Clustering
K-means

Machine Learning – CSE546
Sham Kakade
University of Washington

November 15, 2016

Announcements:

- Project Milestones due date passed.
- HW3 due on Monday
  - It'll be collaborative
- HW2 grades posted today
  - Out of 82 points

- Today:
  - Review: PCA
  - Start: unsupervised learning

Dimensionality Reduction PCA

Machine Learning – CSE546
Sham Kakade
University of Washington

November 15, 2016

Linear projections, a review

- Project a point into a (lower dimensional) space:
  - **point**: $\mathbf{x} = (x_1,\ldots,x_d)$
  - **select a basis** – set of basis vectors – $(\mathbf{u}_1,\ldots,\mathbf{u}_k)$
    - we consider orthonormal basis:
      - $\mathbf{u}_i \bullet \mathbf{u}_i = 1$, and $\mathbf{u}_i \bullet \mathbf{u}_j = 0$ for $i \neq j$
  - **select a center** – $\bar{\mathbf{x}}$, defines offset of space
  - **best coordinates** in lower dimensional space defined by dot-products: $(z_1,\ldots,z_k)$, $z_i = (\mathbf{x}-\bar{\mathbf{x}}) \bullet \mathbf{u}_i$

best reconstruction

$$\hat{X} = \bar{X} + \sum_{i=1}^{K} z_i \vec{u}_i$$

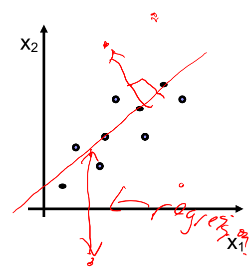
$K \neq d$

$\hat{X} \neq X$

PCA finds projection that minimizes reconstruction error

- Given N data points: $\mathbf{x}^i = (x_1^i,\ldots,x_d^i)$, i=1…N
- Will represent each point as a projection:
  - $\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$ where: $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^i$ and $z_j^i = (\mathbf{x}^i - \bar{\mathbf{x}}) \cdot \mathbf{u}_j$
- PCA:
  - Given k<<d, find $(\mathbf{u}_1,\ldots,\mathbf{u}_k)$ minimizing reconstruction error:

$$error_k = \sum_{i=1}^{N} (\mathbf{x}^i - \hat{\mathbf{x}}^i)^2$$



Understanding the reconstruction error

$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$$

$$z_j^i = (\mathbf{x}^i - \bar{\mathbf{x}}) \cdot \mathbf{u}_j$$

- Given k<<d, find $(\mathbf{u}_1,\ldots,\mathbf{u}_k)$ minimizing reconstruction error:

$$error_k = \sum_{i=1}^{N} (\mathbf{x}^i - \hat{\mathbf{x}}^i)^2$$

- Note that $\mathbf{x}^i$ can be represented exactly by d-dimensional projection:

$$\mathbf{x}^i = \bar{\mathbf{x}} + \sum_{j=1}^{d} z_j^i \mathbf{u}_j$$

full basis: $\underbrace{u_1 \ldots u_k}_{\text{PCA params}}$ $u_{k+1} \ldots u_d$

- Rewriting error:

$$\|x^i - \hat{x}^i\|^2 = \Big\|\sum_{j>k}^{d} z_j^i \vec{u}_j\Big\|^2$$

Reconstruction error and covariance matrix

$$error_k = \sum_{i=1}^{N} \sum_{j=k+1}^{d} [\mathbf{u}_j \cdot (\mathbf{x}^i - \bar{\mathbf{x}})]^2$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^i - \bar{\mathbf{x}})(\mathbf{x}^i - \bar{\mathbf{x}})^T$$

residual error is a function of orthogonal subspace $u^{k+1} \dots u^d$

Minimizing reconstruction error and eigen vectors

- Minimizing reconstruction error equivalent to picking orthonormal basis ($\mathbf{u}_1,\dots,\mathbf{u}_d$) minimizing:

$$error_k = N \sum_{j=k+1}^{d} \mathbf{u}_j^T \Sigma \mathbf{u}_j$$

- Eigen vector definition:

u, λ s.t. $M\vec{u} = \lambda \vec{u}$

- Solution: use the eigenvectors from the SVD

$\Sigma = U \, \mathrm{Diag} \, U^T$ columns of $U = [u_1 | u_2 | \dots u_d]$ are PCA solution.

