

So far, supervised learning:
 $h: \mathcal{X} \rightarrow \mathcal{R}$ ← regression
 $h: \mathcal{X} \rightarrow \{0, 1, \dots, k\}$ ← classification

Unsupervised learning

Clustering

K-means Continued

Machine Learning – CSE446

Carlos Guestrin

University of Washington

May 15, 2013

©Carlos Guestrin 2005-2013

1

Clustering images

no labels given

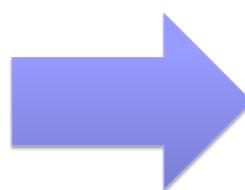


branch

flowers

sky

Set of Images



©Carlos Guestrin 2005-2013

[Goldberger et al.] 2

Clustering web search results

The screenshot shows a web search interface with a sidebar for clusters, sources, and sites. The search term 'race' is highlighted in red. The results list includes:

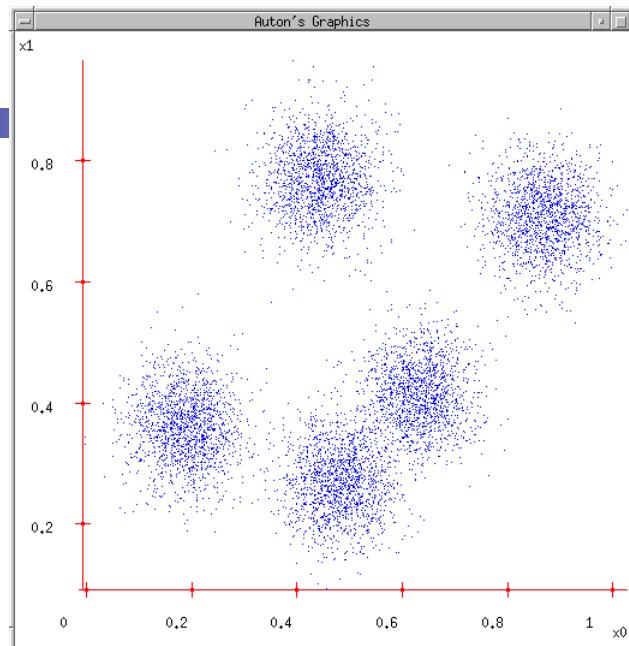
- Race (classification of human beings) - Wikipedia, the free encyclopedia**: A detailed explanation of racial categories based on visible traits like skin color, cranial or facial features, and hair texture.
- Race - Wikipedia, the free encyclopedia**: General racing competitions.
- Publications | Human Rights Watch**: Information about torture, unfair trials, and human rights issues related to race.
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich**: A book by Vincent Sarich and Frank Miele.
- AAPA Statement on Biological Aspects of Race**: A statement from the American Journal of Physical Anthropology.
- race: Definition from Answers.com**: A general definition of race.
- Dopefish.com**: A site for newbies and experienced users of the Dopefish game.

Cluster Human contains 8 documents.

