



Clustering K-means

Machine Learning – CSE546

Sham Kakade

University of Washington

November 15, 2016

©2016 Sham Kakade

©Sham Kakade 2016

1

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

©2016 Sham Kakade

2

Dimensionality Reduction PCA

Machine Learning – CSE546

Sham Kakade

University of Washington

November 15, 2016

©2016 Sham Kakade

©Sham Kakade 2016

3

Linear projections, a review

- Project a point into a (lower dimensional) space:
 - **point:** $\mathbf{x} = (x_1, \dots, x_d)$
 - **select a basis** – set of basis vectors – $(\mathbf{u}_1, \dots, \mathbf{u}_k)$
 - 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, \dots, z_k) , $z_i = (\mathbf{x} - \bar{\mathbf{x}}) \bullet \mathbf{u}_i$

best reconstruction
recons

$$\mathbf{x} = \bar{\mathbf{x}} + \sum_{i=1}^k z_i \mathbf{u}_i \quad \rightarrow \quad k \neq d$$

$\bar{\mathbf{x}} \neq \mathbf{x}$

©2016 Sham Kakade

©Sham Kakade 2016

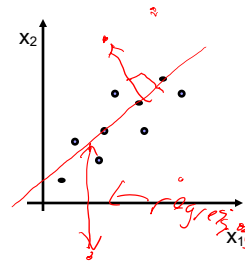
PCA finds projection that minimizes reconstruction error

- Given N data points: $\mathbf{x}^i = (x_1^i, \dots, x_d^i)$, $i=1 \dots N$
- Will represent each point as a projection:

$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^k z_j^i \mathbf{u}_j \quad \text{where: } \bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^i \quad \text{and} \quad z_j^i = (\mathbf{x}^i - \bar{\mathbf{x}}) \cdot \mathbf{u}_j$$

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

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



©2016 Sham Kakade

©Sham Kakade 2016

Understanding the reconstruction error

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

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

basis $\mathbf{u}_1, \dots, \mathbf{u}_k$ $\mathbf{u}_{k+1}, \dots, \mathbf{u}_d$

- Rewriting error:

$$\|\mathbf{x}^i - \hat{\mathbf{x}}^i\|^2 = \left\| \sum_{j>k} z_j^i \mathbf{u}_j \right\|^2$$

PCA terms

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

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

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

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

©2016 Sham Kakade

©Sham Kakade 2016

Reconstruction error and covariance matrix

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

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

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

©2016 Sham Kakade

©Sham Kakade 2016

Minimizing reconstruction error and eigen vectors

- Minimizing reconstruction error equivalent to picking orthonormal basis (u_1, \dots, u_d) minimizing:

$$error_k = \sum_{j=k+1}^d u_j^T \Sigma u_j$$

- Eigen vector definition:

$$u, \lambda \text{ st. } M \vec{u} = \lambda \vec{u}$$

- Solution: use the eigenvectors from the SVD

$$\Sigma = U D \text{diag} U^T \quad U = [u_1 | u_2 | \dots | u_d]$$

columns of U are PCA solution.

©2016 Sham Kakade

©Sham Kakade 2016

Clustering K-means

Machine Learning – CSE546

Sham Kakade

University of Washington

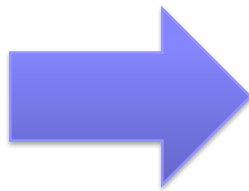
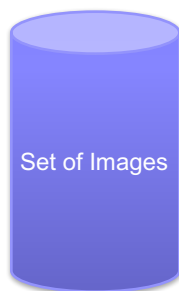
November 14, 2016

