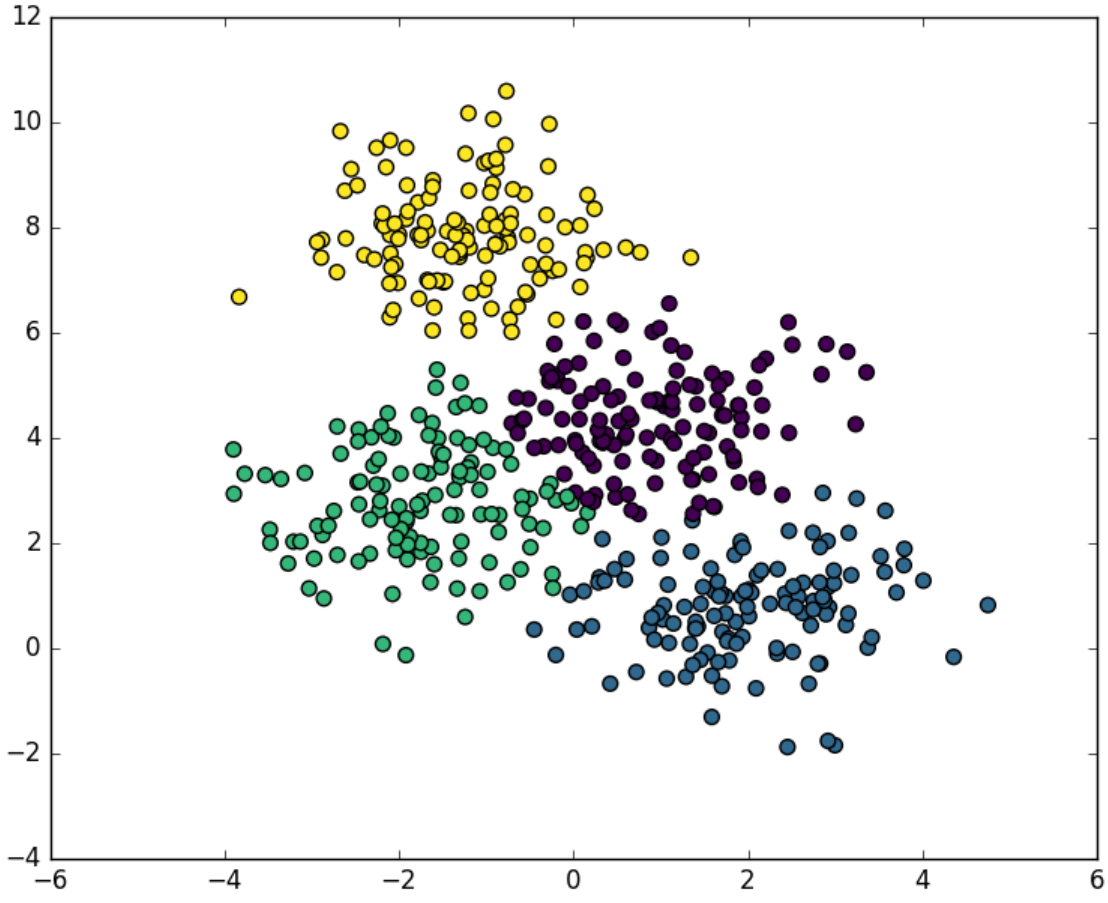
Overview

- Soft k-means
- Soft k-means: probabilistic interpretation
- Gaussian Mixture Model

Toy Example



Soft k-means

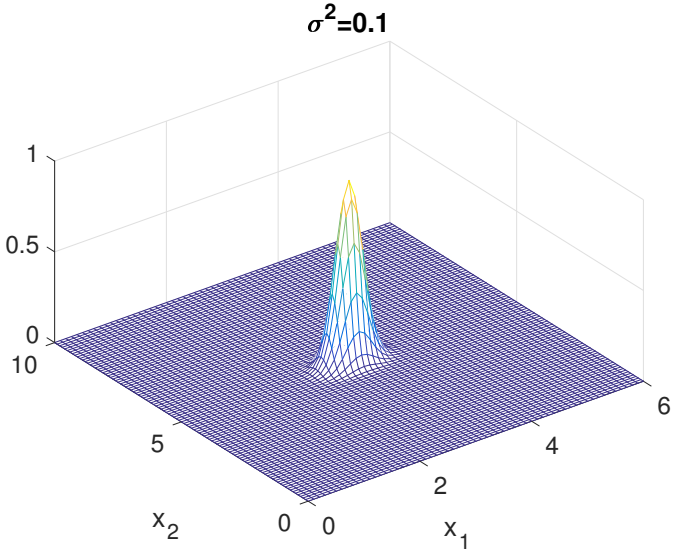
- In the k-means setup we assign a single class label $c^{(i)}$ to point $x^{(i)}$ i.e. make a "hard" assignment to a class.
- We can we extend k-means to estimate the probability that point $x^{(i)}$ is in class j i.e. make a "soft" assignment to a class.

