ECE 5984: Introduction to Machine Learning

Topics:

- (Finish) Expectation Maximization
- Principal Component Analysis (PCA)

Readings: Barber 15.1-15.4

Dhruv Batra

Virginia Tech

Administrativia

- Poster Presentation: Best Project Prize!
  - May 8 1:30-4:00pm
  - 310 Kelly Hall: ICTAS Building
  - Print poster (or bunch of slides)
  - Format:
    - Portrait
    - Make 1 dimen = 36in
    - Board size = 30x40
  - Less text, more pictures.

Administrativia

- Final Exam
  - When: May 11 7:45-9:45am
  - Where: In class

  - Format: Pen-and-paper.
  - Open-book, open-notes, closed-internet.
    - No sharing.

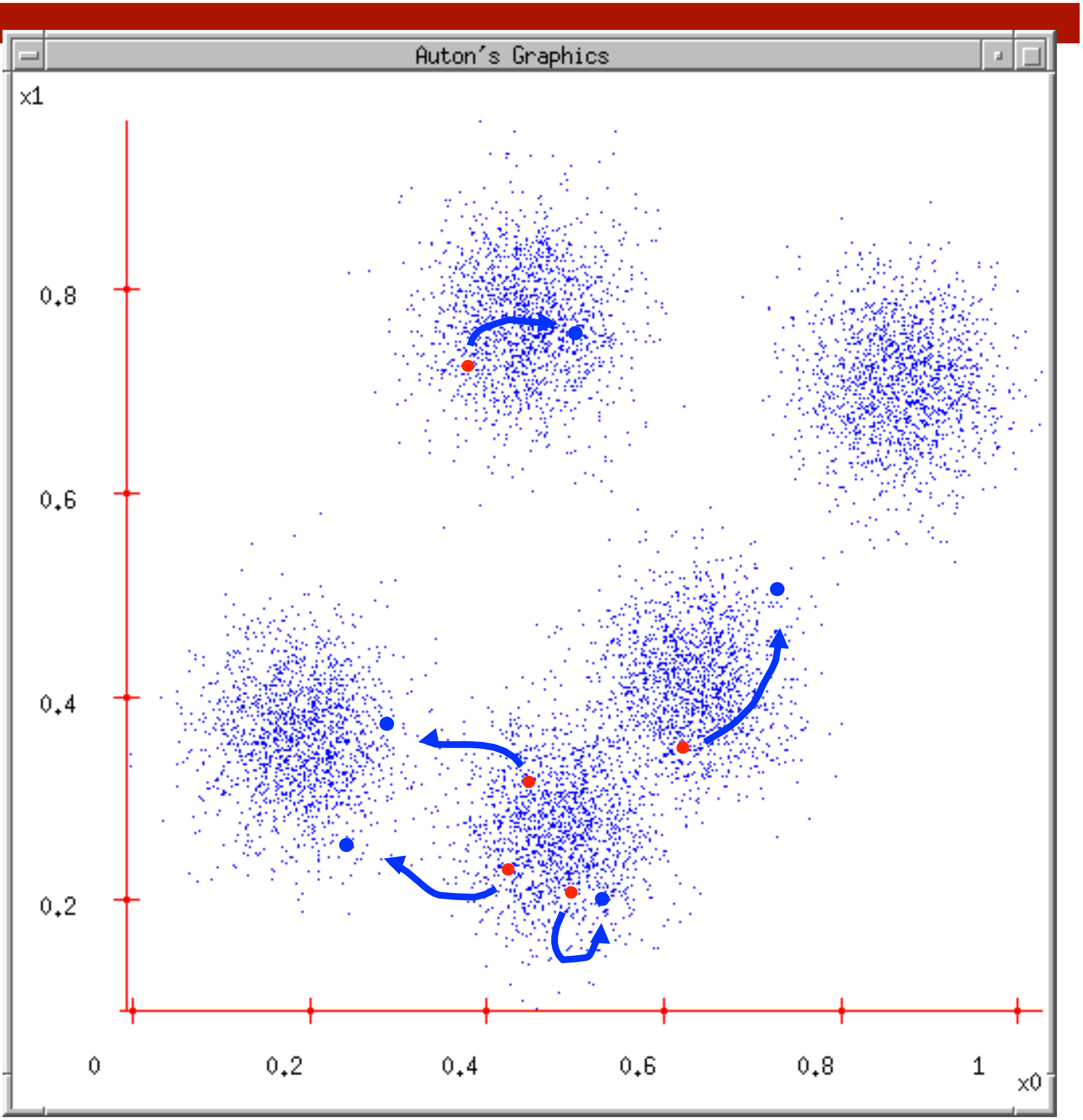
  - What to expect: mix of
    - Multiple Choice or True/False questions
    - “Prove this statement”
    - “What would happen for this dataset?”

  - Material
    - Everything!
    - Focus on the recent stuff.
    - Exponentially decaying weights? Optimal policy?

Recap of Last Time

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!



Slide Credit: Carlos Guestrin

K-means as Co-ordinate Descent

- Optimize objective function:

$$\min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} F(\boldsymbol{\mu}, \boldsymbol{a}) = \min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} \sum_{i=1}^{N} \sum_{j=1}^{k} a_{ij} ||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$$

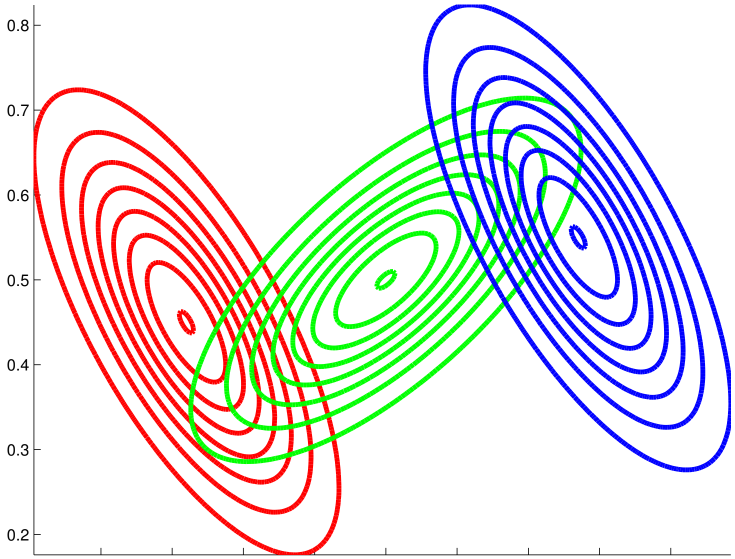
- Fix μ, optimize a (or C)
  - Assignment Step