©Carlos Guestrin 2005-2013

3

Some Data

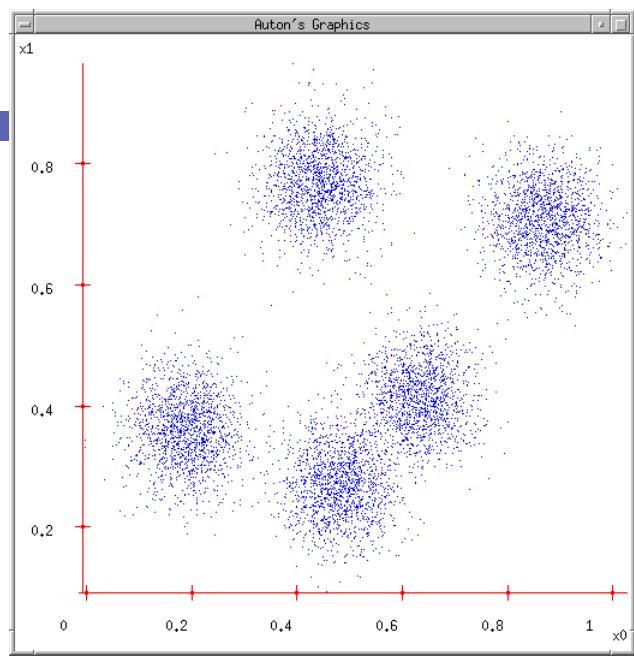


©Carlos Guestrin 2005-2013

4

K-means

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

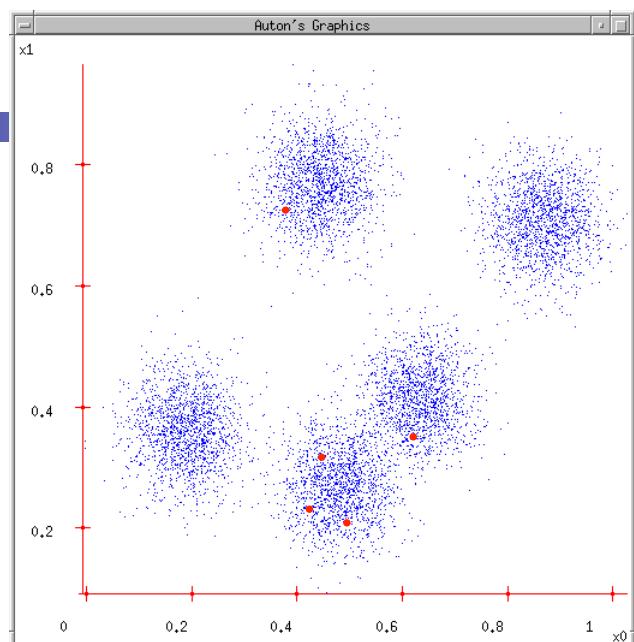


©Carlos Guestrin 2005-2013

5

K-means

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

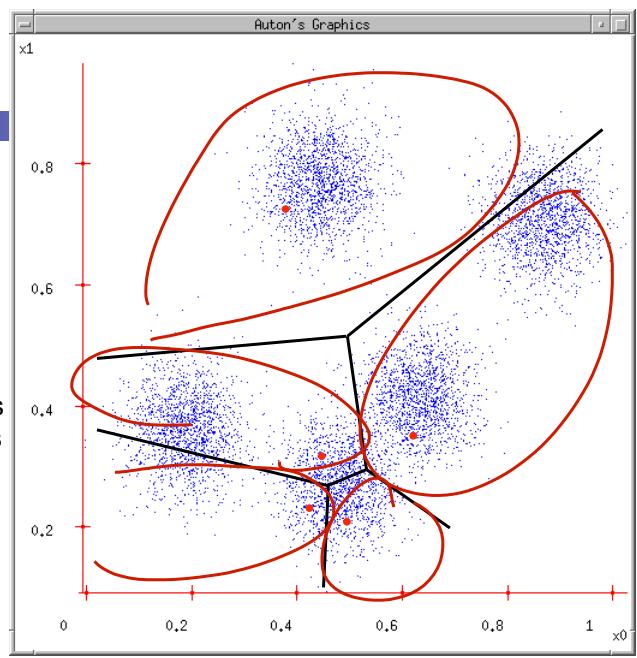


©Carlos Guestrin 2005-2013

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)

