

Shameless plug for my course next quarter

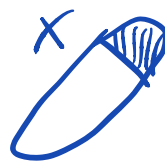
CSE 599: Online and Adaptive Methods for Machine Learning.

Webpage: <https://courses.cs.washington.edu/courses/cse599i/18wi/>

Non-CSE need add-codes: <https://goo.gl/forms/G76D6cOKNtdBlbe62>

The standard approach to machine learning uses a training set of labeled examples to learn a prediction rule that will predict the labels of new examples. Collecting such training sets can be expensive and time-consuming. This course will explore methods that leverage already-collected data to guide future measurements, in a closed loop, to best serve the task at hand. We focus on two paradigms: i) in pure-exploration we desire algorithms that identify or learn a good model using as few measurements as possible (e.g., classification, drug discovery, science), and ii) in regret minimization we desire algorithms that balance taking measurements to learn a model with taking measurements to exploit the model to obtain high reward outcomes (e.g., medical treatment design, ad-serving). The course will assume introductory machine learning (e.g., CSE 546) and maturity in topics like linear algebra, statistics, and calculus. The course will be analysis heavy, with a focus on methods that work well in practice.

Practice



$$XVS^{-1/2}$$



$$XVS^{-1/2}V^T$$



- Fill in the missing plots:

$$\underline{\Sigma} = \mathbf{X}^T \mathbf{J} \mathbf{J} \mathbf{X} = \mathbf{Z}^T \mathbf{J} \mathbf{J} \mathbf{Z}$$

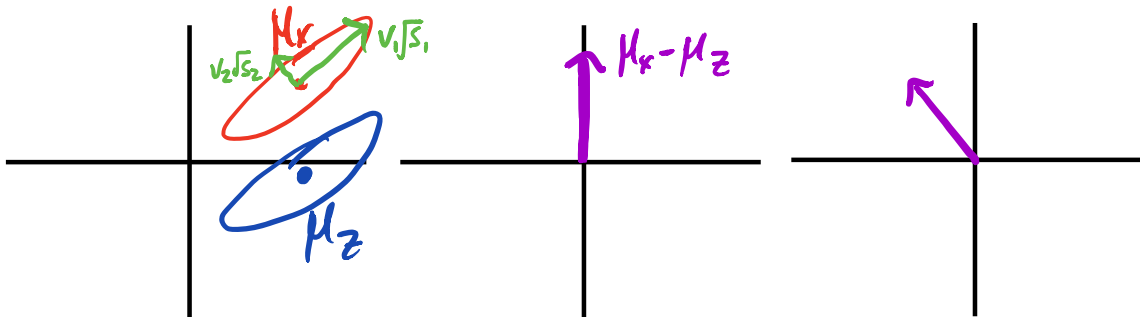
$$\mathbf{V} \mathbf{S} \mathbf{V}^T = \text{eig}(\Sigma) \quad \mathbf{J} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n$$

$$\mu_X = \mathbf{X}^T \mathbf{1}/n \quad \mu_Z = \mathbf{Z}^T \mathbf{1}/n$$

X **Z**

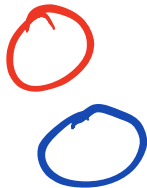
$$\mu_X - \mu_Z$$

$$\mathbf{V} \mathbf{S}^{-1/2} \mathbf{V}^T (\mu_X - \mu_Z) = \hat{\mathbf{w}}$$



$$\hat{f}(x) = \hat{\mathbf{w}}^T x + b$$

Pick b
to minimize
classification
loss





Principal Component Analysis (continued)

Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 16, 2017

Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^N \|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})\|^2.$$

where $\mathbf{V}_q = [v_1, v_2, \dots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

\mathbf{V}_q are the first q eigenvectors of Σ

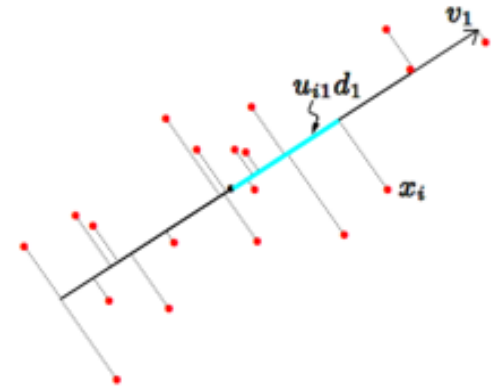
\mathbf{V}_q are the first q principal components

Principal Component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x}^T)$ down onto \mathbf{V}_q

$$(\mathbf{X} - \mathbf{1}\bar{x}^T) \mathbf{V}_q = \mathbf{U}_q \text{diag}(d_1, \dots, d_q) \quad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Singular Value Decomposition defined as

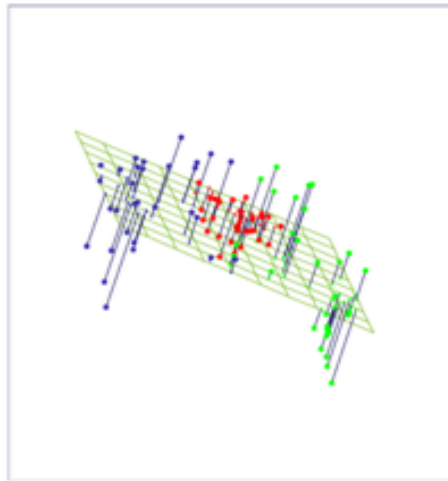
$$\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U} \mathbf{S} \mathbf{V}^T$$



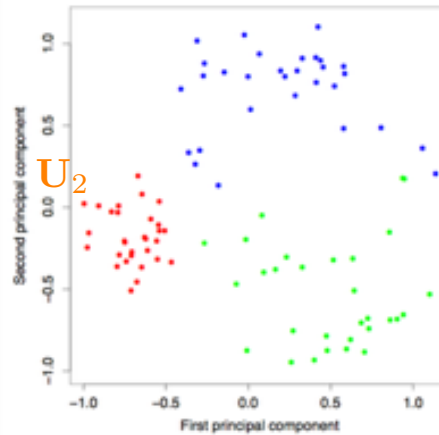
$$\Sigma := \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T$$

Dimensionality reduction

\mathbf{V}_q are the first q eigenvectors of Σ and SVD $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$



$$\mathbf{X} - \mathbf{1}\bar{x}^T$$



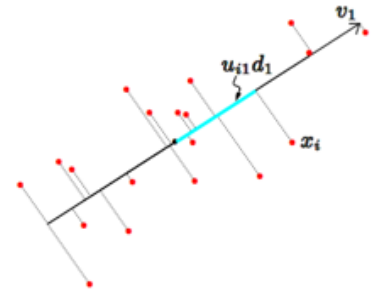
$$\mathbf{U}_1$$

$$\mathbf{U}_2$$

Power method - one at a time

$$\Sigma := \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T \quad v_* = \arg \max_v v^T \Sigma v$$

$$v_{k+1} = \frac{\Sigma v_k}{\|\Sigma v_k\|} \quad v_0 \sim \mathcal{N}(0, I)$$



Matrix completion

Given historical data on how users rated movies in past:



17,700 movies, 480,189 users, 99,072,112 ratings

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for \$1 million prize)

						...
Alice	1	?	?	4	?	
Bob	?	2	5	?	?	
Carol	?	?	4	5	?	
Dave	5	?	?	?	4	
⋮						

Matrix completion

Choose λ by cross-valid.