- Fix a (or C), optimize μ
  - Recenter Step

(C) Dhruv Batra

Slide Credit: Carlos Guestrin

GMM



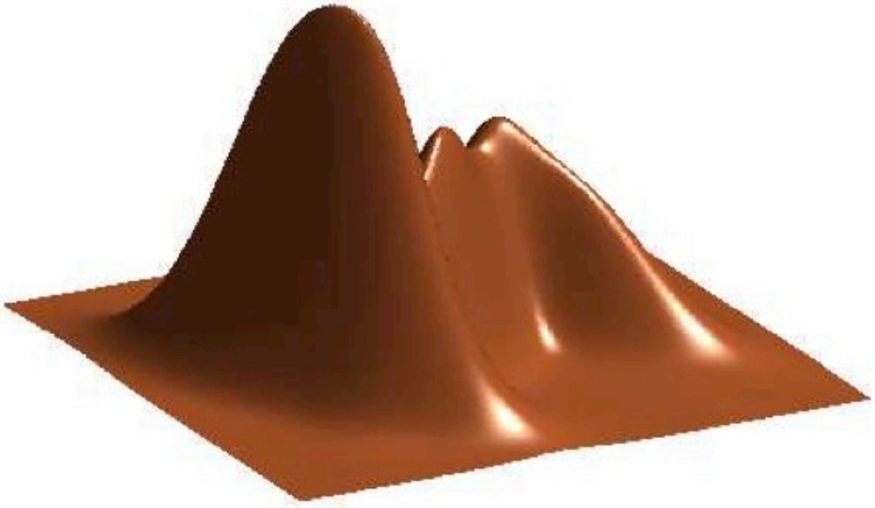


Figure Credit: Kevin Murphy

K-means vs GMM

- K-Means
  - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- GMM
  - http://www.socr.ucla.edu/applets.dir/mixtureem.html

EM

- Expectation Maximization [Dempster ‘77]

- Often looks like “soft” K-means

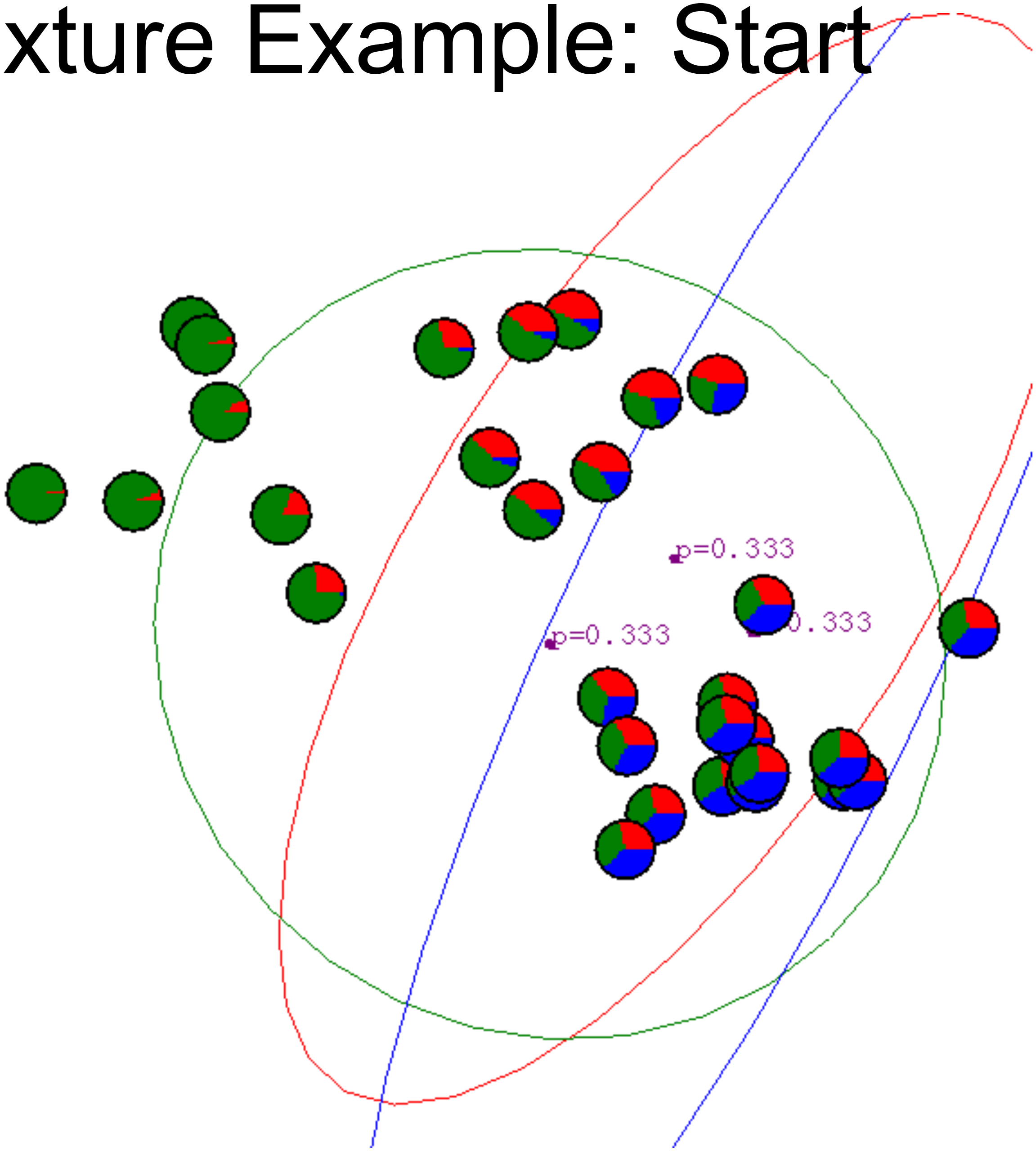
- Extremely general
- Extremely useful algorithm
  - Essentially THE goto algorithm for unsupervised learning