Clustering
K-means

Machine Learning – CSE546
Sham Kakade
University of Washington

November 14, 2016

Clustering images

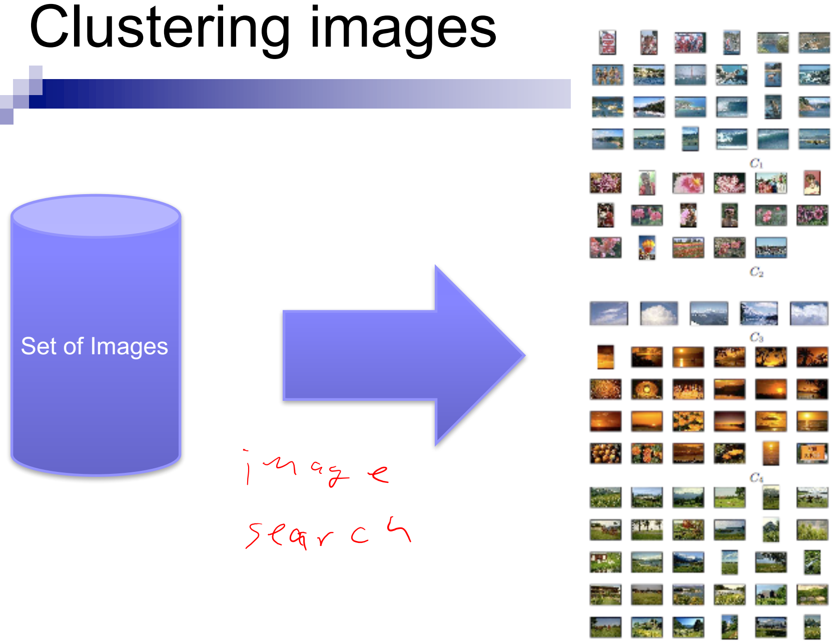
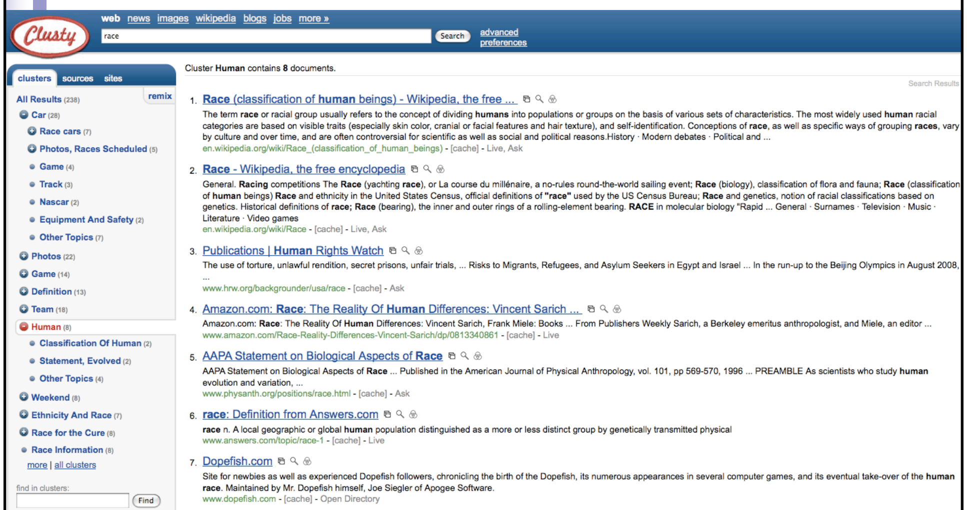


image search

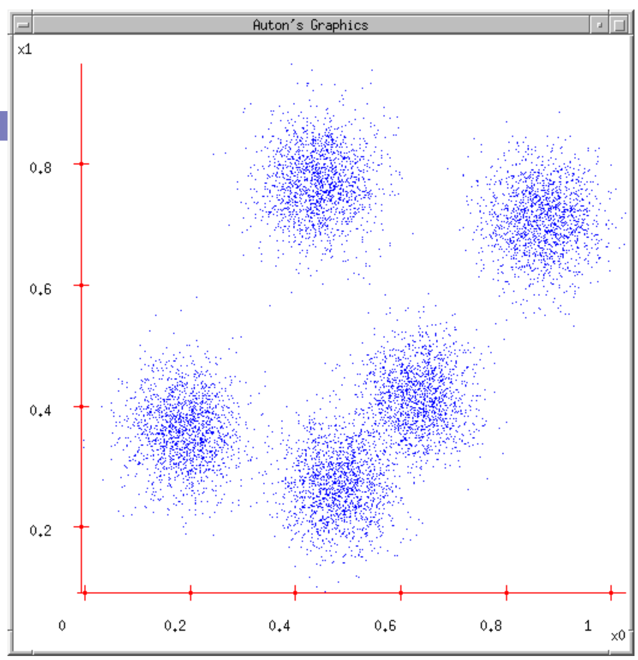
[Goldberger et al.]