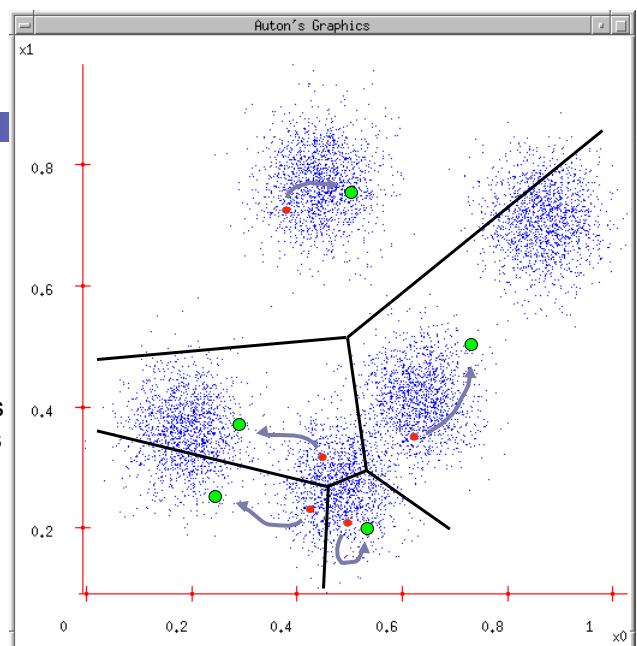


©Carlos Guestrin 2005-2013

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 2005-2013

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 2005-2013

9

K-means

- Randomly initialize k centers *or smartly*
 - $\square \mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$
Converge when nothing moves (no point changes its cluster)
- Classify: Assign each point $j \in \{1, \dots, n\}$ to nearest center:
 - $\square C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$ *in cluster center* *data point*
- Recenter: μ_i becomes centroid of its point:
 - $\square \mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j:C(j)=i} \|\mu - x_j\|^2 \leftarrow \mu_i = \frac{\sum_{j:C(j)=i} x_j}{\text{num of points assigned to cluster } i}$
 - \square Equivalent to $\mu_i \leftarrow \text{average of its points!}$

©Carlos Guestrin 2005-2013

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:

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

©Carlos Guestrin 2005-2013

11

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

©Carlos Guestrin 2005-2013

12

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

©Carlos Guestrin 2005-2013

13

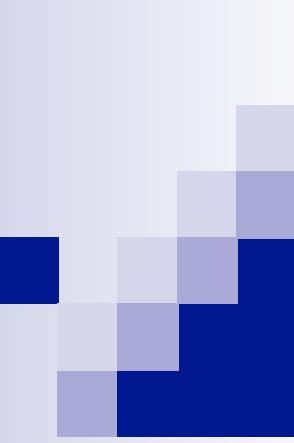
Coordinate descent algorithms

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

- 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 optimum)
- K-means is a coordinate descent algorithm!

©Carlos Guestrin 2005-2013

14



Mixtures of Gaussians

Machine Learning – CSE446

Carlos Guestrin

University of Washington

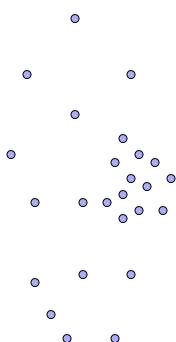
May 15, 2013

©Carlos Guestrin 2005-2013

15

(One) bad case for k-means

- Clusters may overlap
- Some clusters may be “wider” than others



©Carlos Guestrin 2005-2013

16

Gaussians in m Dimensions

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

©Carlos Guestrin 2005-2013

17

Suppose You Have a Gaussian For Each Class

$$P(\mathbf{x} | y = i) \propto \frac{1}{(2\pi)^{m/2} \|\Sigma_i\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right]$$

©Carlos Guestrin 2005-2013

18

Gaussian Bayes Classifier

- You have a Gaussian over \mathbf{x} for each class $y=i$:

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

- But you need probability of class $y=i$ given \mathbf{x} :

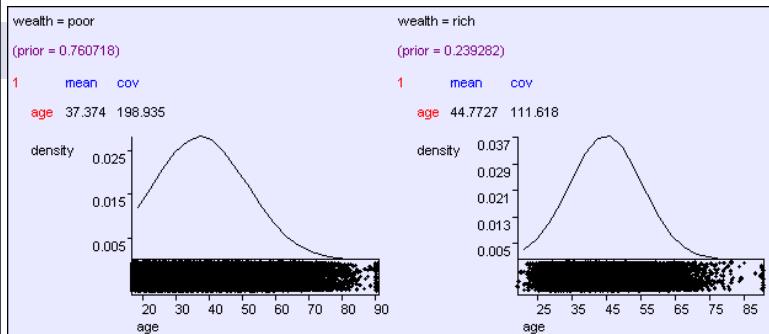
- Thank you Bayes Rule!!