- Plan
  - EM for learning GMM parameters
  - EM for general unsupervised learning problems

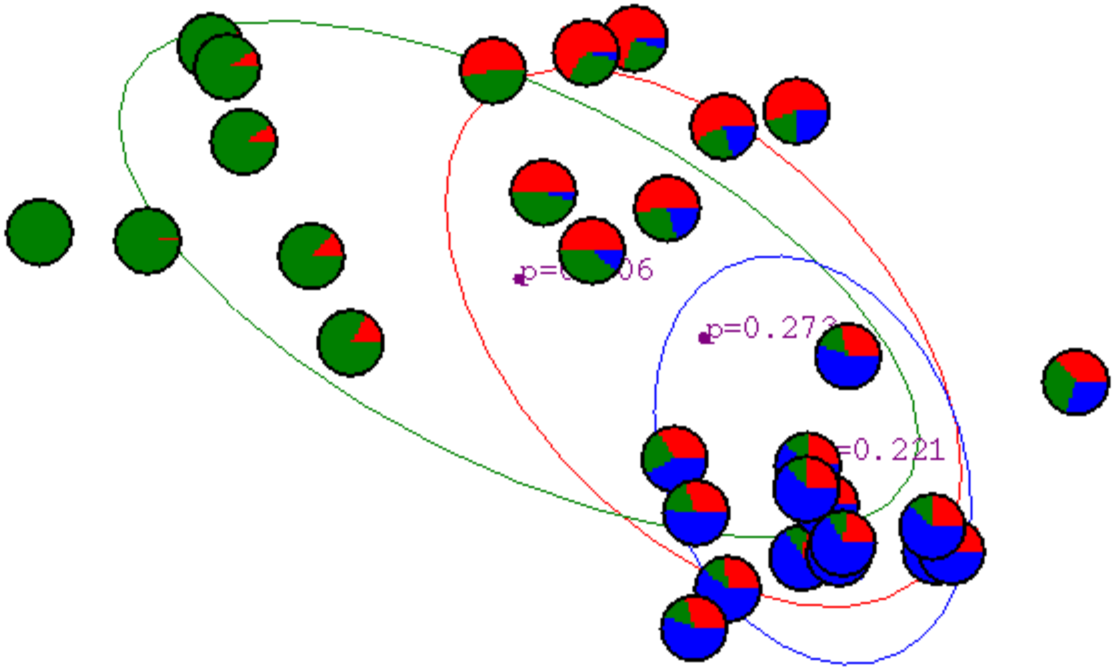
EM for Learning GMMs

- Simple Update Rules
  - E-Step: estimate $P(z_i = j \mid x_i)$
  - M-Step: maximize full likelihood weighted by posterior

