

Clustering K-means

Machine Learning – CSE546

Carlos Guestrin

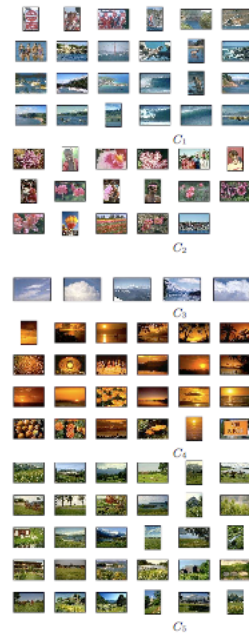
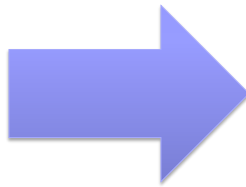
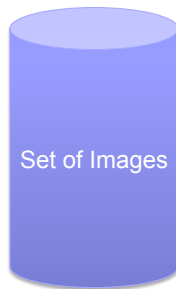
University of Washington

November 4, 2014

©Carlos Guestrin 2005-2014

1

Clustering images



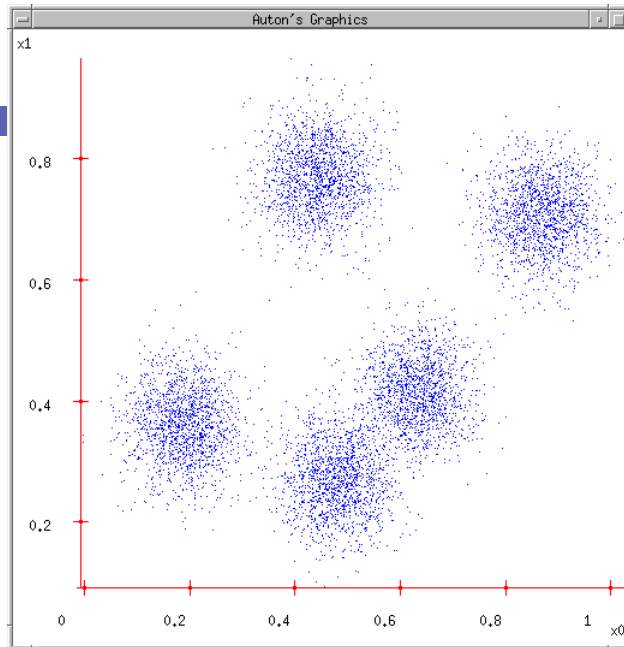
©Carlos Guestrin 2005-2014

[Goldberger et al.] 2

Clustering web search results

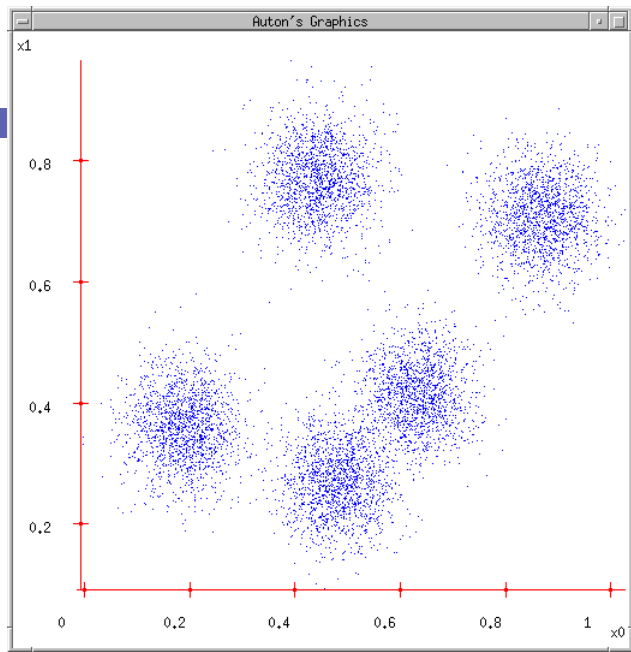
The screenshot shows the Clusty search interface. At the top, there are navigation links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. A search bar contains the word 'race'. Below the search bar, a sidebar on the left lists various categories and their counts: Car (28), Race cars (7), Photos, Races Scheduled (5), Game (4), Track (3), NASCAR (2), Equipment And Safety (2), Other Topics (7), Photos (22), Game (14), Definition (13), Team (18), Human (8), Classification Of Human (2), Statement, Evolved (2), Other Topics (4), Weekend (8), Ethnicity And Race (7), Race for the Cure (8), Race Information (8), and a 'more | all clusters' link. The main content area displays a list of 7 search results, each with a title, a small icon, and a brief description. The results include Wikipedia entries, a book, a statement, and a website. At the bottom of the page, there is a copyright notice: '©Carlos Guestrin 2005-2014' and a page number '3'.