Clustering web search results



©Sham Kakade 2016

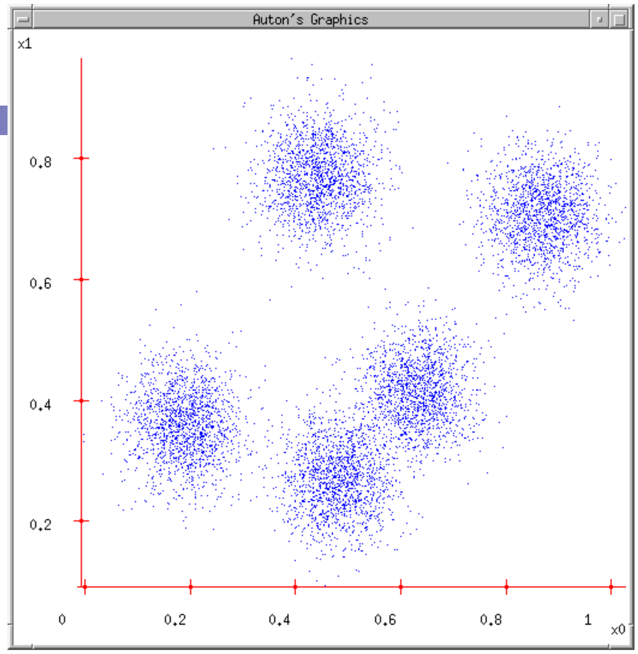
Some Data



©Sham Kakade 2016

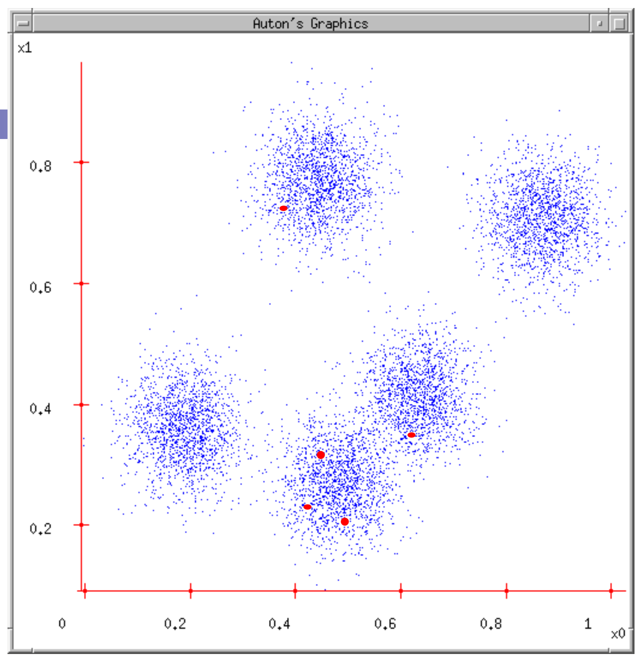
K-means

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



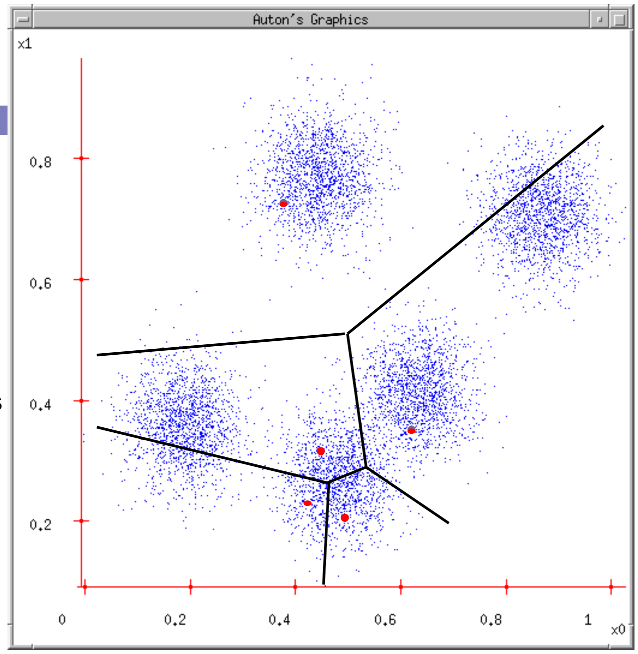
K-means

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



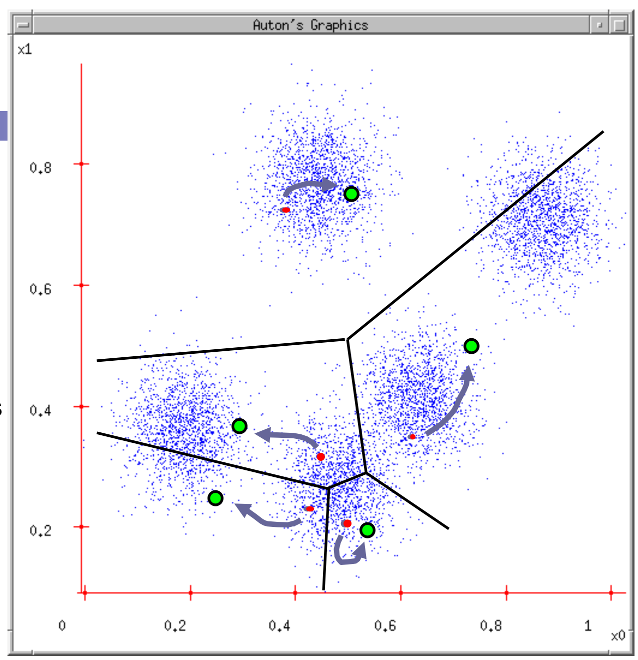
K-means

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)



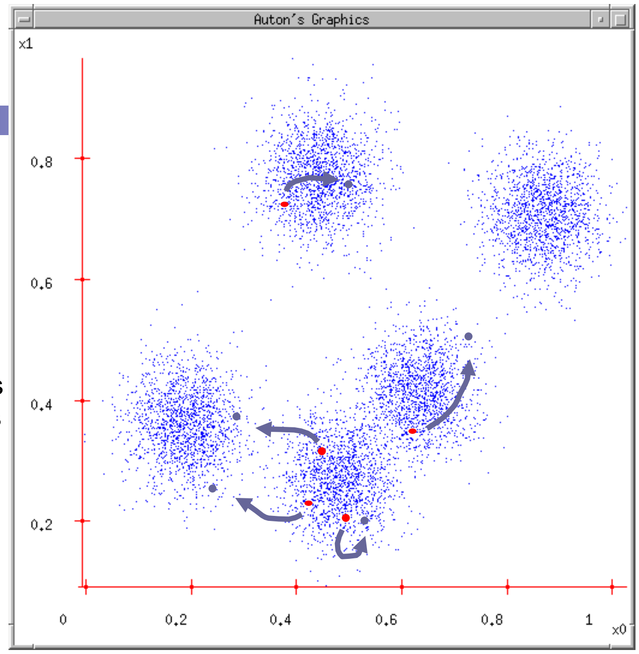
K-means

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



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!



K-means

$\in \mathbb{R}^d$ k-means ++

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

What is K-means optimizing?

- Potential function F(μ,C) of centers μ and point allocations C:
  - $F(\mu, C) = \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2$

means

$\mu_1 \cdots \mu_k$

any assignment function

- 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

$$\min_{C(1) \dots C(j) \dots C(N)} \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2$$

$$= \sum_{j} \min_{C(j)} ||\mu_{C(j)} - x_j||^2$$

indep. opt. problems

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

$$\min_{\mu_1 \ldots \mu_k} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

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

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

(suppose $\forall a,b \quad F(a,b) \geq 0$)

- Converges!!!
  - if F is bounded
  - to a (often good??) local optimum
    - (For LASSO it converged to the global optimum, because of convexity)
  - Some theory of quality of local opt…

$\exists F_\infty$ s.t. $F(a_t, b_t) \to F_\infty$

if points "well-separated"

being global

- K-means is a coordinate descent algorithm!