

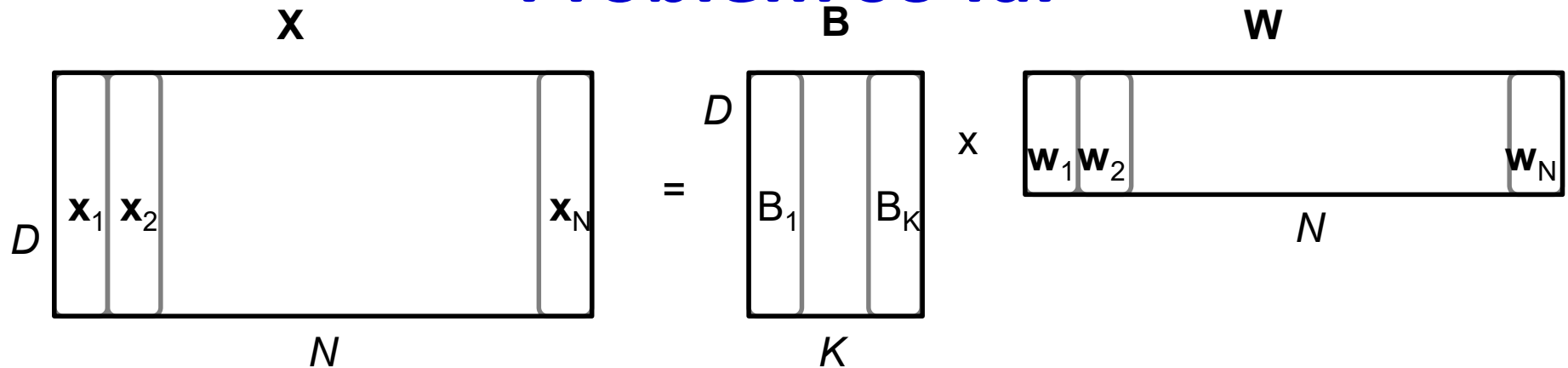
Machine Learning for Signal Processing

Quantization and Clustering

Bhiksha Raj

Learning Representations:

Problem so far



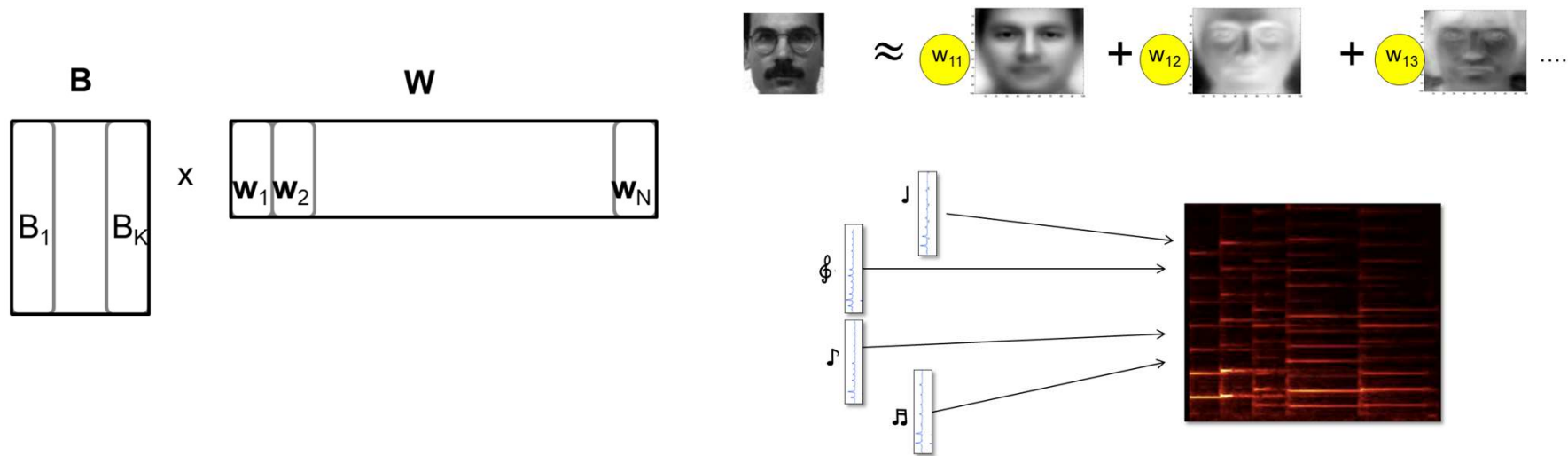
$$x_i = Bw_i$$



$$x_i = w_{11}B_1 + \dots + w_{1K}B_K$$


- **Problem:** Given a collection of data X , find a set of “bases” B , such that each vector x_i can be expressed as a weighted combination of the bases

Why is this important?