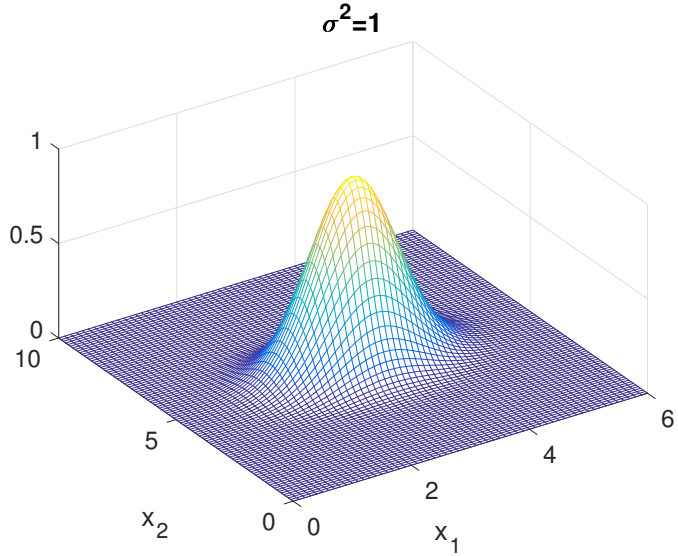
Gaussian kernel

Recall Gaussian kernel $K(x,z) = e^{\frac{\sum_{j=1}^{n}(x_j - z_j)^2}{\sigma^2}} = e^{\frac{\|x-z\|^2}{\sigma^2}}$, where for vectors x and z we define $\|x - z\|^2 = \sum_{j=1}^{n}(x_j - z_j)^2$

Parameter σ_i controls how quickly the weighting decays i.e how narrow or wide the bell shape is.

- Example: $z = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$





Soft k-means

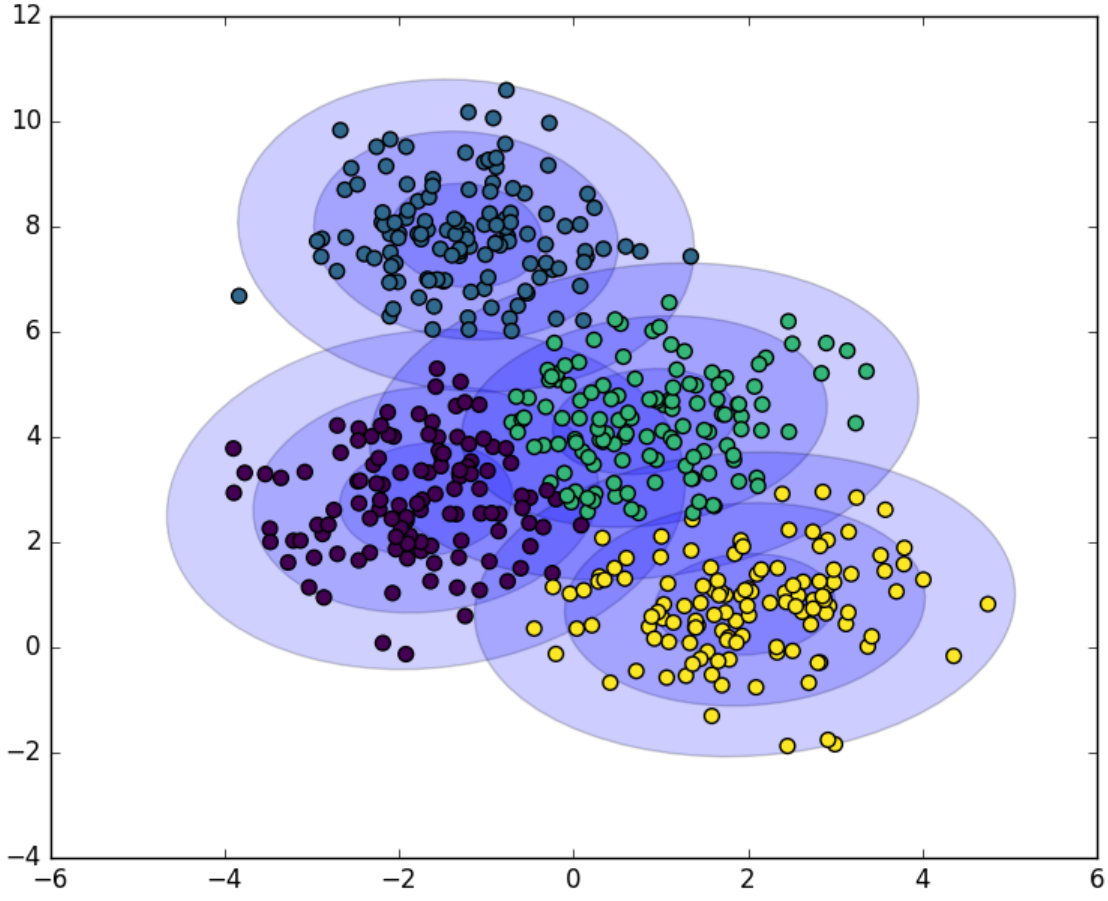
One approach:

- Modify the k-means setup so that a point x is associated with cluster j with weight $K(x, \mu^{(j)})$.
- So weight is high when x is close to the cluster centre $\mu^{(j)}$ and low when x is far from $\mu^{(j)}$.