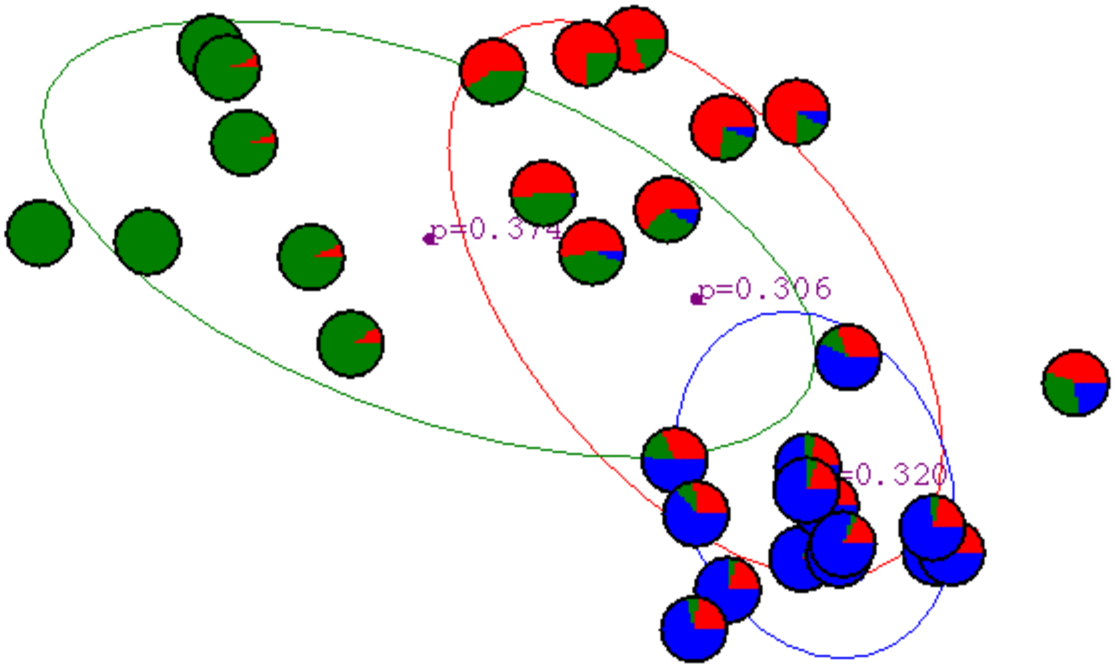
Gaussian Mixture Example: Start



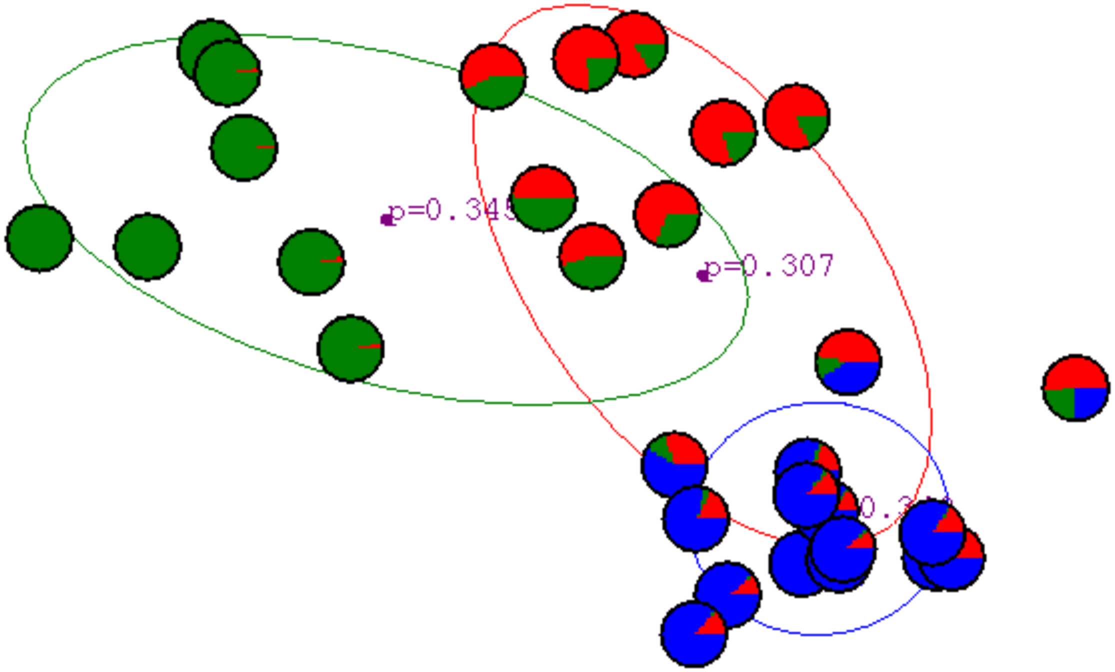
After 1st iteration



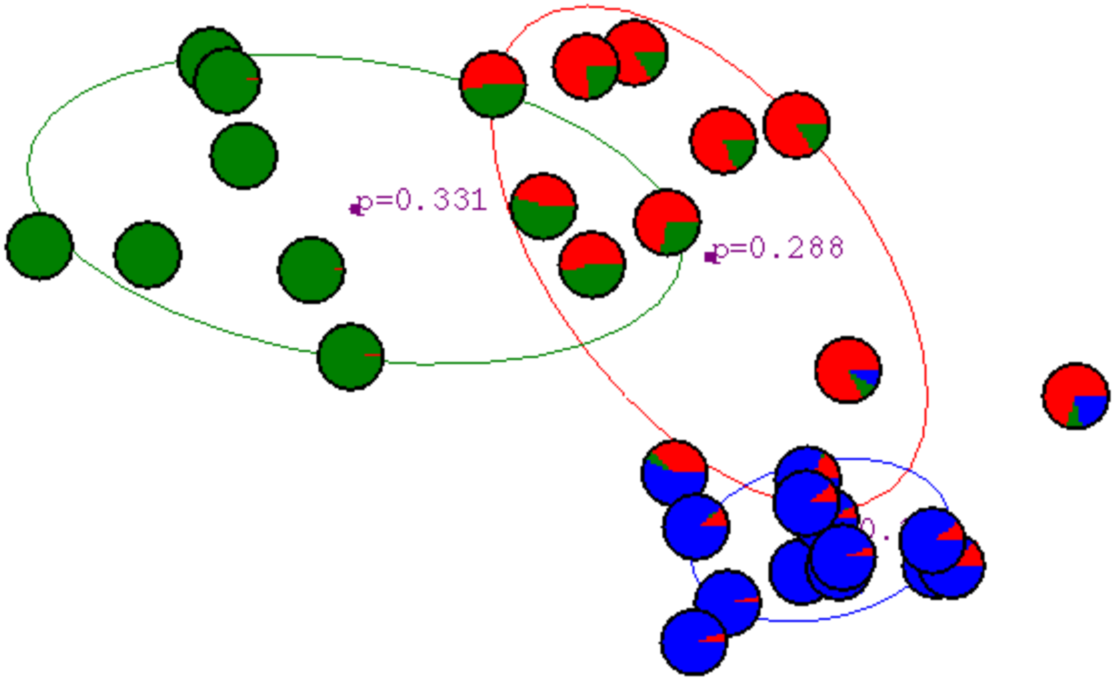
After 2nd iteration



After 3rd iteration

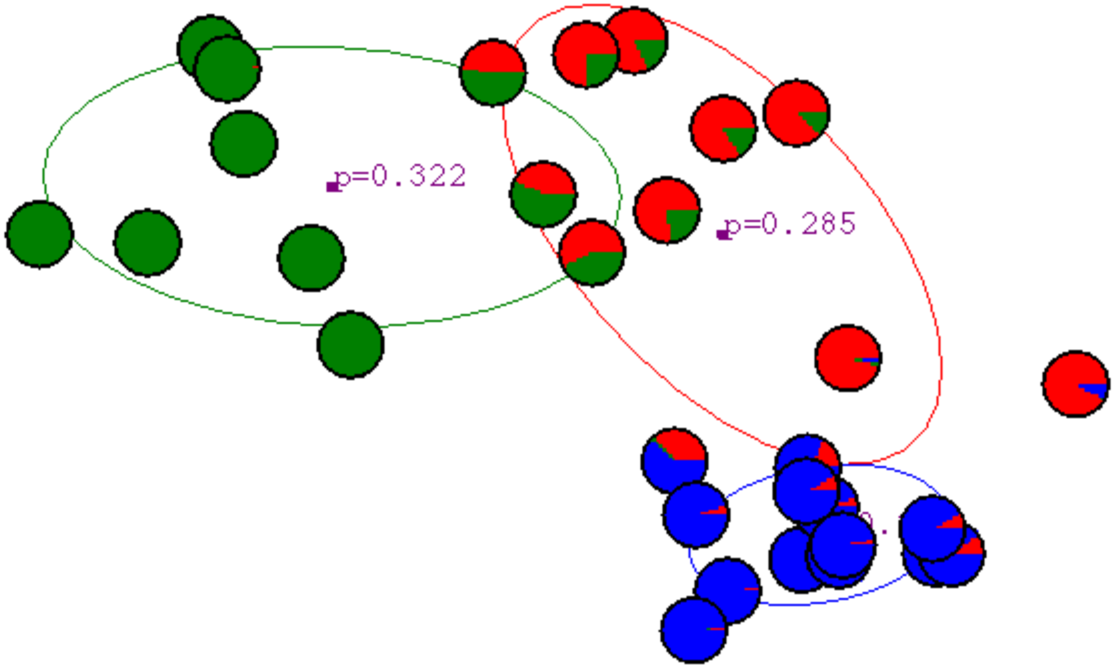


After 4th iteration



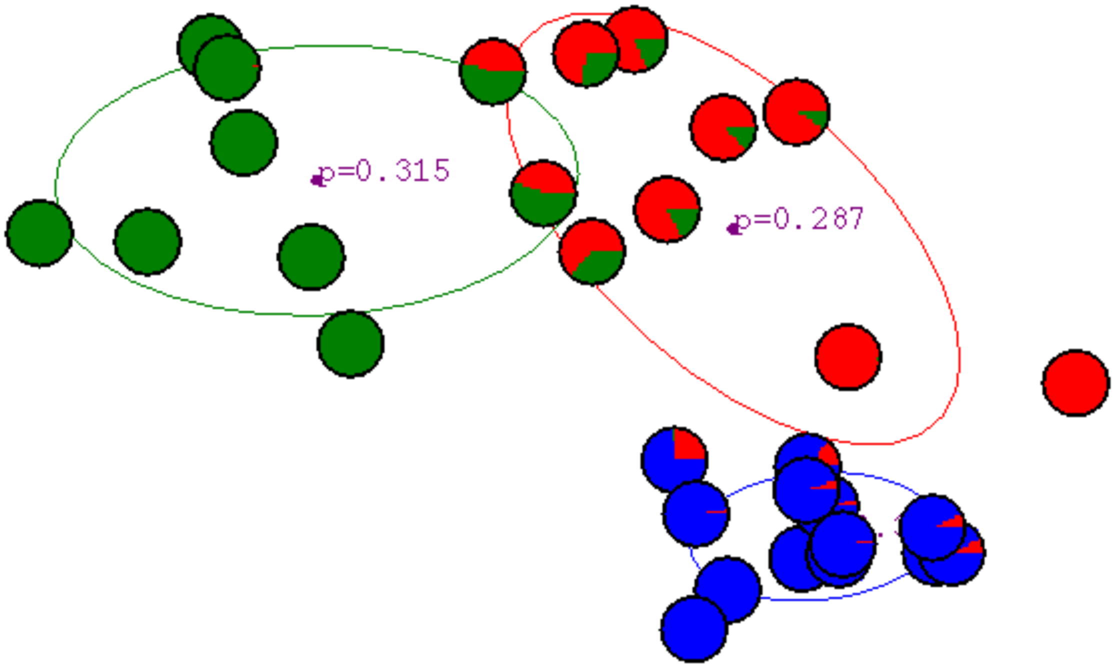
Slide Credit: Carlos Guestrin

After 5th iteration



Slide Credit: Carlos Guestrin

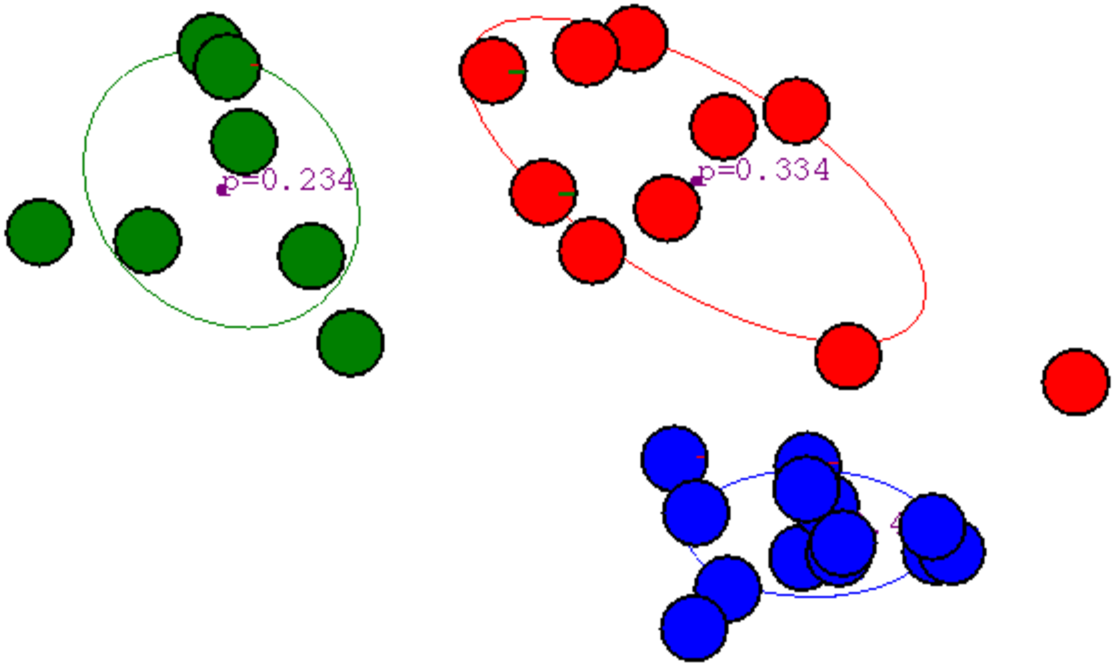
After 6th iteration



(C) Dhruv Batra

Slide Credit: Carlos Guestrin

After 20th iteration

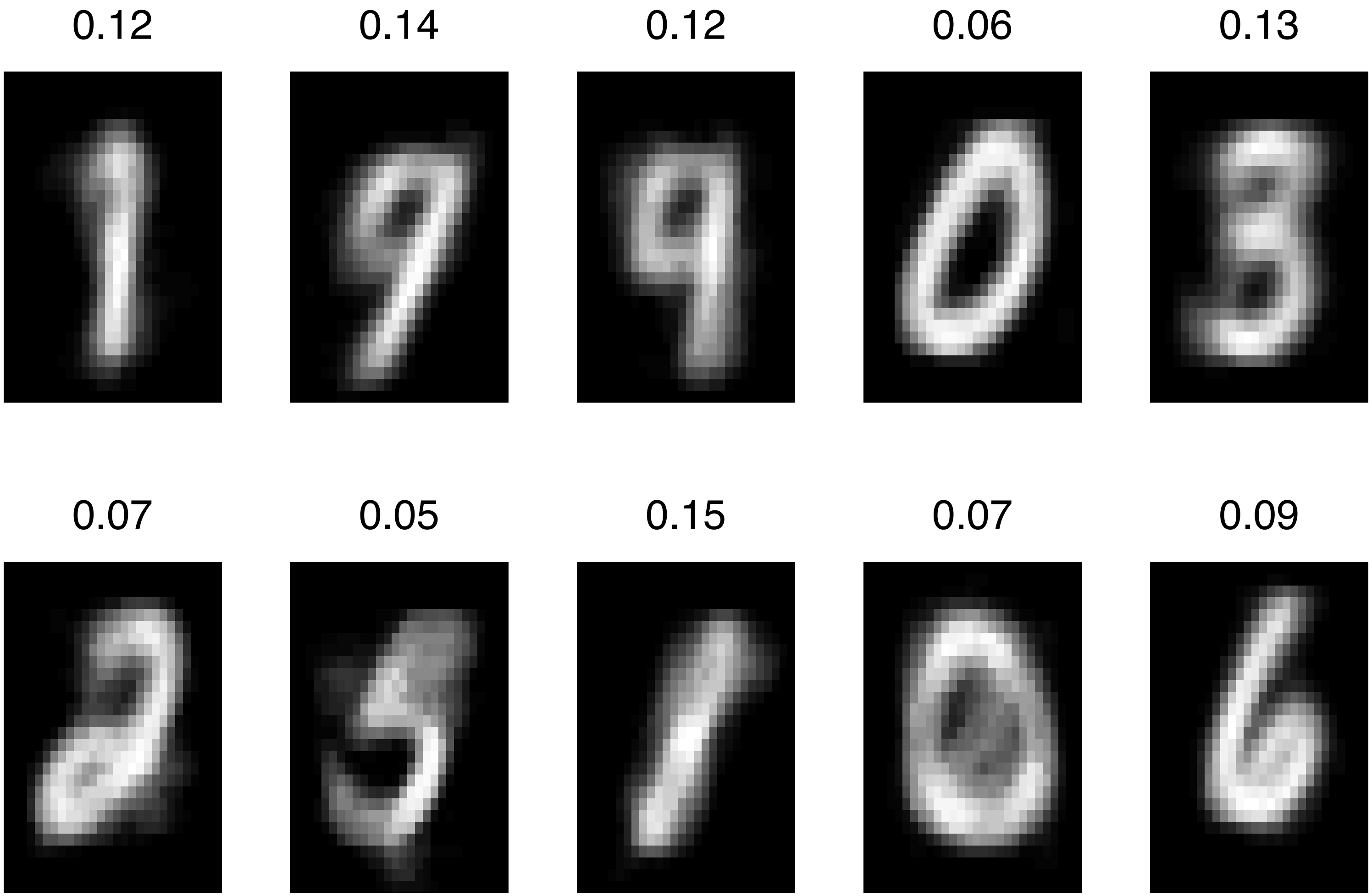


Slide Credit: Carlos Guestrin

General Mixture Models

$P(x \mid z)$	$P(z)$	Name
Gaussian	Categorial	GMM
Multinomial	Categorical	Mixture of Multinomials
Categorical	Dirichlet	Latent Dirichlet Allocation

Mixture of Bernoullis