σ^2 is chosen by test loss

$$l(U, V)$$

$$X \in \mathbb{R}^{m \times n}$$

n movies, m users, |S| ratings

$$U_t^{(i,j)} \sim \mathcal{N}(0, \sigma^2)$$

$$V_t^{(i,j)} \sim \mathcal{N}(0, \sigma^2)$$

$$\arg \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}}$$

$$\sum_{(i,j,s) \in S} \|(UV^T)_{i,j} - s_{i,j}\|_2^2$$

SGD: How do we solve it? With full information? Sample (i,j,s) from training set

$$U_{t+1} = U_t - \nabla l_{ijs}(U_t, V_t) \eta_t$$

$$V_{t+1} = V_t - \nabla l_{ijs}(U_t, V_t) \eta_t$$

$$l_{ijs} = (\langle U_i, V_j \rangle - s_{ijs})^2$$

Alternating Min.

$$U_{t+1} = \arg \min_U l(U, V_t)$$

$$V_{t+1} = \arg \min_V l(U_{t+1}, V)$$

$$X = \tilde{U} \tilde{S} \tilde{V}^T$$

$$l(U, V) = \|UV^T - X\|_F^2$$

$$= \text{Tr}((UV^T - X)^T (UV^T - X))$$

$$\propto -\text{Tr}(U^T X V)$$

Matrix completion

n movies, m users, |S| ratings

$$\arg \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \sum_{(i,j,s) \in \mathcal{S}} \|(UV^T)_{i,j} - s_{i,j}\|_2^2$$

Random projections

PCA finds a low-dimensional representation that reduces population variance

$$\min_{\mathbf{V}_q} \sum_{i=1}^N \|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})\|^2.$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

\mathbf{V}_q are the first q eigenvectors of Σ

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

But what if I care about the reconstruction of the *individual* points?

$$\min_{\mathbf{W}_q} \max_{i=1, \dots, n} \|(x_i - \bar{x}) - \mathbf{W}_q \mathbf{W}_q^T (x_i - \bar{x})\|^2$$

Random projections

$$\min_{\mathbf{W}_q} \max_{i=1, \dots, n} \|(x_i - \bar{x}) - \mathbf{W}_q \mathbf{W}_q^T (x_i - \bar{x})\|^2$$

Johnson-Lindenstrauss (1983)

Theorem 1.1. (Johnson-Lindenstrauss) Let $\epsilon \in (0, 1/2)$. Let $Q \subset \mathbb{R}^d$ be a set of n points and $k = \frac{20 \log n}{\epsilon^2}$. There exists a Lipschitz mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ such that for all $u, v \in Q$:

(independent of d)

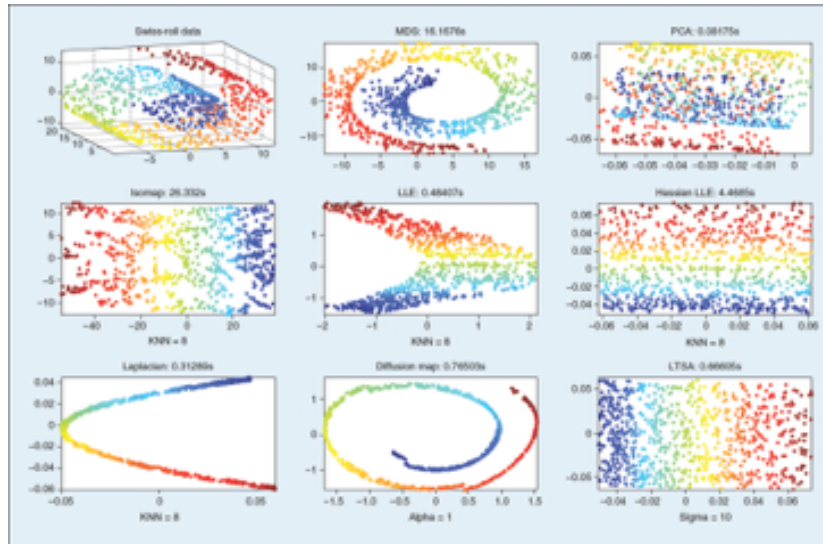
$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

Theorem 1.2. (Norm preservation) Let $x \in \mathbb{R}^d$. Assume that the entries in $A \subset \mathbb{R}^{k \times d}$ are sampled independently from $N(0, 1)$. Then,

$$\Pr\left((1 - \epsilon)\|x\|^2 \leq \left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \leq (1 + \epsilon)\|x\|^2\right) \geq 1 - 2e^{-(\epsilon^2 - \epsilon^3)k/4}$$

Nonlinear dimensionality reduction

Find a low dimensional representation that respects “local distances” in the higher dimensional space



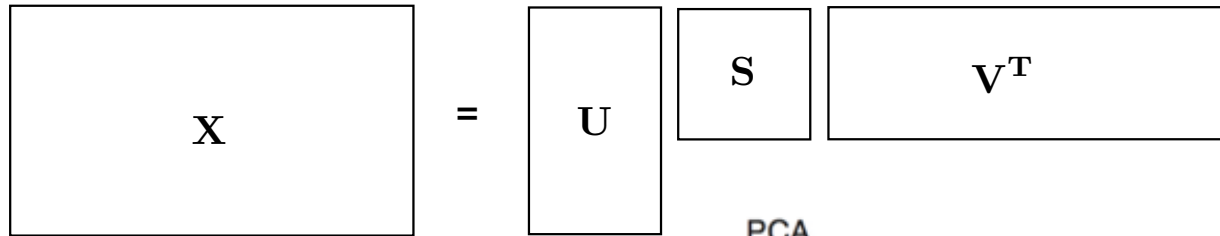
Zhang et al 2010

Many methods:

- Kernel PCA
- ISOMAP
- Local linear embedding
- Maximum volume unfolding
- Non-metric multidimensional scaling
- Laplacian
- Neural network auto encoder
- ...