Randomly initialise k cluster centres $\mu^{(1)}, \ldots, \mu^{(k)}$. e.g. choose k points from training set and use these (need $k < m$).

- Repeat:
  update cluster weights (probabilities ?):
  for $j = 1$ to m,
  for $j = 1$ to k
  $$w_j^{(i)} := \frac{K(x^{(i)}, \mu^{(j)})}{\sum_{l=1}^{k} K(x^{(i)}, \mu^{(k)})}$$
  update centres (take weighted average):
  for $j = 1$ to k
  $$\mu^{(j)} := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

Toy Example (cont)



Here we've labelled points with the closest centre, so its like k-means, but can use distance information to refine this.

Soft k-means: probabilistic interpretation

What about a probabilistic interpretation of soft k-means ?

Notation: we'll use μ to denote the set of vectors $\{\mu^{(1)}, \ldots, \mu^{(k)}\}$ to keep things tidier. Using Bayes Rule we have:

$$P(c^{(i)} = j | x^{(i)}, \mu) = \frac{P(x^{(i)} | c^{(i)} = j, \mu) P(c^{(i)} = j | \mu)}{P(x^{(i)} | \mu)}$$

We already have an expression for $P(x^{(i)} | c^{(i)} = j, \mu)$, so we need $P(c^{(i)} = j | \mu)$ and $P(x^{(i)} | \mu)$.

Soft k-means: probabilistic interpretation

What about $P(x^{(i)}|\mu)$?

- Use fact that $\sum_{l=1}^{k} P(c^{(i)} = l|x^{(i)}, \mu) = 1$ (since its a probability).
- That is,

$$\sum_{l=1}^{k} \frac{P(x^{(i)}|c^{(i)} = l, \mu)P(c^{(i)} = l|\mu)}{P(x^{(i)}|\mu)} = \frac{\sum_{l=1}^{k} P(x^{(i)}|c^{(i)} = l, \mu)P(c^{(i)} = l|\mu)}{P(x^{(i)}|\mu)} = 1$$

and so

$$P(x^{(i)}|\mu) = \sum_{l=1}^{k} P(x^{(i)}|c^{(i)} = l, \mu)P(c^{(i)} = l|\mu)$$

Soft k-means: probabilistic interpretation

$$P(c^{(i)} = j | x^{(i)}, \mu) = \frac{P(x^{(i)} | c^{(i)} = j, \mu) P(c^{(i)} = j | \mu)}{\sum_{l=1}^{k} P(x^{(i)} | c^{(i)} = l, \mu) P(c^{(i)} = l | \mu)}$$

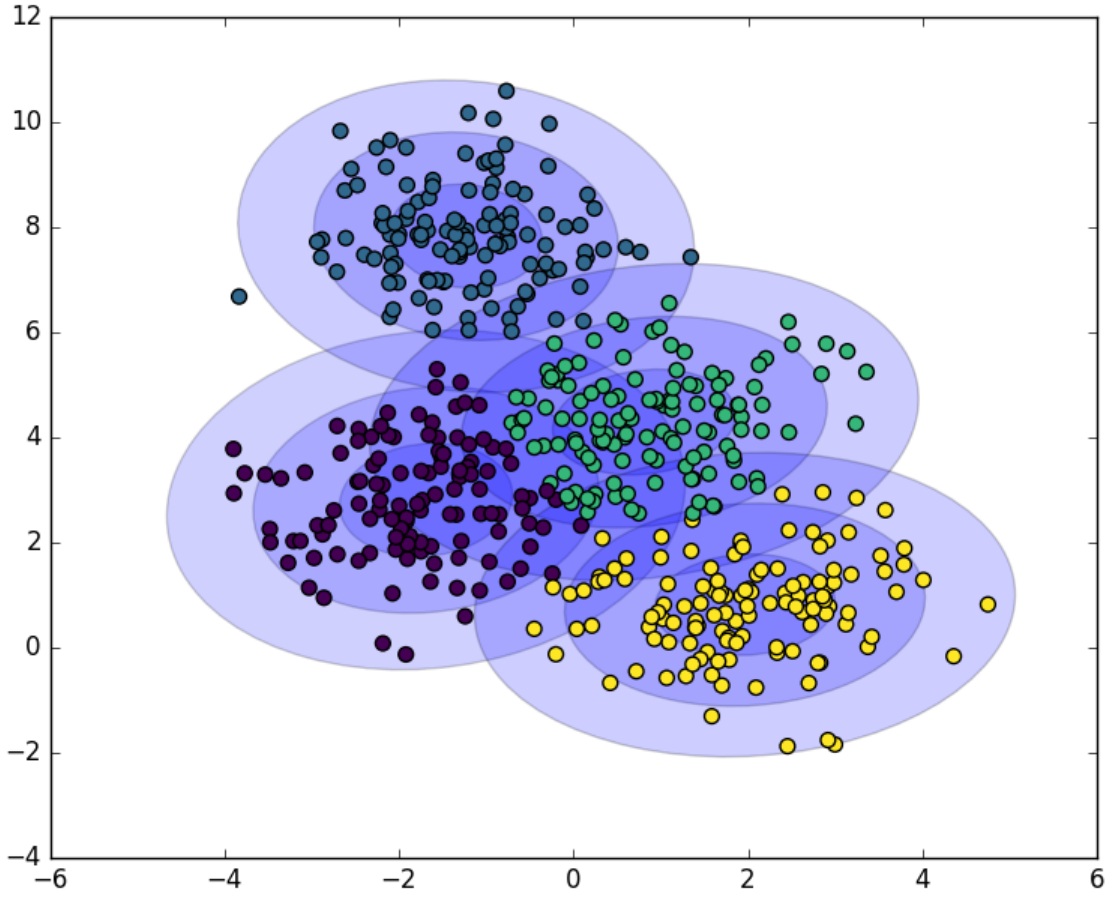
Assuming a uniform prior $P(c^{(i)} = j | \mu) = \frac{1}{k}$ for all $j = 1, \ldots, k$ then