$$P(y = i | \mathbf{x}) = \frac{p(\mathbf{x} | y = i)P(y = i)}{p(\mathbf{x})}$$

©Carlos Guestrin 2005-2013

19

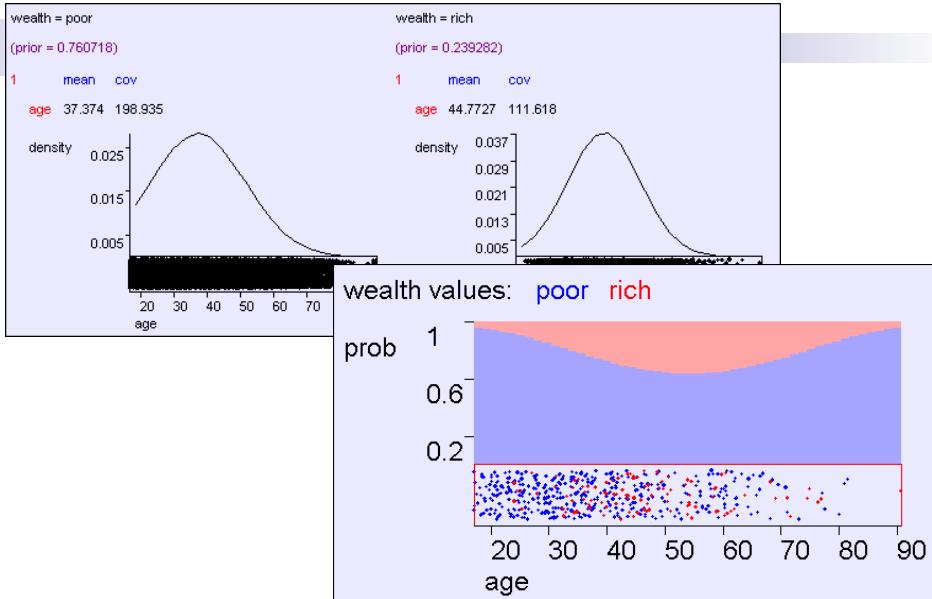
Predicting wealth from age



©Carlos Guestrin 2005-2013

20

Predicting wealth from age

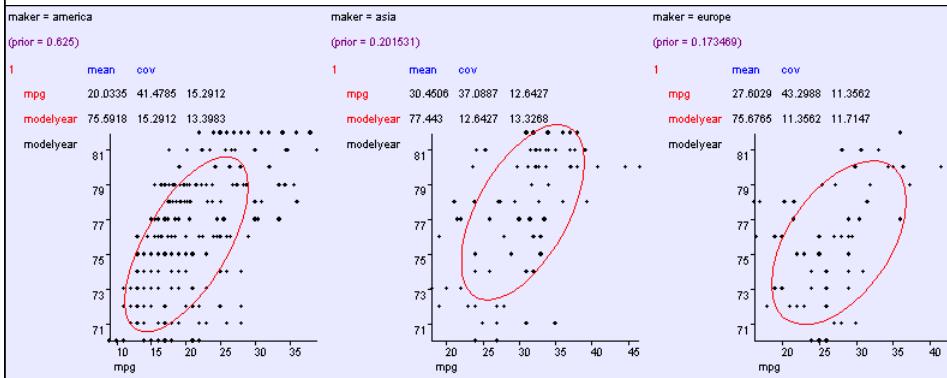


©Carlos Guestrin 2005-2013

21

Learning modelyear , mpg ---> maker