Some Data



K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)

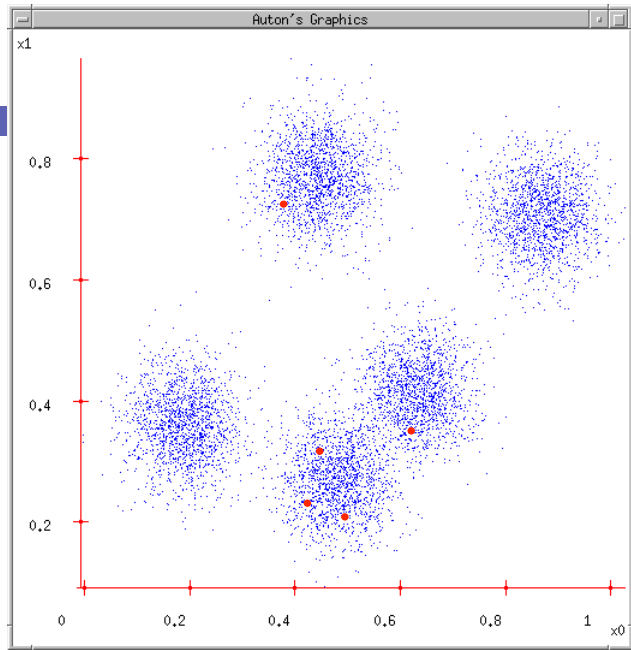


©Carlos Guestrin 2005-2014

5

K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations

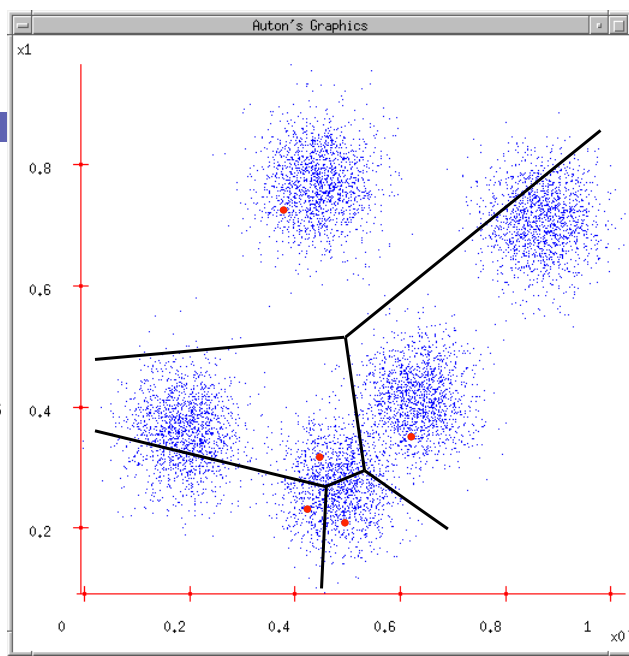


©Carlos Guestrin 2005-2014

6

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

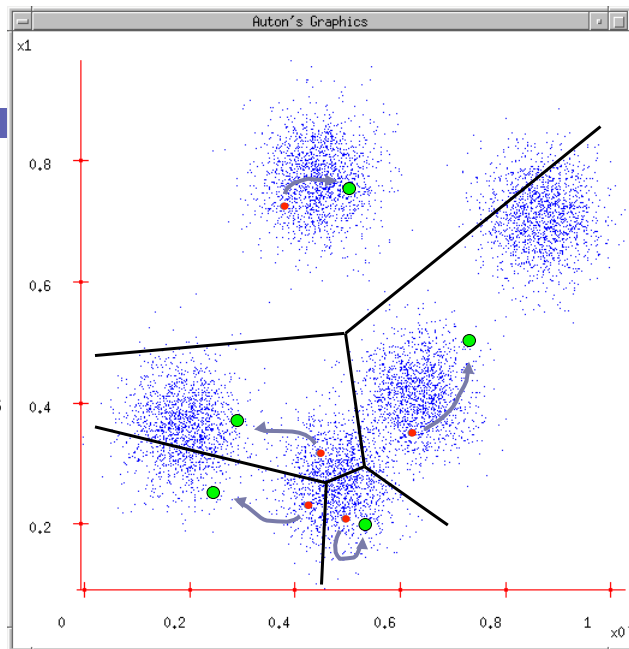


©Carlos Guestrin 2005-2014

7

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

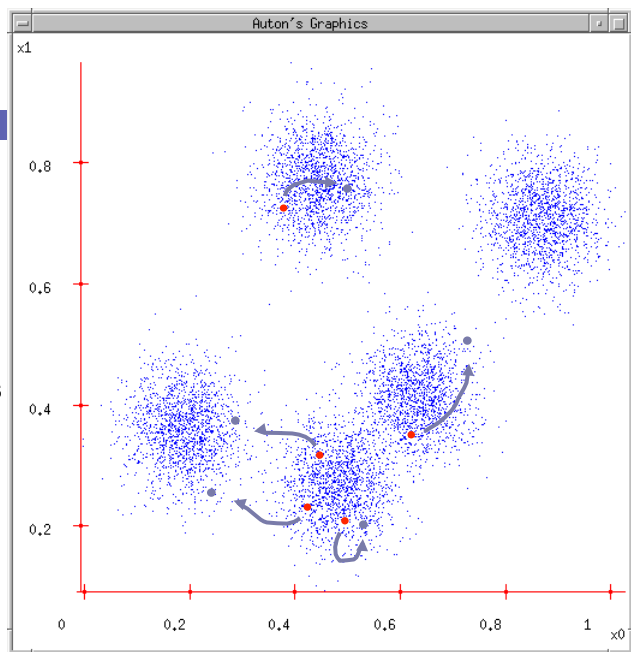


©Carlos Guestrin 2005-2014

8

K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



©Carlos Guestrin 2005-2014

9

K-means

- Randomly initialize k centers
 - $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$
- **Classify:** Assign each point $j \in \{1, \dots, N\}$ to nearest center:
 - $C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$
- **Recenter:** μ_i becomes centroid of its point:
 - $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j)=i} \|\mu - x_j\|^2$
 - Equivalent to $\mu_i \leftarrow$ average of its points!

©Carlos Guestrin 2005-2014

10

What is K-means optimizing?

- Potential function $F(\mu, C)$ of centers μ and point allocations C :

- $F(\mu, C) = \sum_{j=1}^N \|\mu_{C(j)} - x_j\|^2$

- Optimal K-means:

- $\min_{\mu} \min_C F(\mu, C)$

Does K-means converge??? Part 1

- Optimize potential function:

- $$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Fix μ , optimize C

Does K-means converge??? Part 2

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Fix C, optimize μ

Coordinate descent algorithms

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Want: $\min_a \min_b F(a,b)$
- Coordinate descent:
 - fix a, minimize b
 - fix b, minimize a
 - repeat
- Converges!!!
 - if F is bounded
 - to a (often good) local optimum
 - as we saw in applet (play with it!)
 - (For LASSO it converged to the global optimum, because of convexity)
- K-means is a coordinate descent algorithm!