Due to lack of agreed upon metrics, it is very hard to judge which is best. Also, results from 3 to 2 dims is probably not representative of 1000 to 2 dimensions.

Other matrix factorizations

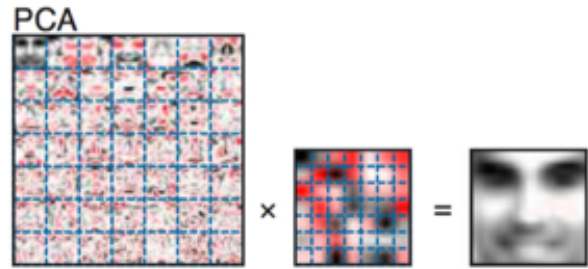


Singular value decomposition

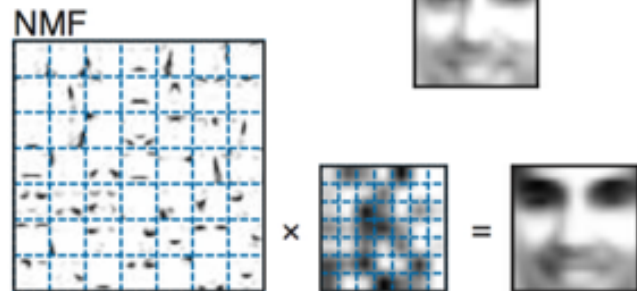
$$\mathbf{U}^T \mathbf{U} = \mathbf{I}, \mathbf{V}^T \mathbf{V} = \mathbf{I}, \mathbf{S} = \text{diag}(s)$$

$$\mathbf{U} \in \mathbb{R}^{n \times q}, \mathbf{V} \in \mathbb{R}^{m \times q}, s \in \mathbb{R}_+^q$$

$$\mathbf{X} \approx \mathbf{U}_q \mathbf{S}_q \mathbf{V}_q^T$$



Original



Nonnegative matrix factorization (NMF)

$$\mathbf{W} \in \mathbb{R}_+^{n \times q} \text{ with } \mathbf{W}\mathbf{1} = \mathbf{1}$$

$$\mathbf{B} \in \mathbb{R}_+^{q \times n} \text{ with } \mathbf{B}\mathbf{1} = \mathbf{1}$$

$$\mathbf{X} \approx \mathbf{W} \mathbf{B} \mathbf{X}$$

H



Clustering

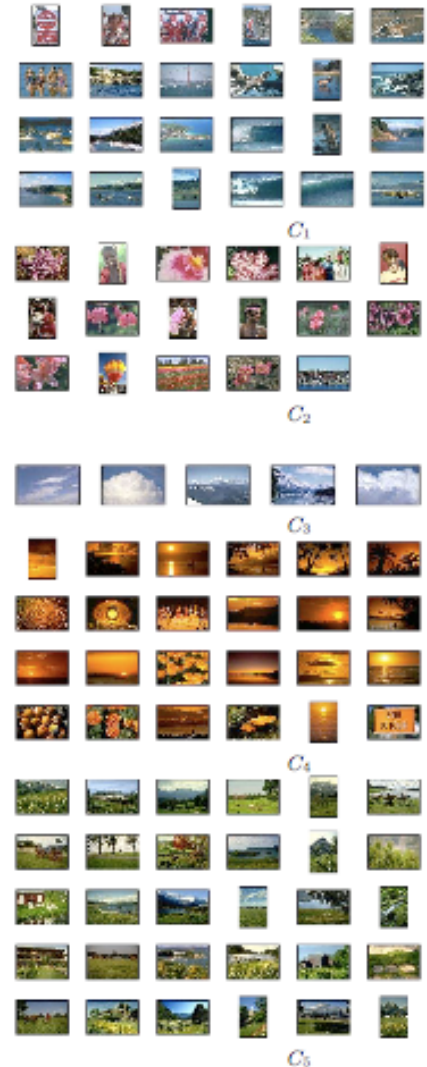
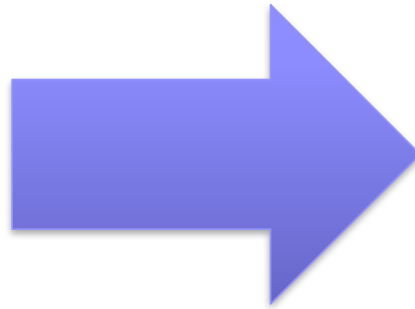
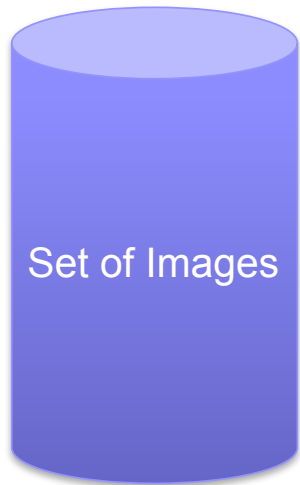
Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 16, 2016

Clustering images



Clustering web search results

The screenshot shows the Clusty search engine interface. At the top, there are navigation links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. The search bar contains the word 'race' and has a 'Search' button and a link to 'advanced preferences'. On the left side, there is a sidebar with 'clusters', 'sources', and 'sites' tabs. Under 'clusters', there are several categories with document counts: Car (20), Race cars (7), Photos, Races Scheduled (10), Game (4), Track (3), Nascar (2), Equipment And Safety (2), Other Topics (7), Photos (22), Game (14), Definition (13), Team (18), and Human (8). The 'Human (8)' category is selected, and it lists sub-categories: Classification Of Human (2), Statement, Evolved (2), and Other Topics (4). Below these are 'Weekend (8)', 'Ethnicity And Race (7)', 'Race for the Cure (8)', and 'Race Information (8)'. At the bottom of the sidebar, there is a 'more | all clusters' link and a search box for finding clusters.

Cluster Human contains 8 documents.