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

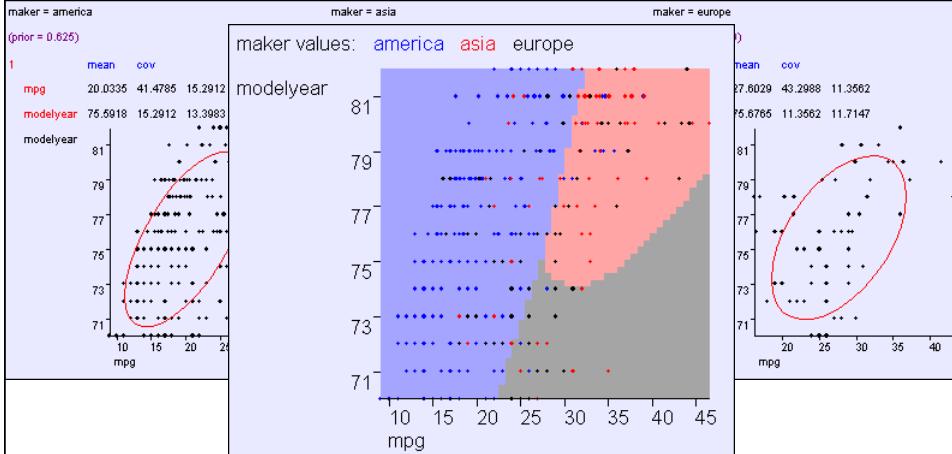


©Carlos Guestrin 2005-2013

22

General: $O(m^2)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

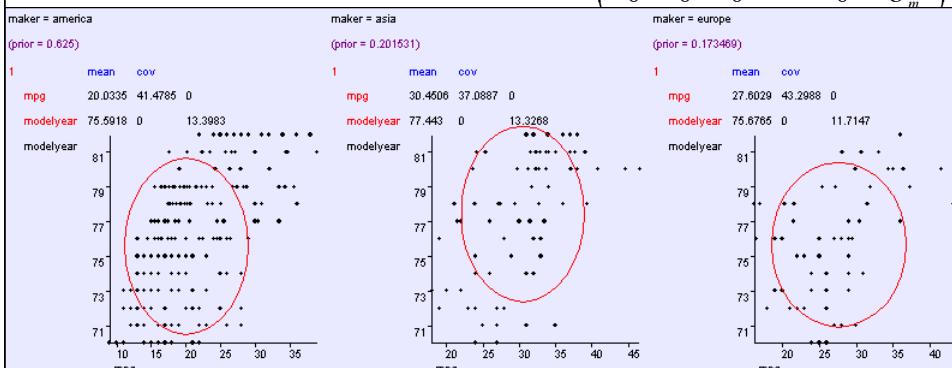


©Carlos Guestrin 2005-2013

23

Aligned: $O(m)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_m^2 \end{pmatrix}$$