Mixtures of Gaussians

Machine Learning – CSE546

Carlos Guestrin

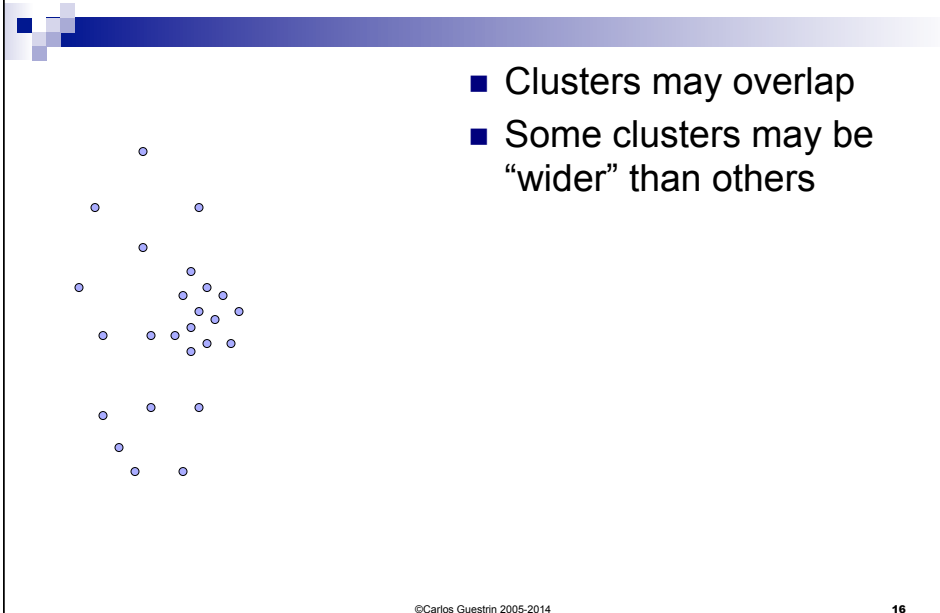
University of Washington

November 4, 2014

©Carlos Guestrin 2005-2014

15

(One) bad case for k-means

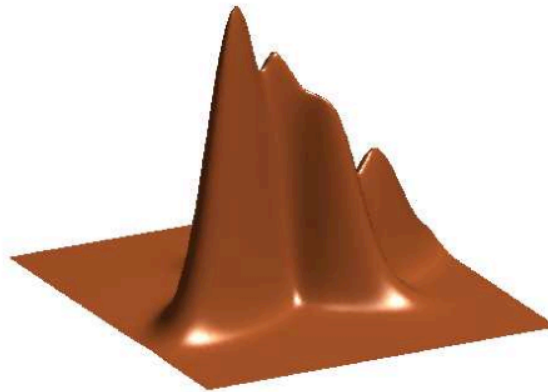


©Carlos Guestrin 2005-2014

16

Density Estimation

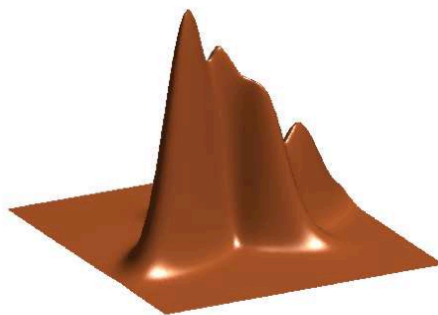
- Estimate a density based on x^1, \dots, x^N



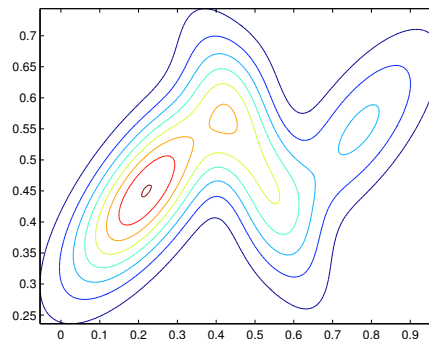
©Carlos Guestrin 2005-2014

17

Density Estimation



Contour Plot of Joint Density



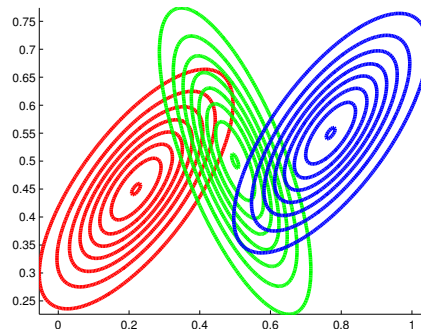
©Carlos Guestrin 2005-2014

18

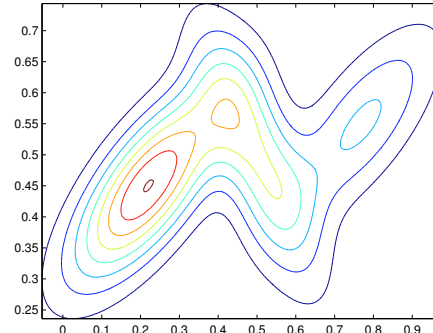
Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