$$P(c^{(i)} = j | x^{(i)}, \mu) = \frac{P(x^{(i)} | c^{(i)} = j, \mu)}{\sum_{l=1}^{k} P(x^{(i)} | c^{(i)} = l, \mu)}$$

Using $P(x^{(i)} | c^{(i)} = k, \mu^{(k)}) \propto K(x^{(i)}, \mu^{(k)})$ then

$$P(c^{(i)} = j | x^{(i)}, \mu) = \frac{K(x^{(i)}, \mu^{(j)})}{\sum_{l=1}^{k} K(x^{(i)}, \mu^{(j)})}$$

Toy Example (cont)



Soft k-means: probabilistic interpretation

- Additionally:

$$P(x^{(i)}|\mu) = \sum_{j=1}^{k} P(x^{(i)}|c^{(i)} = j, \mu) P(c^{(i)} = j|\mu) \propto \sum_{j=1}^{k} K(x^{(i)}, \mu^{(j)})$$

  assuming uniform prior $P(c^{(i)} = j|\mu) = \frac{1}{k}$
- So the log-likelihood of training data d is

$$\log P(d|\mu) \propto \sum_{i=1}^{m} \log \left(\sum_{j=1}^{k} K(x^{(i)}, \mu^{(j)}) \right)$$

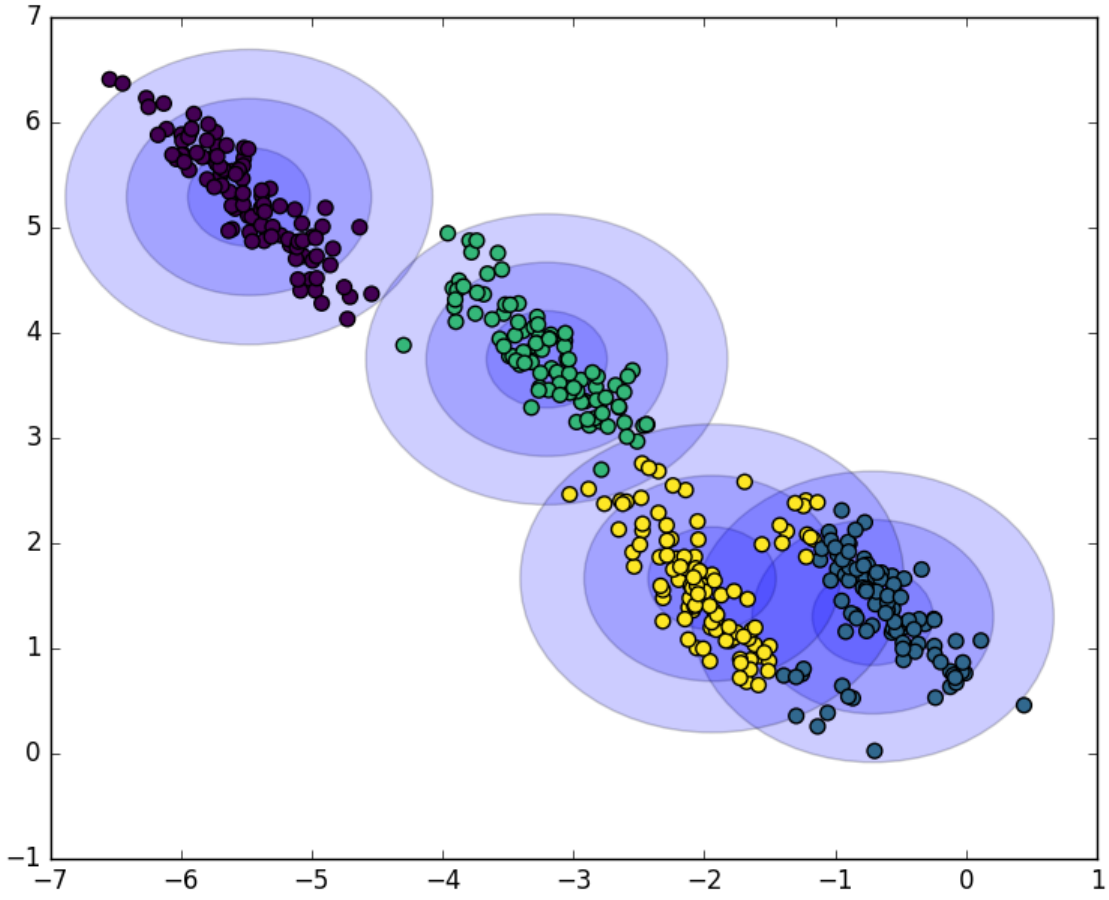
Soft k-means: probabilistic interpretation