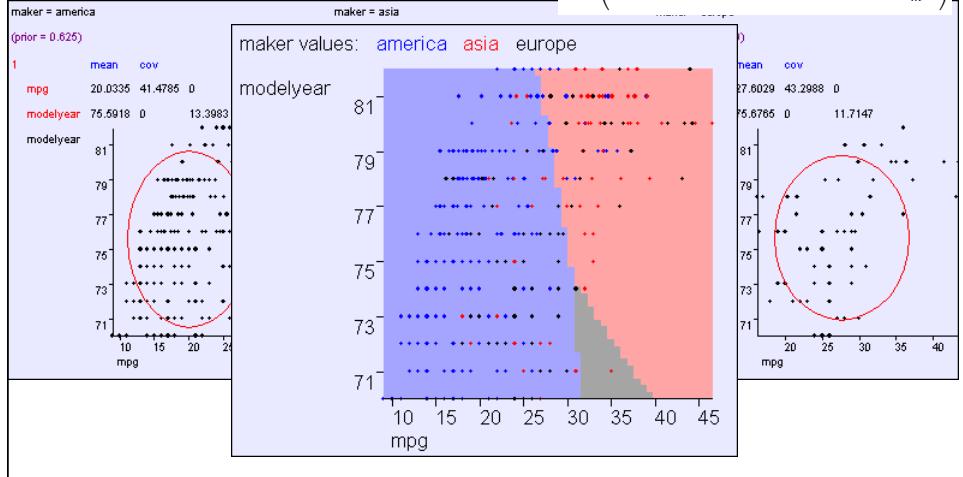


©Carlos Guestrin 2005-2013

24

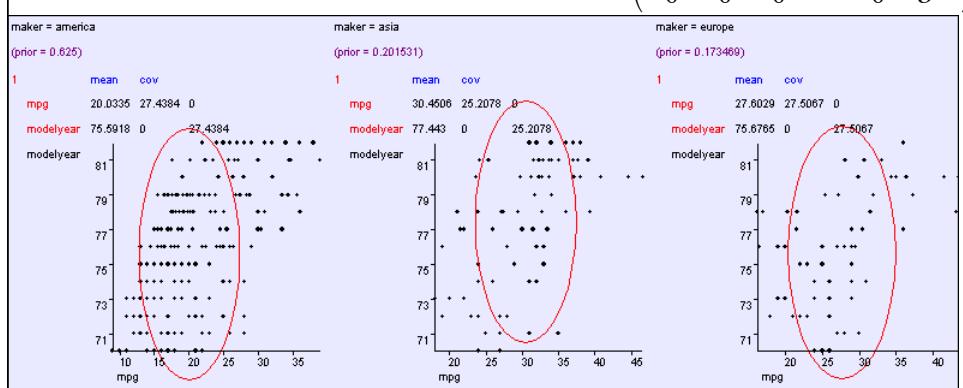
Aligned: $O(m)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_m^2 \end{pmatrix}$$



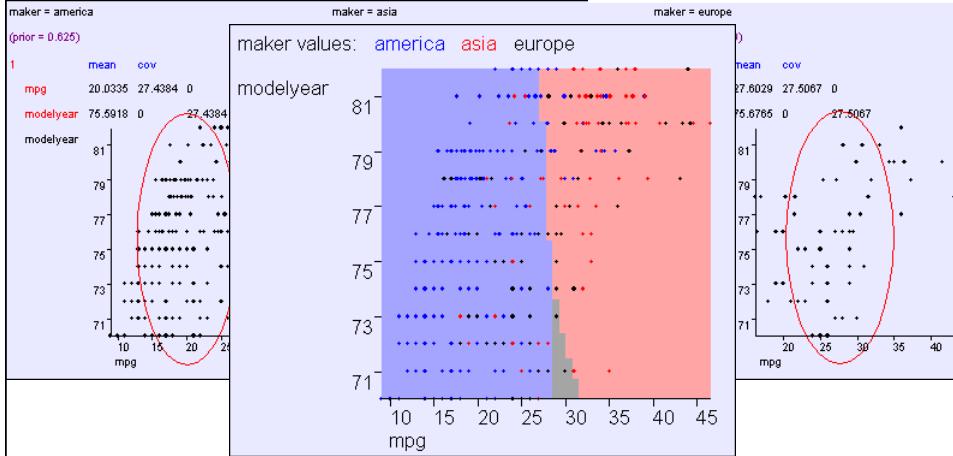
Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$



Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

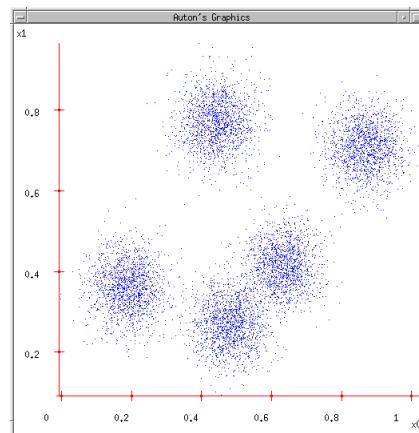


©Carlos Guestrin 2005-2013

27

Next... back to Density Estimation

What if we want to do density estimation with multimodal or clumpy data?



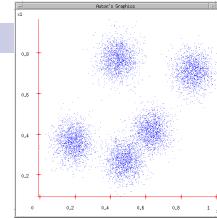
©Carlos Guestrin 2005-2013

28

But we don't see class labels!!!