Contour Plot of Joint Density



©Carlos Guestrin 2005-2014

19

Gaussians in d Dimensions

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \|\Sigma\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right]$$

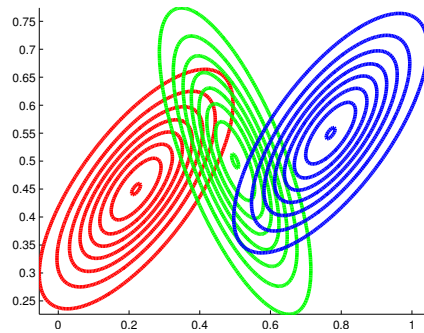
©Carlos Guestrin 2005-2014

20

Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



$$p(x^i | \pi, \mu, \Sigma) =$$

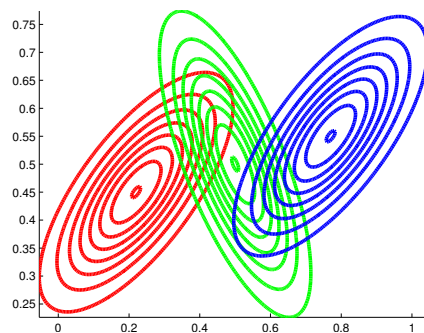
©Carlos Guestrin 2005-2014

21

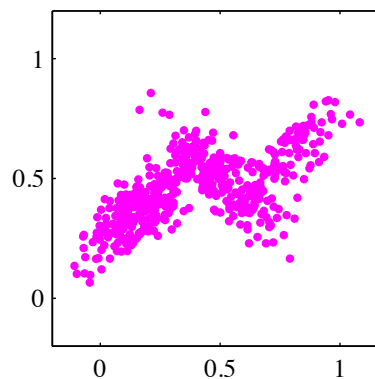
Density as Mixture of Gaussians

- Approximate with density with a mixture of Gaussians

Mixture of 3 Gaussians



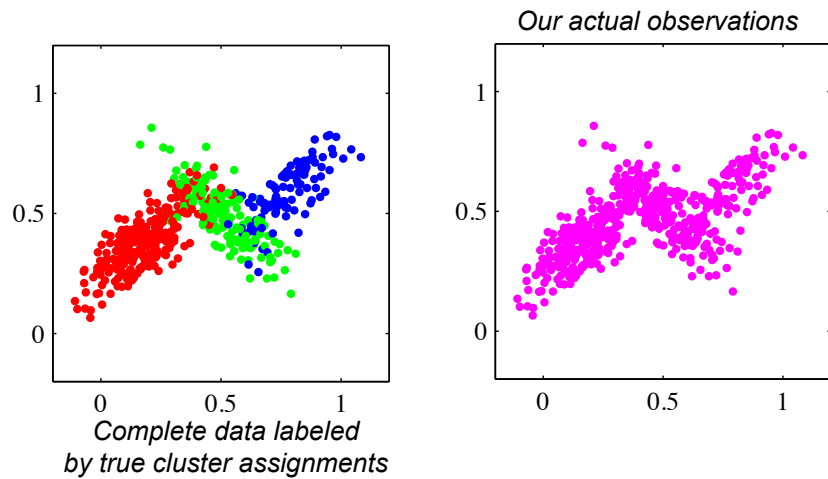
Our actual observations



C. Bishop, *Pattern Recognition & Machine Learning*

Clustering our Observations

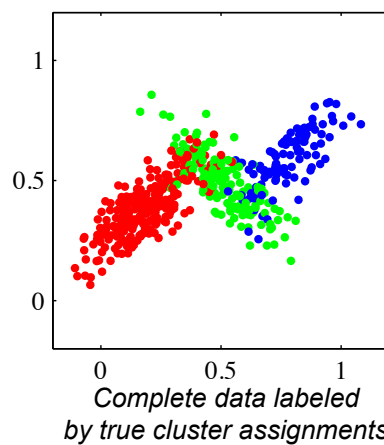
- Imagine we have an assignment of each x^i to a Gaussian