- [Race \(classification of human beings\) - Wikipedia, the free ...](#)
The term **race** or racial group usually refers to the concept of dividing humans into populations or groups on the basis of various sets of characteristics. The most widely used human racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **races**, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History · Modern debates · Political and ...
[en.wikipedia.org/wiki/Race_\(classification_of_human_beings\)](#) - [cache] - Live, Ask
- [Race - Wikipedia, the free encyclopedia](#)
General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of human beings) **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games
[en.wikipedia.org/wiki/Race](#) - [cache] - Live, Ask
- [Publications | Human Rights Watch](#)
The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
[www.hrw.org/background/en/usa/race](#) - [cache] - Ask
- [Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...](#)
Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books** ... From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
[www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861](#) - [cache] - Live
- [AAPA Statement on Biological Aspects of Race](#)
AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study human evolution and variation, ...
[www.physanth.org/positions/race.html](#) - [cache] - Ask
- [race: Definition from Answers.com](#)
race n. A local geographic or global human population distinguished as a more or less distinct group by genetically transmitted physical
[www.answers.com/topic/race-1](#) - [cache] - Live
- [Dopefish.com](#)
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the human **race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
[www.dopefish.com](#) - [cache] - Open Directory

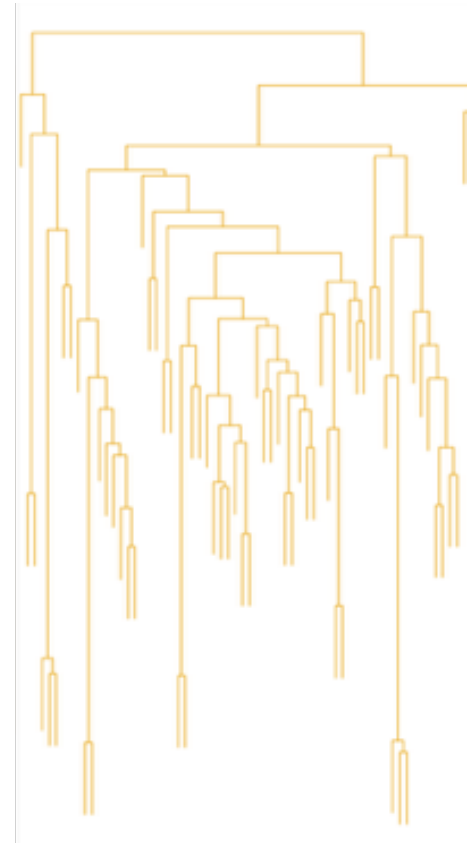
Hierarchical Clustering

Pick one:

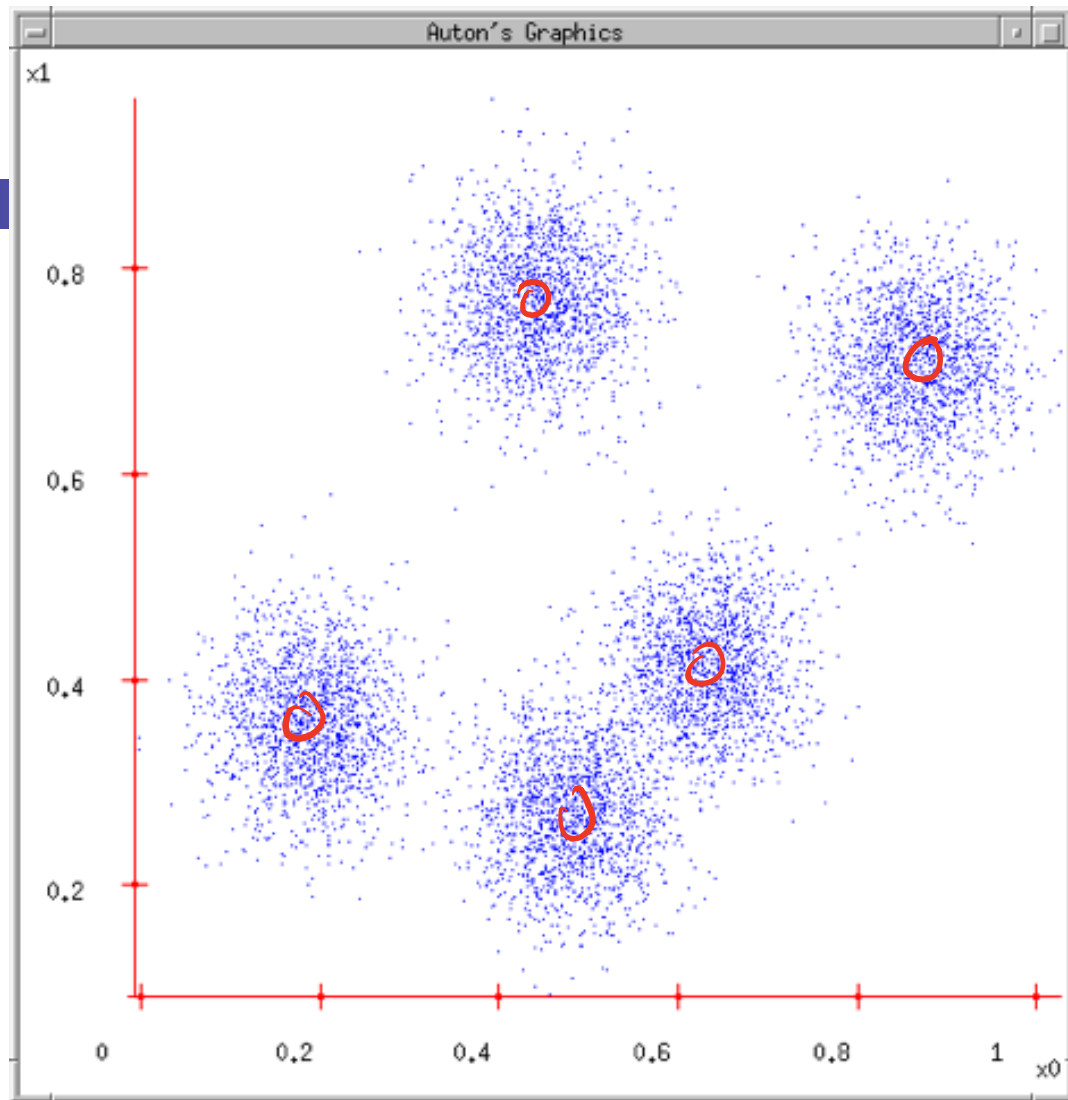
- Bottom up: start with every point as a cluster and merge
- Top down: start with a single cluster containing all points and split

Different rules for splitting/merging, no “right answer”

Gives apparently interpretable tree representation. However, warning: even random data with no structure will produce a tree that “appears” to be structured.