- MLE:

- $\operatorname{argmax} \prod_j P(y^j, x^j)$



- But we don't know y^j !!!
- Maximize marginal likelihood:
 - $\operatorname{argmax} \prod_j P(x^j) = \operatorname{argmax} \prod_j \sum_{i=1}^k P(y^j=i, x^j)$

©Carlos Guestrin 2005-2013

29

Special case: spherical Gaussians and hard assignments

- $P(y = i | \mathbf{x}^j) \propto \frac{1}{(2\pi)^{m/2} \|\Sigma_i\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}^j - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}^j - \mu_i)\right] P(y = i)$

- If $P(X|Y=i)$ is spherical, with same σ for all classes:

$$P(\mathbf{x}^j | y = i) \propto \exp\left[-\frac{1}{2\sigma^2} \|\mathbf{x}^j - \mu_i\|^2\right]$$
- If each x_j belongs to one class $C(j)$ (hard assignment), marginal likelihood:

$$\prod_{j=1}^m \sum_{i=1}^k P(\mathbf{x}^j, y = i) \propto \prod_{j=1}^m \exp\left[-\frac{1}{2\sigma^2} \|\mathbf{x}^j - \mu_{C(j)}\|^2\right]$$

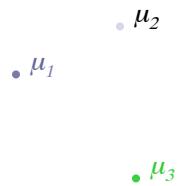
- Same as K-means!!!

©Carlos Guestrin 2005-2013

30

The GMM assumption

- There are k components
- Component i has an associated mean vector μ_i



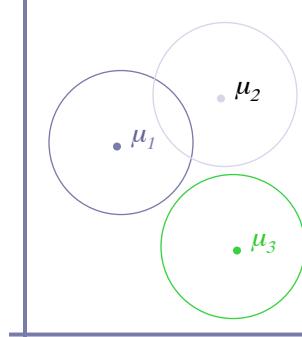
©Carlos Guestrin 2005-2013

31

The GMM assumption

- There are k components
- Component i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 I$

Each data point is generated according to the following recipe:



©Carlos Guestrin 2005-2013

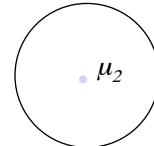
32

The GMM assumption

- There are k components
- Component i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 I$

Each data point is generated according to the following recipe:

1. Pick a component at random:
Choose component i with probability $P(y=i)$



©Carlos Guestrin 2005-2013

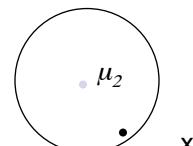
33

The GMM assumption

- There are k components
- Component i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 I$

Each data point is generated according to the following recipe:

1. Pick a component at random:
Choose component i with probability $P(y=i)$
2. Datapoint $\sim N(\mu_i, \sigma^2 I)$



©Carlos Guestrin 2005-2013

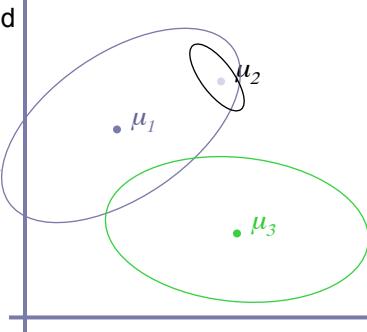
34

The General GMM assumption

- There are k components
- Component i has an associated mean vector m_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix Σ_i

Each data point is generated according to the following recipe:

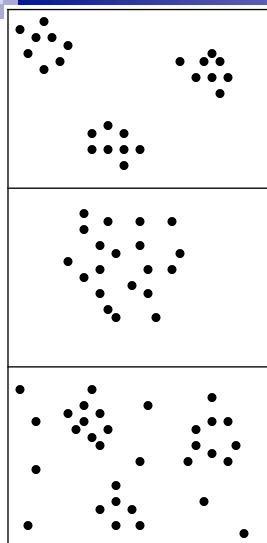
1. Pick a component at random:
Choose component i with probability $P(y=i)$
2. Datapoint $\sim N(m_i, \Sigma_i)$



©Carlos Guestrin 2005-2013

35

Unsupervised Learning: not as hard as it looks



Sometimes easy

Sometimes impossible

and sometimes in between

IN CASE YOU'RE WONDERING WHAT THESE DIAGRAMS ARE, THEY SHOW 2-d UNLABELED DATA (X VECTORS) DISTRIBUTED IN 2-d SPACE. THE TOP ONE HAS THREE VERY CLEAR GAUSSIAN CENTERS

©Carlos Guestrin 2005-2013

36