- Select parameters $\mu^{(1)}, \ldots, \mu^{(k)}$ to maximise this log-likelihood.
- We can use gradient descent, as usual, which will converge to a local maximum.
- Or use soft k-means iteration:

  1. $w_j^{(i)} := \frac{K(x^{(i)}, \mu^{(j)})}{\sum_{l=1}^{k} K(x^{(i)}, \mu^{(k)})}$
  2. $\mu^{(j)} := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$

  (proving this converges to a local maximum is beyond this module)

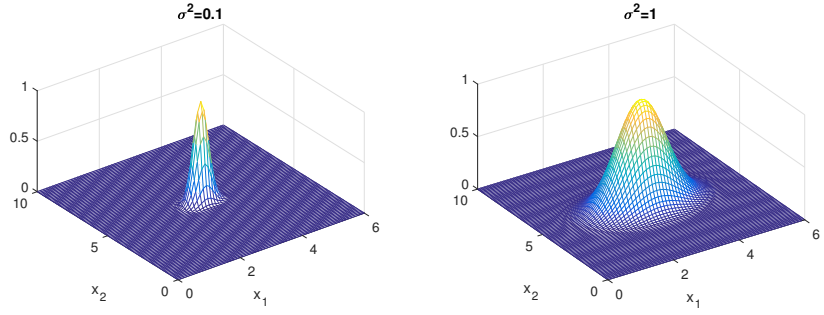
A Nastier Toy Example



Soft k-means: probabilistic interpretation

The soft k-means statistical model with $K(x,z) = e^{\frac{\|x-z\|^2}{\sigma^2}}$ assumes:

- The Gaussian noise has the same variance σ^2 for every class, but some classes might be more or less concentrated than others.



- A uniform prior $P(c^{(i)} = j|\mu) = \frac{1}{k}$ on the class membership

Extensions:

- Use kernels $K_{\sigma^{(j)}}(x,z) = e^{\frac{\|x-z\|^2}{(\sigma^{(j)})^2}}$ with different variance $\sigma^{(j)}$ for each cluster. Select value of $\sigma^{(j)}$ that maximises the log-likelihood
- Introduce $P(c^{(i)} = j|\mu)$ as another parameter to be estimated, again to maximise the log-likelihood

Soft k-means: probabilistic interpretation

Extended algorithm

Repeat:

update cluster probabilities $w_j = P(c^{(i)} = j | x^{(i)}, \mu)$**:**

for $j = 1$ to m, for $j = 1$ to k $\{ \; w_j^{(i)} := \frac{K(x^{(i)}, \mu^{(j)}) v_j}{\sum_{l=1}^{k} K(x^{(i)}, \mu^{(k)}) v_l} \}$

update centres $\mu^{(j)}$**:**

for $j = 1$ to k $\{ \mu^{(j)} := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \}$

update variance $(\sigma^{(j)})^2$**:**

for $j = 1$ to k $\{ (\sigma^{(j)})^2 := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu^{(j)})^T (x^{(i)} - \mu^{(j)})}{\sum_{i=1}^{m} w_j^{(i)}} \}$