The general learning problem with missing data

- Marginal likelihood – **x** is observed, **z** is missing:

$$\begin{aligned} ll(\theta : \mathcal{D}) &= \log \prod_{i=1}^{N} P(\mathbf{x}_i \mid \theta) \\ &= \sum_{i=1}^{N} \log P(\mathbf{x}_i \mid \theta) \\ &= \sum_{i=1}^{N} \log \sum_{\mathbf{z}} P(\mathbf{x}_i, \mathbf{z} \mid \theta) \end{aligned}$$

Applying Jensen's inequality

- Use: log $\sum_{\mathbf{z}}$ P($\mathbf{z}$) f($\mathbf{z}$) ≥ $\sum_{\mathbf{z}}$ P($\mathbf{z}$) log f($\mathbf{z}$)

$$ll(\theta : \mathcal{D}) = \sum_{i=1}^{N} \log \sum_{\mathbf{z}} Q_i(\mathbf{z}) \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

Convergence of EM

- Define potential function F(θ,Q):

$$ll(\theta : \mathcal{D}) \;\geq\; F(\theta, Q_i) \;=\; \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- EM corresponds to coordinate ascent on F
  - Thus, maximizes lower bound on marginal log likelihood

EM is coordinate ascent

$$ll(\theta : \mathcal{D}) \geq F(\theta, Q_i) = \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- **E-step**: Fix $\theta^{(t)}$, maximize F over Q:

$$\begin{aligned}
F(\theta, Q_i) &= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{z} \mid \mathbf{x}_i, \theta) P(\mathbf{x}_i \mid \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log P(\mathbf{x}_i \mid \theta^{(t)}) + \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= ll(\theta^{(t)} : \mathcal{D}) - \sum_{i=1}^{N} KL(Q_i(\mathbf{z}) || P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})
\end{aligned}$$

  - “Realigns” F with likelihood:

$$Q_i^{(t)}(\mathbf{z}) = P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})$$