©2016 Sham Kakade

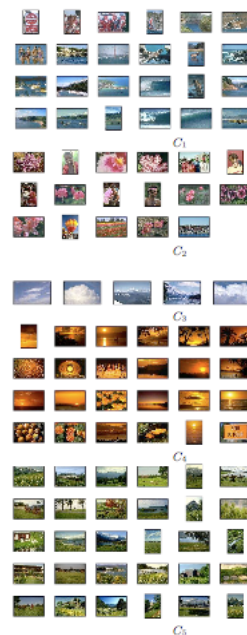
©Sham Kakade 2016

9

Clustering images



*image
search*



©2016 Sham Kakade

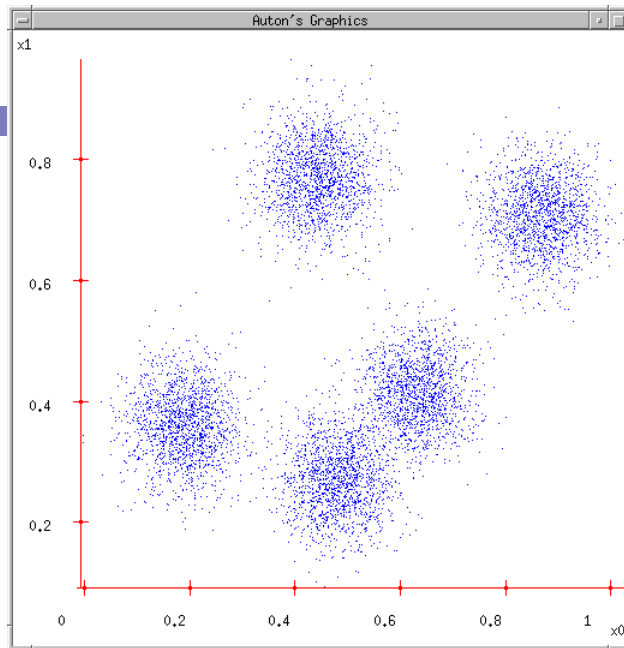
©Sham Kakade 2016

[Goldberger et al.]₁₀

Clustering web search results

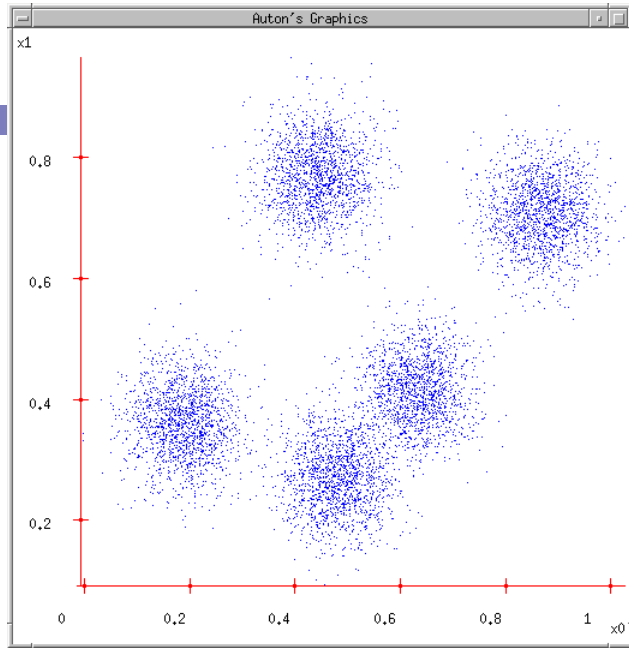
The screenshot shows the Clusty search interface. At the top, there are navigation links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. The search bar contains the word 'race'. Below the search bar, there are tabs for 'clusters', 'sources', and 'sites'. The main content area displays a list of search results, each with a title, a brief description, and a link to the source. The results are clustered around the word 'race'. The first result is 'Race (classification of human beings) - Wikipedia, the free encyclopedia'. The second result is 'Race - Wikipedia, the free encyclopedia'. The third result is 'Publications | Human Rights Watch'. The fourth result is 'Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books...'. The fifth result is 'AAPA Statement on Biological Aspects of Race'. The sixth result is 'race: Definition from Answers.com'. The seventh result is 'Dopefish.com'. On the left side, there is a sidebar with a list of clusters, including 'Car (23)', 'Race cars (7)', 'Photos, Races Scheduled (3)', 'Game (4)', 'Track (3)', 'Nascar (2)', 'Equipment And Safety (2)', 'Other Topics (7)', 'Photos (2)', '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)', and 'Race Information (8)'. At the bottom, there are copyright notices: '©2016 Sham Kakade' and '©Sham Kakade 2016', and a page number '11'.