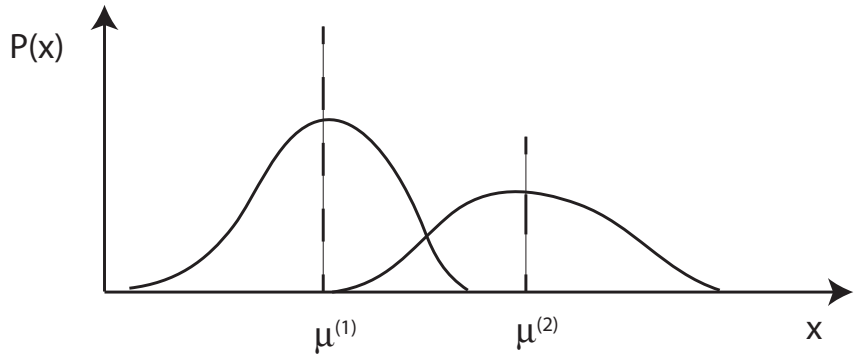
update prior $v_j = P(c^{(i)} = j | \mu)$**:**

for $j = 1$ to k $\{ \; v_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \}$

Gaussian Mixture Model

This statistical model is commonly known as a **Gaussian Mixture Model** or a **Mixture of Gaussian Model**. Why ?

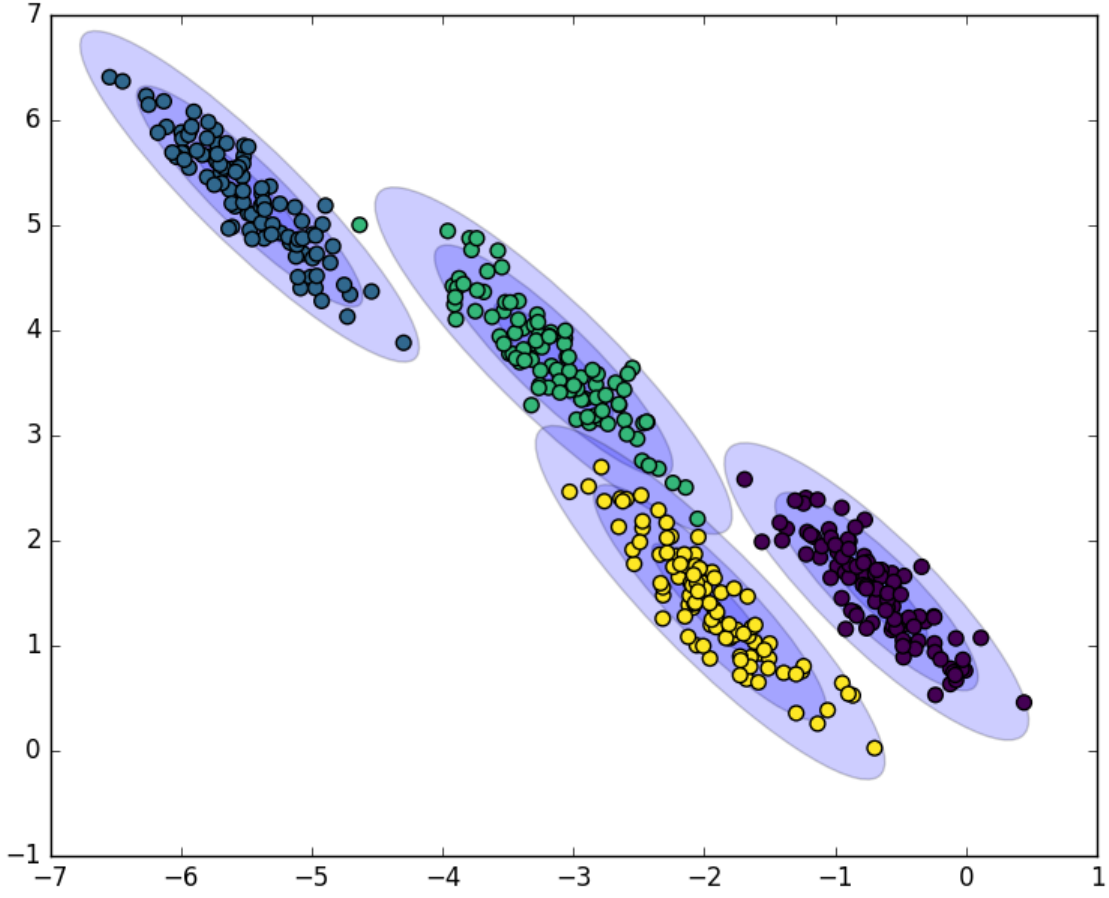
- $P(x|\mu) \propto \sum_{j=1}^{k} e^{\frac{\|x-\mu^{(j)}\|^2}{2(\sigma^{(j)})^2}} P(c=j)$
- The likelihood is expressed as the weighted sum of gaussians.



The update algorithm is an example of the **Expectation-Maximisation** algorithm, a variant of gradient descent.