$$F(\theta^{(t)}, Q^{(t)}) = ll(\theta^{(t)} : \mathcal{D})$$

EM is coordinate ascent

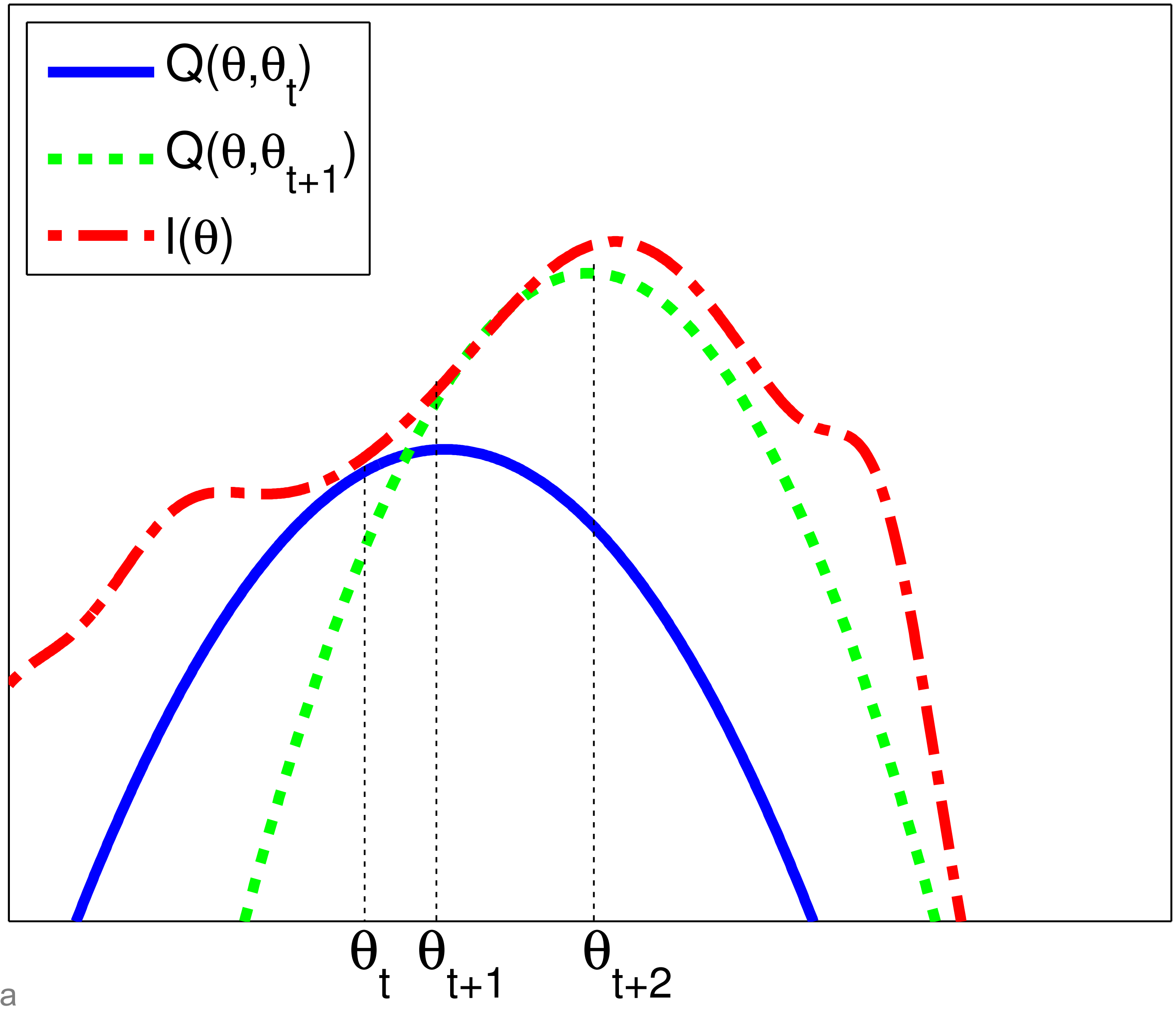
$$ll(\theta : \mathcal{D}) \;\geq\; F(\theta, Q_i) \;=\; \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- **M-step**: Fix $Q^{(t)}$, maximize F over θ

$$\begin{aligned}
F(\theta, Q_i) &= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i^{(t)}(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log P(\mathbf{x}_i, \mathbf{z} \mid \theta) - \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log Q_i^{(t)}(\mathbf{z}) \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log P(\mathbf{x}_i, \mathbf{z} \mid \theta) + \sum_{i=1}^{N} \underbrace{H(Q_i^{(t)})}_{\text{constant}}
\end{aligned}$$

- Corresponds to weighted dataset:
  - $<\mathbf{x}_1, \mathbf{z}=1>$ with weight $Q^{(t+1)}(\mathbf{z}=1|\mathbf{x}_1)$
  - $<\mathbf{x}_1, \mathbf{z}=2>$ with weight $Q^{(t+1)}(\mathbf{z}=2|\mathbf{x}_1)$
  - $<\mathbf{x}_1, \mathbf{z}=3>$ with weight $Q^{(t+1)}(\mathbf{z}=3|\mathbf{x}_1)$
  - $<\mathbf{x}_2, \mathbf{z}=1>$ with weight $Q^{(t+1)}(\mathbf{z}=1|\mathbf{x}_2)$
  - $<\mathbf{x}_2, \mathbf{z}=2>$ with weight $Q^{(t+1)}(\mathbf{z}=2|\mathbf{x}_2)$
  - $<\mathbf{x}_2, \mathbf{z}=3>$ with weight $Q^{(t+1)}(\mathbf{z}=3|\mathbf{x}_2)$
  - …

EM Intuition



What you should know

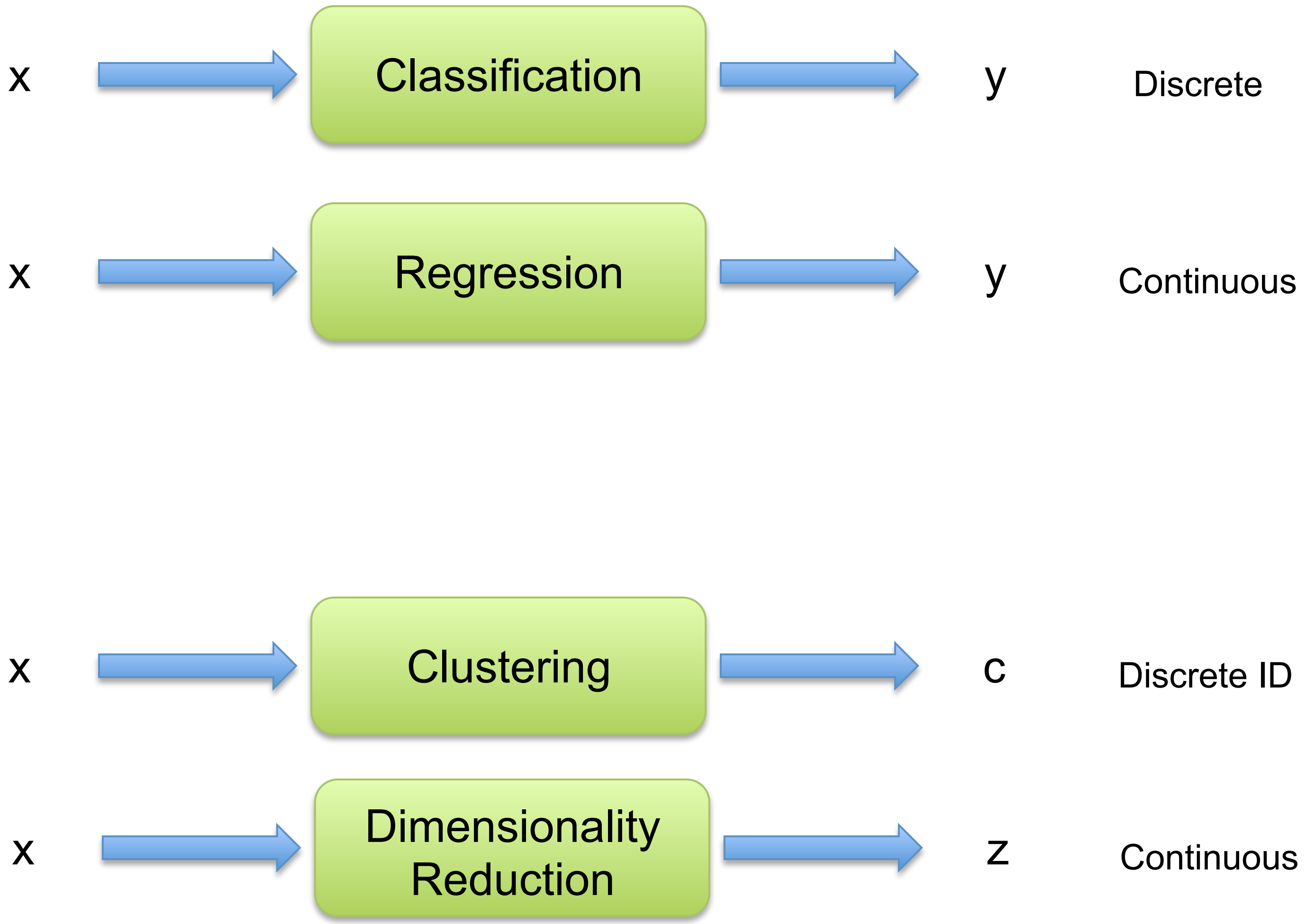
- K-means for clustering:
  - algorithm
  - converges because it's coordinate ascent