Some Data



K-means

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



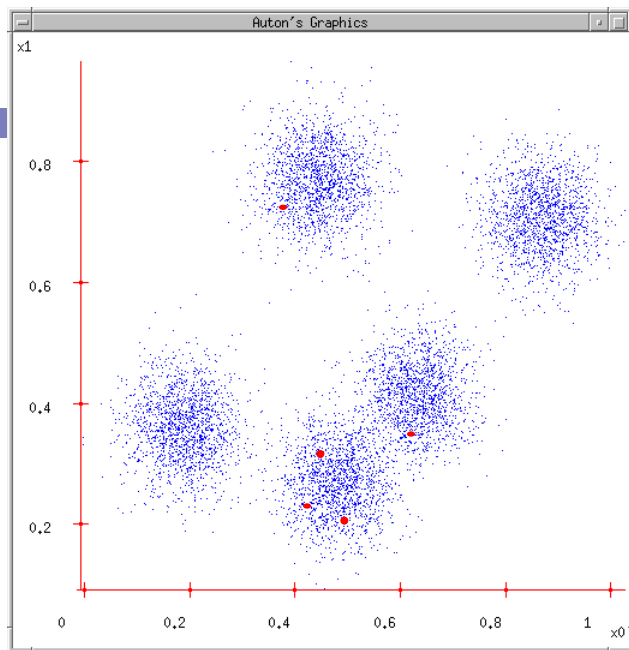
©2016 Sham Kakade

©Sham Kakade 2016

13

K-means

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



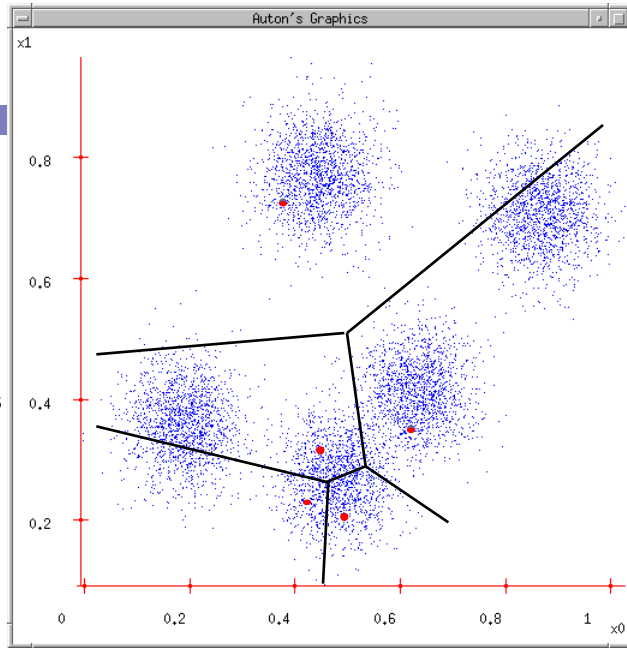
©2016 Sham Kakade

©Sham Kakade 2016

14

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)