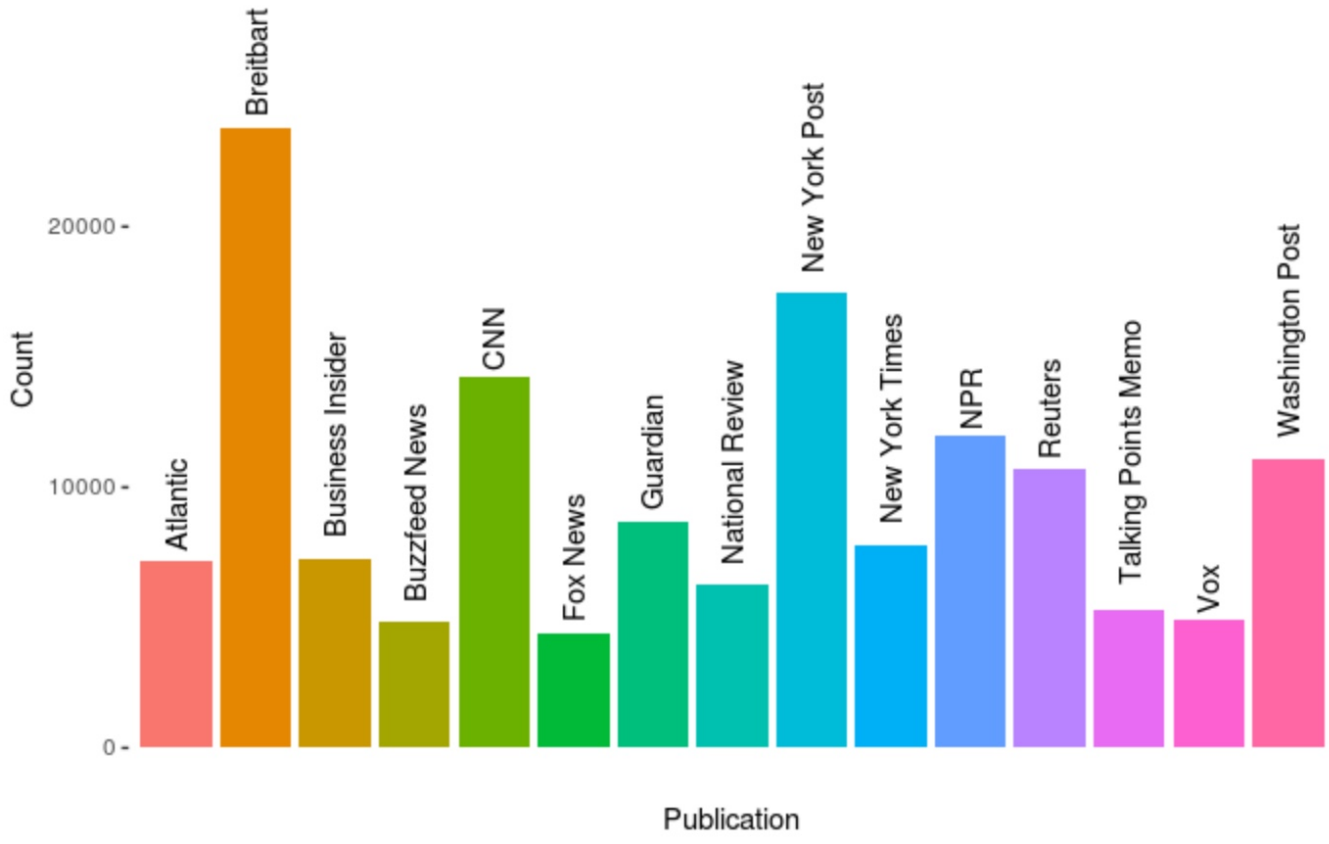
The algorithm outputs $P(c^{(i)} = j|x^{(i)}, \mu)$. The class labels $c^{(i)}$, $i = 1, \ldots, m$ are often called **latent variables** since they are not observed (only $x^{(i)}$ is observed).

A Nastier Toy Example (cont)



Example: Clustering News Articles

- Dataset: https://www.kaggle.com/snapcrack/all-the-news/home
- Fields: id, title, publication name, author, date, year, month, url, content
- We'll use 50,000 articles from articles1.csv