- EM for mixture of Gaussians:
  - How to "learn" maximum likelihood parameters (locally max. like.) in the case of unlabeled data

- EM is coordinate ascent
- Remember, E.M. can get stuck in local minima, and empirically it DOES

- General case for EM

Tasks



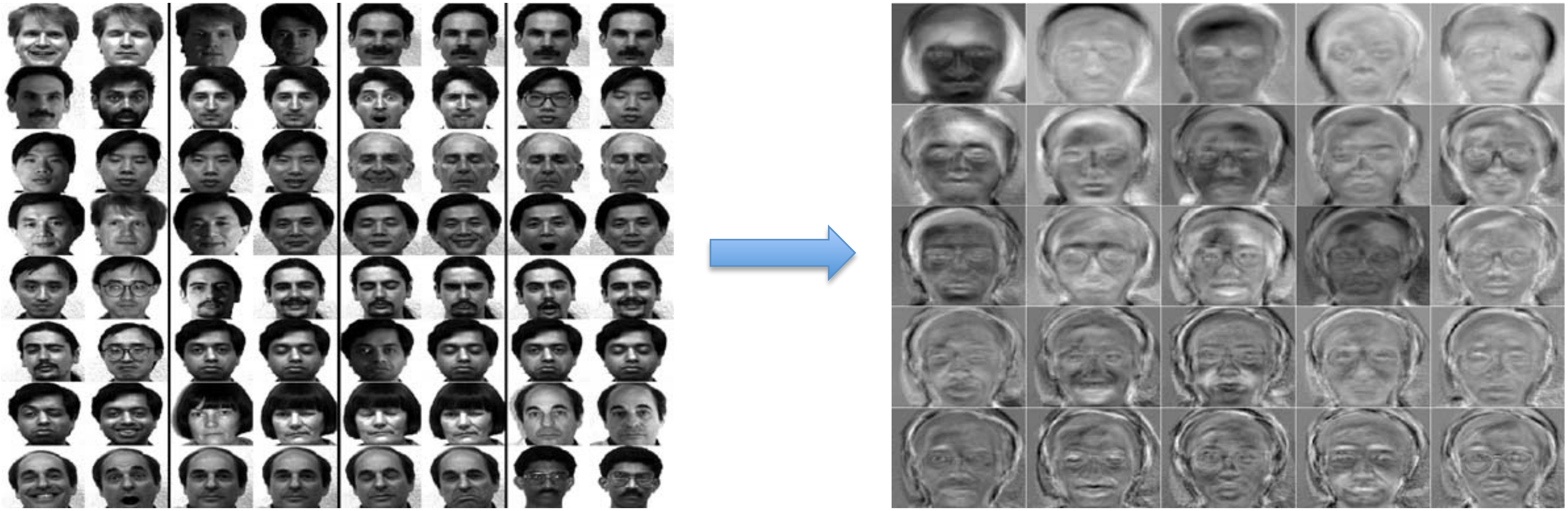
x → Classification → y Discrete

x → Regression → y Continuous

x → Clustering → c Discrete ID

x → Dimensionality Reduction → z Continuous

New Topic: PCA



Synonyms

- Principal Component Analysis
- Karhunen–Loève transform

- Eigen-Faces
- Eigen-<Insert-your-problem-domain>

- PCA is a Dimensionality Reduction Algorithm

- Other Dimensionality Reduction algorithms
  - Linear Discriminant Analysis (LDA)
  - Independent Component Analysis (ICA)
  - Local Linear Embedding (LLE)
  - …

Dimensionality reduction

- Input data may have thousands or millions of dimensions!
  - e.g., images have 5M pixels

- **Dimensionality reduction**:
  represent data with fewer dimensions
  - easier learning – fewer parameters
  - visualization – hard to visualize more than 3D or 4D
  - discover "intrinsic dimensionality" of data
    - high dimensional data that is truly lower dimensional

Dimensionality Reduction

- Demo
  - http://lcn.epfl.ch/tutorial/english/pca/html/
  - http://setosa.io/ev/principal-component-analysis/

PCA / KL-Transform

- De-correlation view
  - Make features uncorrelated
  - No projection yet

- Max-variance view:
  - Project data to lower dimensions
  - Maximize variance in lower dimensions

- Synthesis / Min-error view:
  - Project data to lower dimensions
  - Minimize reconstruction error

- All views lead to same solution