C. Bishop, *Pattern Recognition & Machine Learning*

Clustering our Observations

- Imagine we have an assignment of each x^i to a Gaussian



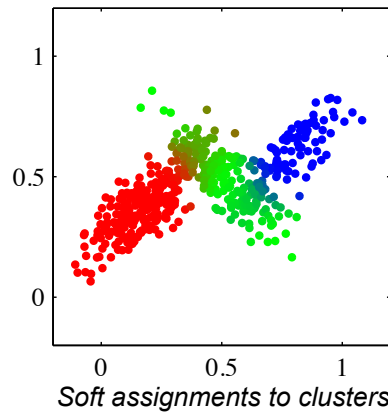
- Introduce latent cluster indicator variable z^i

- Then we have
$$p(x^i | z^i, \pi, \mu, \Sigma) =$$

C. Bishop, *Pattern Recognition & Machine Learning*

Clustering our Observations

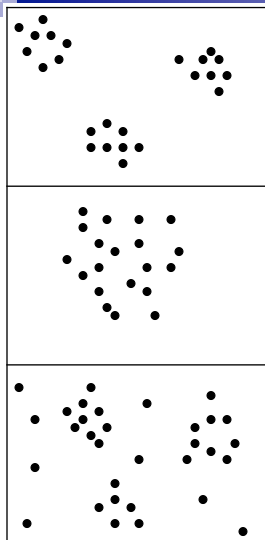
- We must infer the cluster assignments from the observations



- Posterior probabilities of assignments to each cluster *given* model parameters:
 $r_{ik} = p(z^i = k | x^i, \pi, \mu, \Sigma) =$

C. Bishop, *Pattern Recognition & Machine Learning*

Unsupervised Learning: not as hard as it looks



Sometimes easy

Sometimes impossible

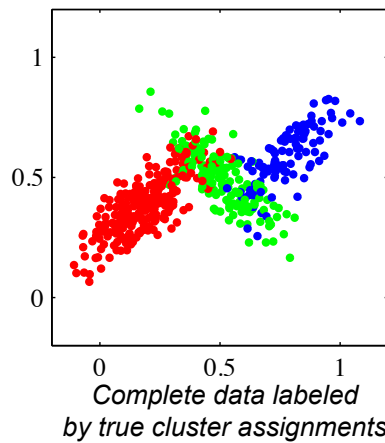
and sometimes in between

©Carlos Guestrin 2005-2014

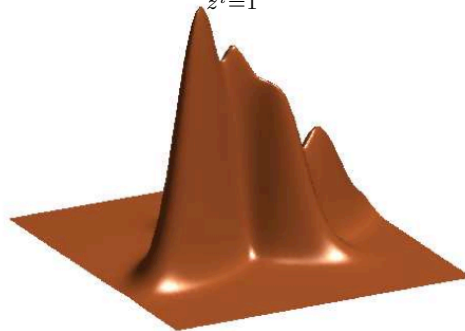
26

Summary of GMM Concept

- Estimate a density based on x^1, \dots, x^N



$$p(x^i | \pi, \mu, \Sigma) = \sum_{z^i=1}^K \pi_{z^i} \mathcal{N}(x^i | \mu_{z^i}, \Sigma_{z^i})$$



©Carlos Guestrin 2005-2014

27

Summary of GMM Components

- Observations $x^i \in \mathbb{R}^d, \quad i = 1, 2, \dots, N$
- Hidden cluster labels $z_i \in \{1, 2, \dots, K\}, \quad i = 1, 2, \dots, N$
- Hidden mixture means $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \dots, K$
- Hidden mixture covariances $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \dots, K$
- Hidden mixture probabilities $\pi_k, \quad \sum_{k=1}^K \pi_k = 1$

Gaussian mixture marginal and conditional likelihood :

$$p(x^i | \pi, \mu, \Sigma) = \sum_{z^i=1}^K \pi_{z^i} p(x^i | z^i, \mu, \Sigma)$$

$$p(x^i | z^i, \mu, \Sigma) = \mathcal{N}(x^i | \mu_{z^i}, \Sigma_{z^i})$$

©Carlos Guestrin 2005-2014

28