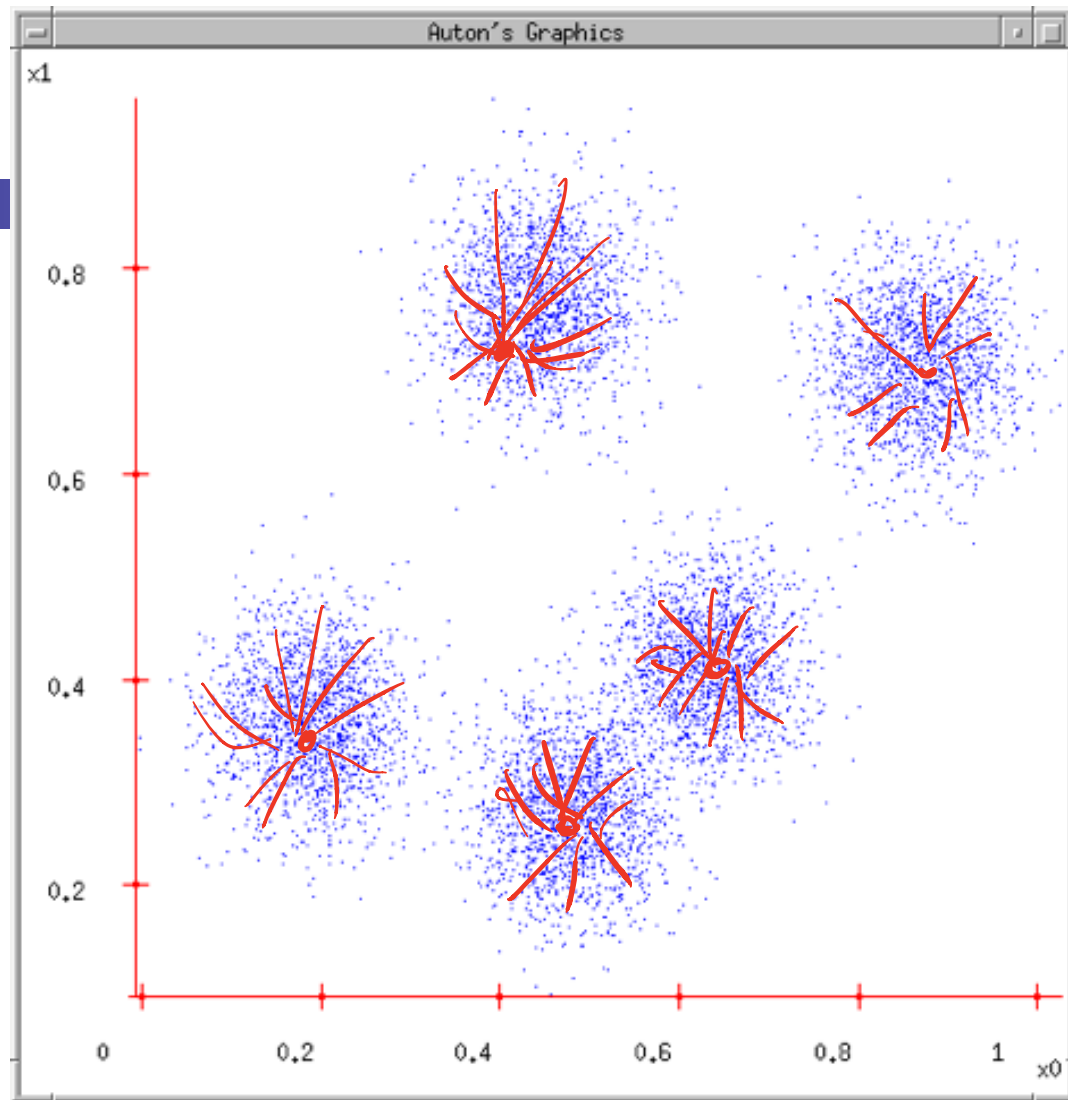


Some Data



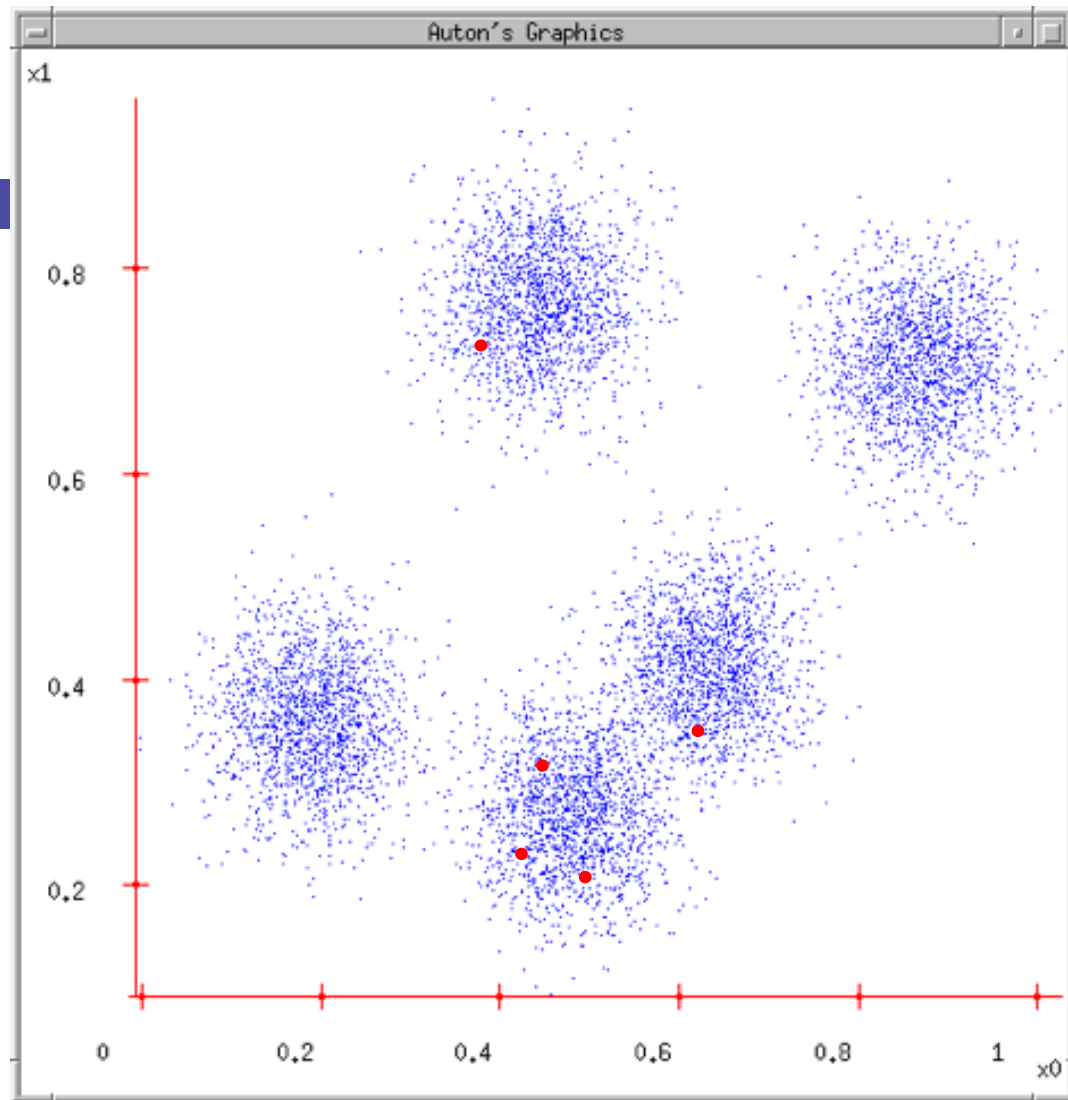
K-means

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



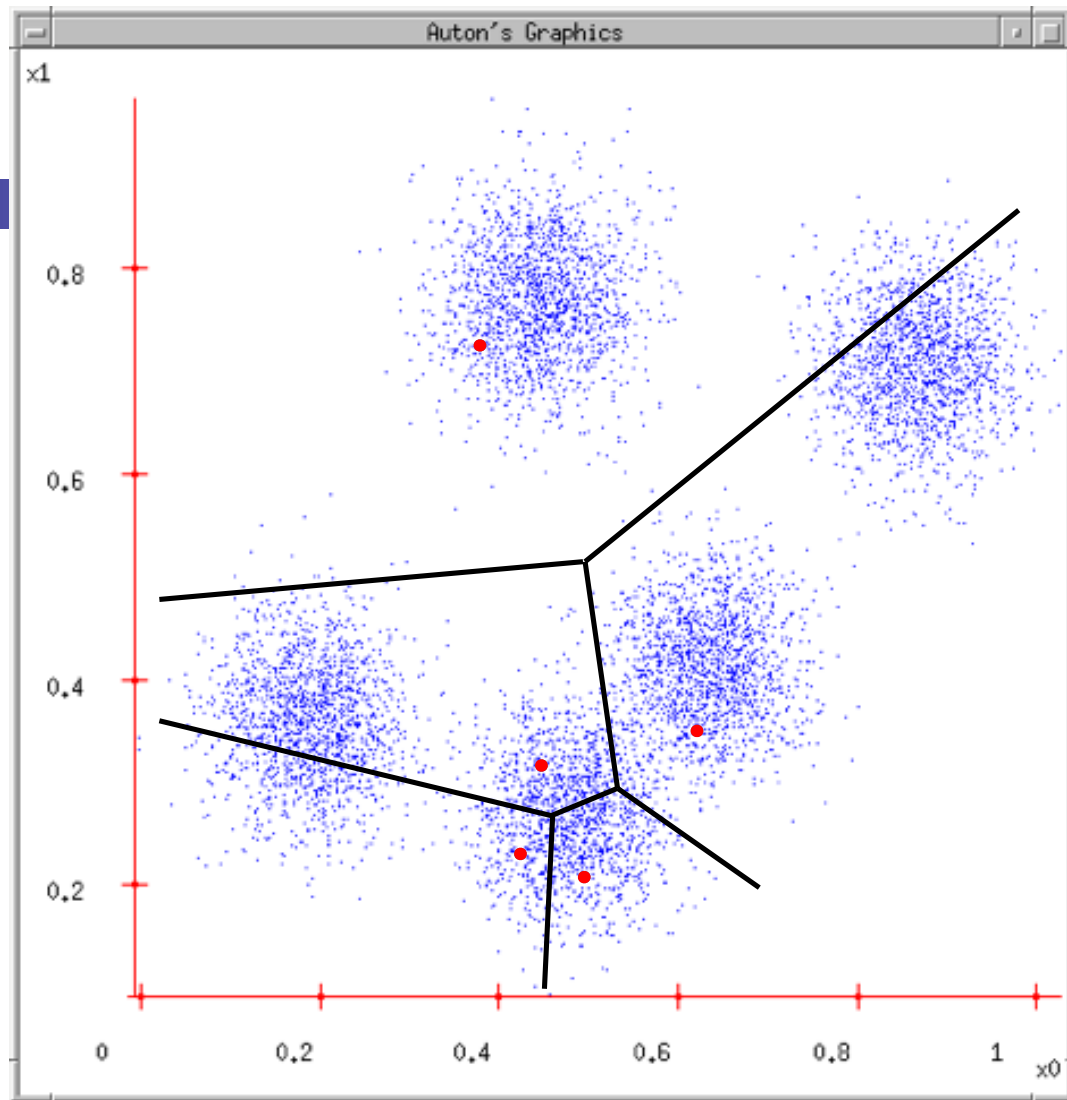
K-means

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