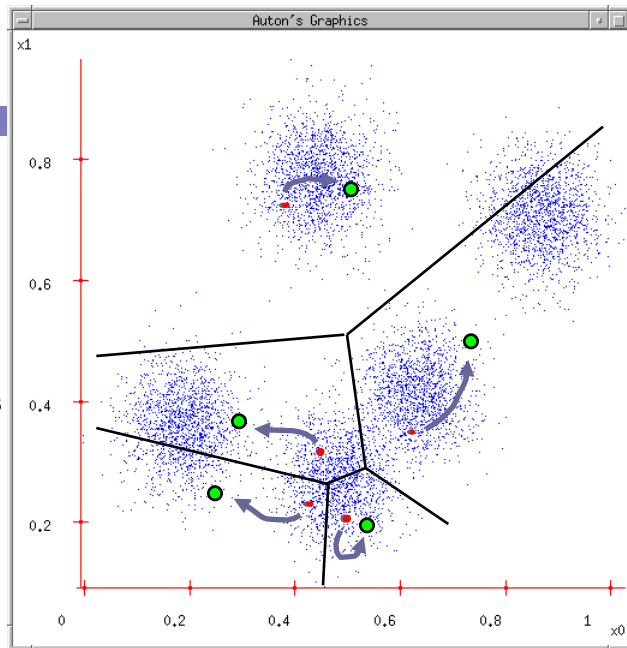
©2016 Sham Kakade

©Sham Kakade 2016

15

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



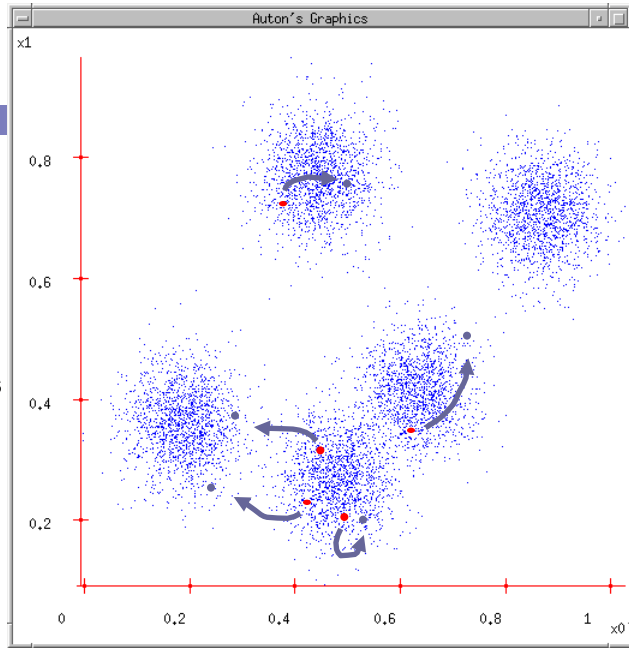
©2016 Sham Kakade

©Sham Kakade 2016

16

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!



©2016 Sham Kakade

©Sham Kakade 2016

17

K-means

$\in \mathbb{R}^d$

K-means $t+1$

- Randomly initialize k centers

□ $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$

*← smartly??
(use points in your dataset?)*

- **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:

□ $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j)=i} \|\mu - x_j\|^2$

$\mu_i^{(t+1)} = \frac{\sum_{j: C(j)=i} x_j}{|\{j: C(j)=i\}|}$

- Equivalent to $\mu_i \leftarrow$ average of its points!

©2016 Sham Kakade

©Sham Kakade 2016

18

What is K-means optimizing?

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

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

means
 μ_1, \dots, μ_k

assignment function

- Optimal K-means:

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

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

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

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

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

$$\begin{aligned} \min_{\mu_1, \dots, \mu_k} \sum_{i=1}^k \sum_{j: C(j)=i} \|\mu_i - x_j\|^2 \\ \sum_{i=1}^k \min_{\mu_i} \sum_{j: C(j)=i} \|\mu_i - x_j\|^2 \end{aligned}$$

©2016 Sham Kakade

©Sham Kakade 2016

21

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
 - (For LASSO it converged to the global optimum, because of convexity)
- Some theory of quality of local opt...

being global

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

$\exists F_{\infty}$ s.t. $F(a_{\epsilon}, b_{\epsilon}) \rightarrow F_{\infty}$

if points

"well-separated"

- K-means is a coordinate descent algorithm!

©2016 Sham Kakade

©Sham Kakade 2016

22