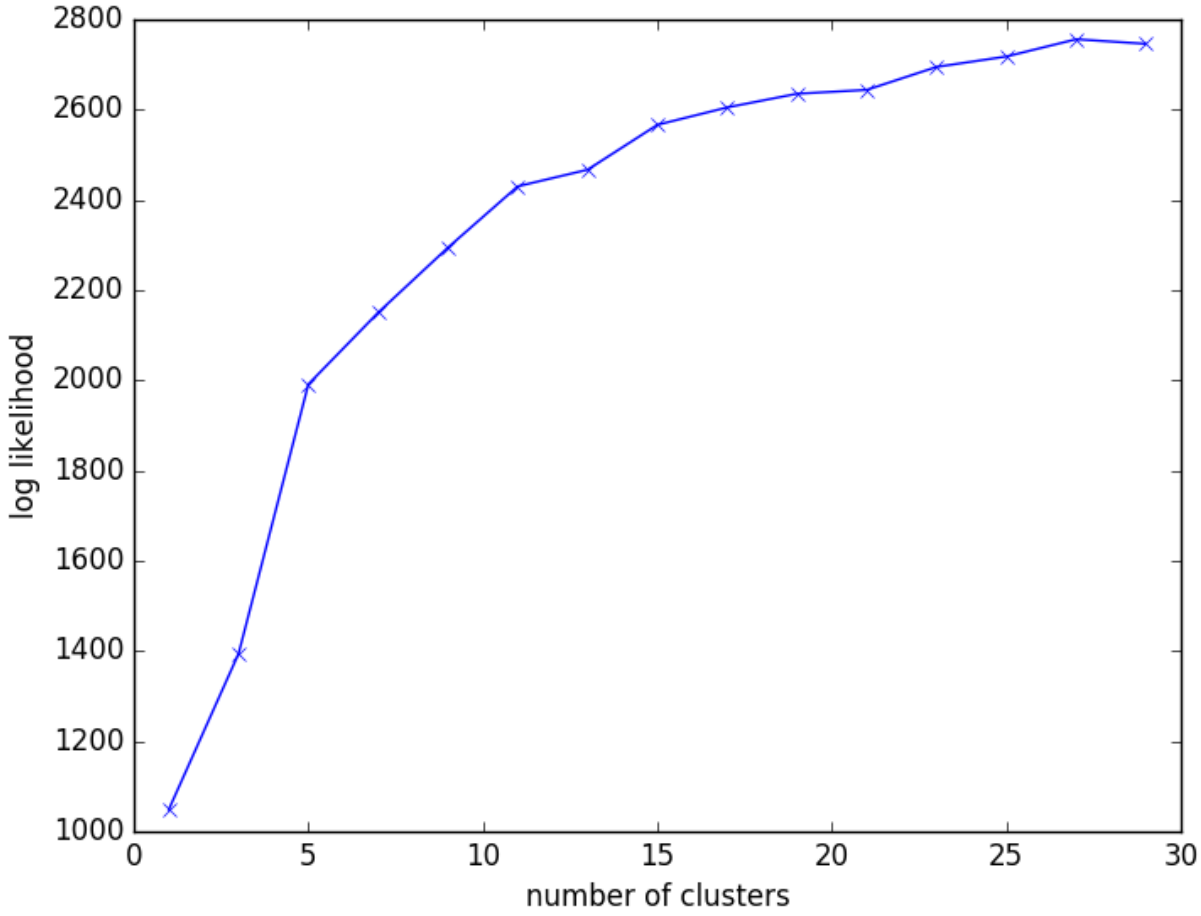
Example: Clustering News Articles

First five headlines:

- Rift Between Officers and Residents as Killings Persist in South Bronx - The New York Times
- Tyrus Wong, Bambi Artist Thwarted by Racial Bias, Dies at 106 - The New York Times
- Among Deaths in 2016, a Heavy Toll in Pop Music - The New York Times
- Kim Jong-un Says North Korea Is Preparing to Test Long-Range Missile - The New York Times

Example: Clustering News Articles

- Remove stop words, use steming
- Bag of words model
- Use 500 most frequent terms (to speed things up)



Example: Clustering News Articles

First 25 most frequent terms in mean of some clusters:

- cluster0: ['order', '', 'iran', 'united', 'refugees', 'briefing', 'ban', 'countries', 'american', 'immigration', 'executive', 'nations', 'muslim', 'administration', 'security', 'iraq', 'state', 'world', 'obama', 'officials', 'country', 'seven', 'syria', 'policy', 'department']
- cluster6: ['ms', 'family', 'mother', 'husband', 'school', 'children', 'york', 'daughter', 'life', 'times', 'father', 'home', 'women', 'told', 'food', 'work', 'education', 'public', 'job', 'news', 'day', 'love', 'month', 'son', 'later']
- cluster14: ['european', 'mrs', 'britain', 'union', 'british', 'minister', 'trade', 'prime', 'deal', 'sunday', 'foreign', 'husband', 'letter', 'turkey', 'ms', 'american', 'party', 'member', 'world', 'country', 'day', 'statement', 'market', 'united', 'service']

Example: Clustering News Articles

Titles of some articles in cluster 0:

- With New Congress Poised to Convene, Obamas Policies Are in Peril - The New York Times
- Republicans Stonewalled Obama. Now the Ball Is in Their Court. - The New York Times
- House Republicans, Under Fire, Back Down on Gutting Ethics Office - The New York Times
- In Republicans Ethics Office Gambit, a Spectacle of Tweets and Retreats - The New York Times
- Senate Confirmation Hearings to Begin Without All Background Checks - The New York Times

Example: Clustering News Articles

Titles of some articles in cluster 6:

- After The Biggest Loser, Their Bodies Fought to Regain Weight - The New York Times
- Work. Walk 5 Minutes. Work. - The New York Times
- Is Your Workout Not Working? Maybe Youre a Non-Responder - The New York Times
- Scientists Say the Clock of Aging May Be Reversible - The New York Times
- Light Pillars, a Million-Mirror Optical Illusion on Winter Nights - The New York Times