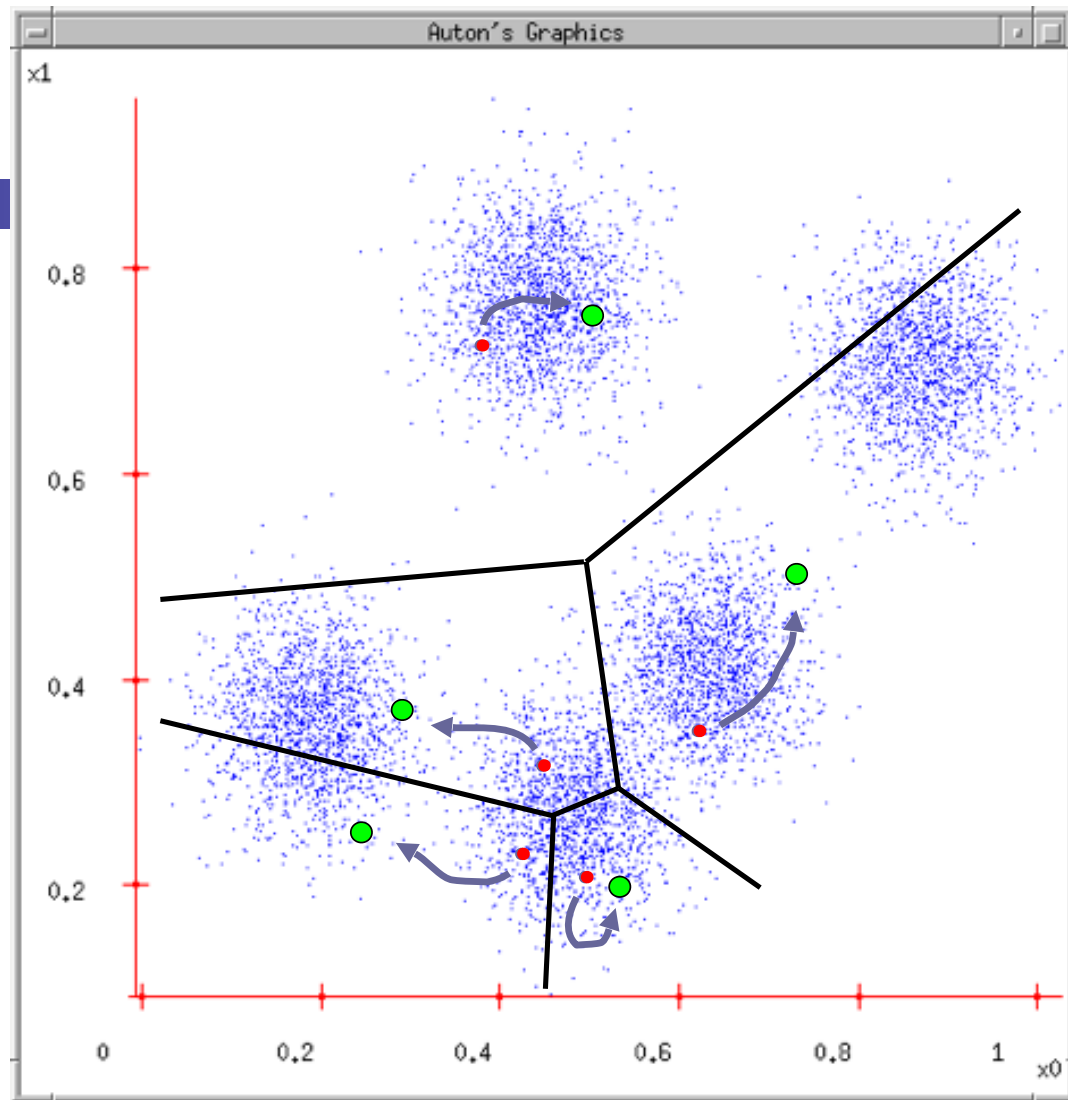
K-means

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)



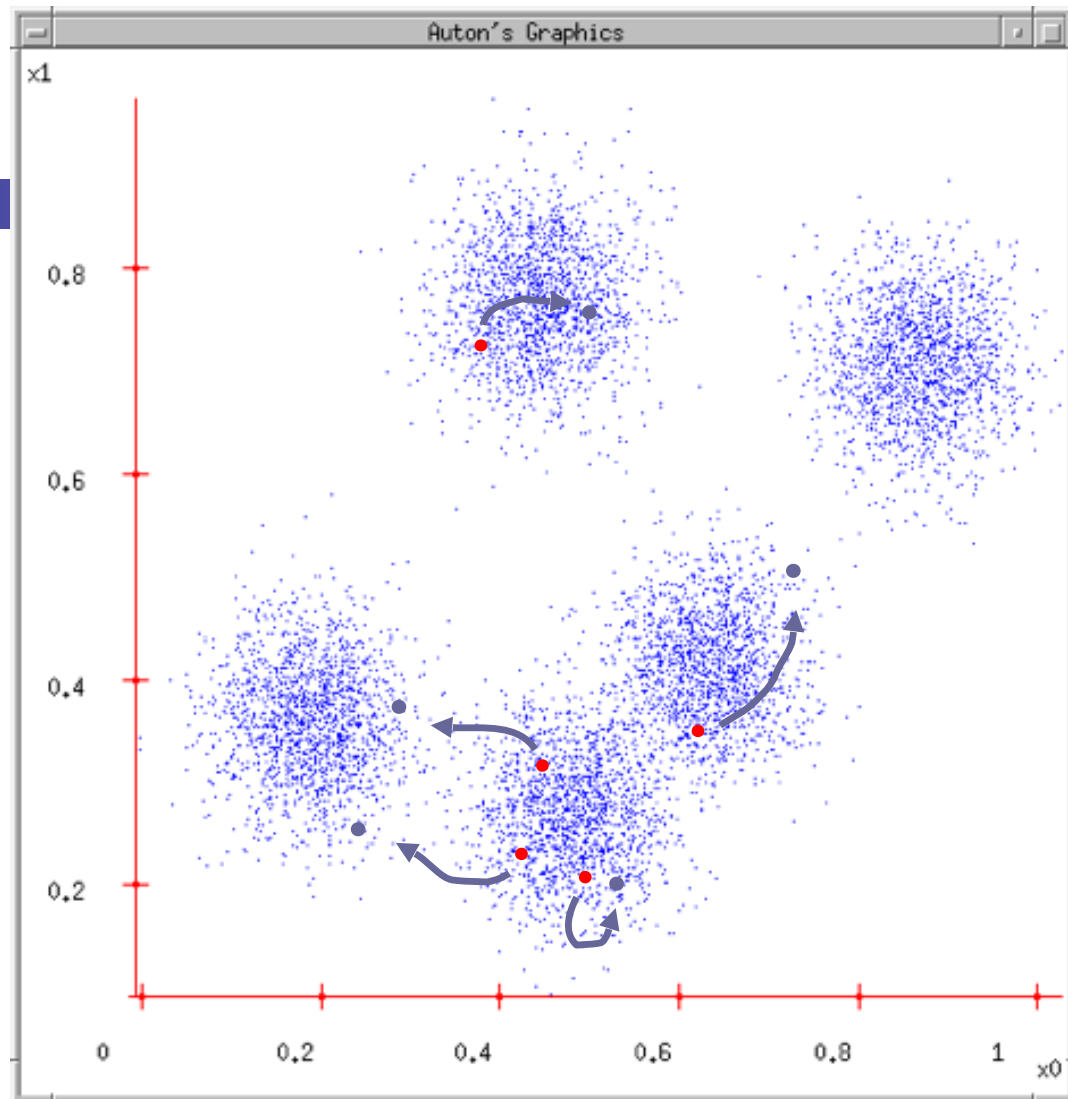
K-means

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

- 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!

What is K-means optimizing?

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

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

- 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

$$\min_{C_i} \|x_j - \mu_i\|_2^2$$

Does K-means converge???

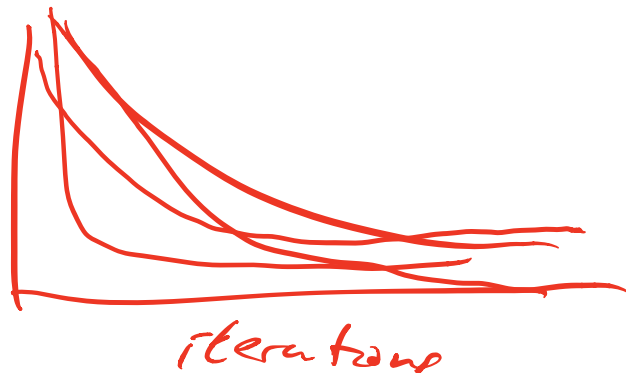
Part 2

- 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

$$\operatorname{argmin}_{\mu} \sum_{i \in C} \|\mu - x_i\|_2^2 \Rightarrow \mu = \frac{1}{n} \sum_{i \in C} x_i$$



Vector Quantization, Fisher Vectors

Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel

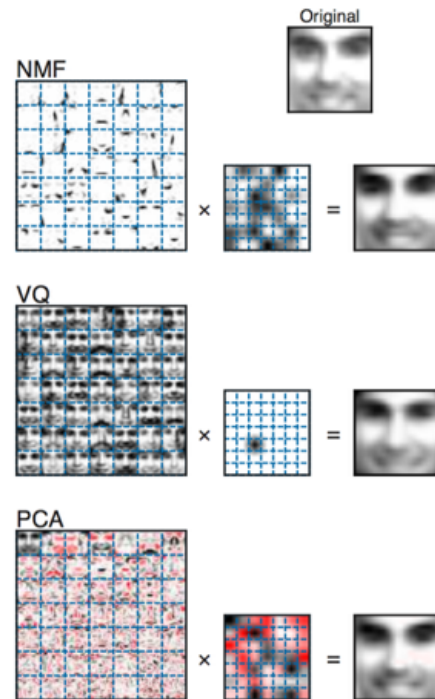
Vector Quantization, Fisher Vectors

Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel



Vector Quantization, Fisher Vectors

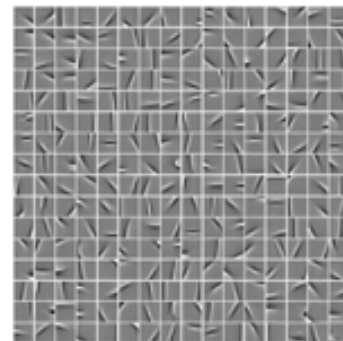
Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel

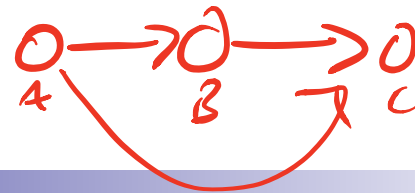
Typical output of k-means on patches



Similar reduced representation can be used as a feature vector

Coates, Ng, *Learning Feature Representations with K-means*, 2012

Spectral Clustering



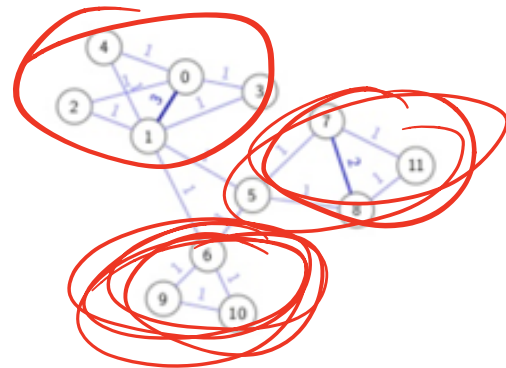
Adjacency matrix: \mathbf{W}

$\mathbf{W}_{i,j}$ = weight of edge (i, j)

$$\mathbf{D}_{i,i} = \sum_{j=1}^n \mathbf{W}_{i,j} \quad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

Given feature vectors, could construct:

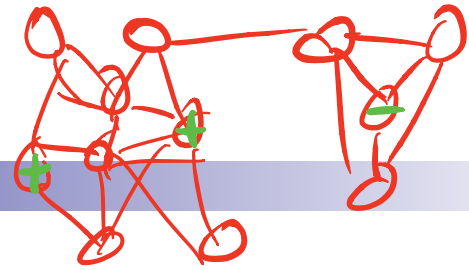
- k-nearest neighbor graph with weights in $\{0, 1\}$
- weighted graph with arbitrary *similarities* $\mathbf{W}_{i,j} = e^{-\gamma \|x_i - x_j\|^2}$



Let $f \in \mathbb{R}^n$ be a function over the nodes

$$\begin{aligned} \mathbf{f}^T \mathbf{L} \mathbf{f} &= \sum_{i=1}^N g_i f_i^2 - \sum_{i=1}^N \sum_{i'=1}^N f_i f_{i'} w_{ii'} \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N w_{ii'} (f_i - f_{i'})^2. \end{aligned}$$

Spectral Clustering



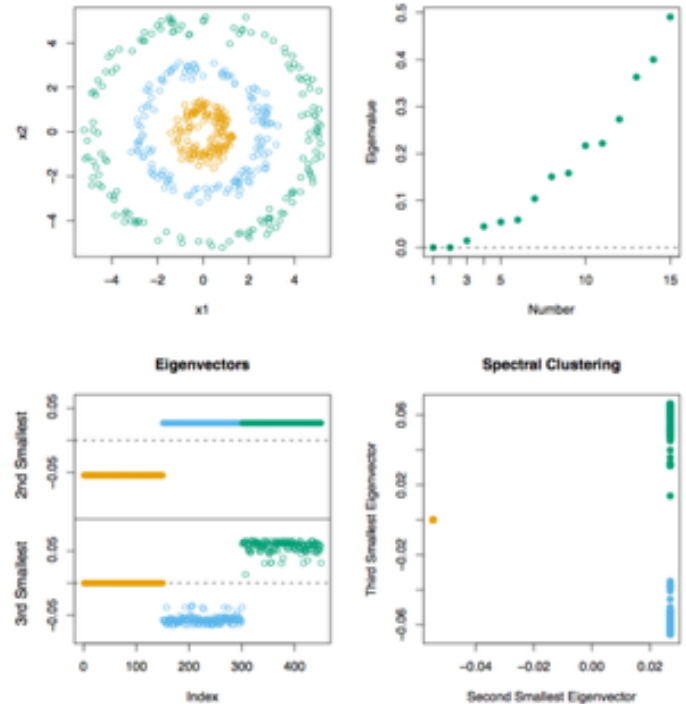
Adjacency matrix: \mathbf{W}

$\mathbf{W}_{i,j}$ = weight of edge (i, j)

$$\mathbf{D}_{i,i} = \sum_{j=1}^n \mathbf{W}_{i,j} \quad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

Given feature vectors, could construct:

- (k=10)-nearest neighbor graph with weights in $\{0,1\}$



Popular to use the Laplacian \mathbf{L} or its normalized form $\tilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$ as a regularizer for learning over graphs

$$w_{cc} = \sum_{i \in \text{observed}} (y_i - f_i)^2 + \lambda f^T \mathbf{L} f$$