- With the right set of bases, the weights represent the data most effectively
 - We can now use the weights to represent the data
 - E.g. with notes as bases, the weights would be the score
- If the bases are agreed upon, we can also *communicate* the information about the data most efficiently
 - Just communicate the weights
 - E.g. enough to store eigen face weights to reconstruct face
 - E.g. just reading the score is sufficient for anyone to recreate music

What is the most accurate way to represent data


$$f = \sum_i w_i d_i$$

D $w_k = 1, w_j = 0 \text{ for } j \neq k$ f




↑
Selecting the kth face in the collection

- If, instead of bases, we had a *dictionary* of all possible data
 - A matrix that included every possible data vector as a column
 - And the weights vector simply selected the correct data instance
 - I.e. \mathbf{w} was *one-sparse* vector

$$|\mathbf{w}|_0 = 1$$

(actually a one-hot vector because the one non-zero entry of $\mathbf{w} = 1$, i.e. $\sum_i w_i = 1$)

What is the most accurate way to represent data




D

$$f = \sum_i w_i d_i$$

$w_k = 1, w_j = 0 \text{ for } j \neq k$


Selecting the k th face in the collection



f

- If, instead of bases, we had a *dictionary* of all possible data
 - A matrix that included every possible data vector as a column
 - And the weights vector simply selected the correct data instance
- **Problem:** Infeasible to construct such a dictionary!
 - Will require infinite entries
 - And our \mathbf{w} vector too will require infinite bits to represent
 - Alternately, will require storing the entire training data
 - And will not be useful to represent data outside the training set


Approximate representation with a dictionary

 ...

$f \approx \sum_i w_i d_i$

$w_k = 1, w_j = 0 \text{ for } j \neq k$

f




↑

Selecting the kth face in the collection


- **Problem:** Infeasible to construct a perfect dictionary
 - Will require too many (potentially infinite) entries
- **Solution:** Can we instead construct a smaller *finite* dictionary such that all data can be approximated well by one of the entries in the dictionary?
 - E.g. “The guy looks a lot like the 7th face in the dictionary”
 - E.g. The vector x looks a lot like the d_i , the i-th entry in the dictionary.
- **Questions:**
 - What do we mean by “looks a lot like”
 - How do we construct the dictionary?

Approximate representation with a dictionary

 ...

$$f \approx \sum_i w_i d_i$$


$w_k = 1, w_j = 0 \text{ for } j \neq k$



Selecting the kth face in the collection

- **Problem:** Infeasible to construct a perfect dictionary
 - Will require too many (potentially infinite) entries
- **Solution:** Can we instead construct a smaller *finite* dictionary such that all data can be approximated well by one of the entries in the dictionary?
 - E.g. “The guy looks a lot like the 7th face in the dictionary”
 - E.g. The vector x looks a lot like the d_i , the i-th entry in the dictionary.
- **Questions:**
 - What do we mean by “looks a lot like”
 - How do we construct the dictionary?

Quantifying the error




$$f \approx \sum_i w_i d_i$$

D

$$w_k = 1, w_j = 0 \text{ for } j \neq k$$

Selecting the kth face in the collection



f

- Different error metrics will result in different solutions
- Let's generically represent the error as $div()$


$$\hat{f} = D\mathbf{w}, \quad \|\mathbf{w}\|_0 = 1, \quad \sum_i w_i = 1$$

$$Error(f) = div(f, \hat{f})$$

- A common choice is the L_2 error


$$Error(f) = \|f - \hat{f}\|^2$$

Approximate representation with a dictionary

 ...

$$f \approx \sum_i w_i d_i$$

$w_k = 1, w_j = 0 \text{ for } j \neq k$



Selecting the kth face in the collection

- **Problem:** Infeasible to construct a perfect dictionary
 - Will require too many (potentially infinite) entries
- **Solution:** Can we instead construct a smaller *finite* dictionary such that all data can be approximated well by one of the entries in the dictionary?
 - E.g. “The guy looks a lot like the 7th face in the dictionary”
 - E.g. The vector x looks a lot like the d_i , the i-th entry in the dictionary.
- **Questions:**
 - What do we mean by “looks a lot like”
 - How do we construct the dictionary?

Learning the Dictionary

- $V = [V_1, V_2, V_3, \dots]$ are the data for which the dictionary is being learned
- $D = [d_1, d_2, \dots, d_K]$ is the matrix of dictionary vectors
- $W = [w_1, w_2, w_3, \dots]$ is a set of *one-hot* vectors
- Learning: Learn D and W to minimize total error on V

$$\hat{D}, \hat{W} = \underset{D, W}{\operatorname{argmin}} \operatorname{div}(V, DW) = \underset{D, W}{\operatorname{argmin}} \sum_i \operatorname{div}(V_i, Dw_i),$$

s.t. $w_i = \text{one hot}$

- If we're only interested in learning the dictionary

$$\hat{D} = \underset{D}{\operatorname{argmin}} \min_W \sum_i \operatorname{div}(V_i, Dw_i), \quad \text{s.t. } w_i = \text{one hot}$$

Learning the Dictionary

- $\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \min_{\mathbf{W}} \sum_i \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$

$$= \operatorname{argmin}_{\mathbf{D}} \sum_i \min_{\mathbf{w}_i} \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$$

- Generally does not have a closed form solution, but can be solved with the following iteration that provably reduces error in each step

$$\mathbf{w}_i = \operatorname{argmin}_{\mathbf{w}} \operatorname{div}(V_i, \mathbf{D}\mathbf{w})$$

$$\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$$

Learning the Dictionary

- $\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(\mathbf{V}_i, \mathbf{D}\mathbf{w}_i)$

For $\operatorname{div}(\cdot) = \|\mathbf{V}_i - \mathbf{D}\mathbf{w}_i\|^2$ this gives us the well-known K-means algorithm

- Generally does not have a closed form solution, but can be solved with the following iteration that provably reduces error in each step

$$\mathbf{w}_i = \operatorname{argmin}_{\mathbf{w}} \operatorname{div}(\mathbf{V}_i, \mathbf{D}\mathbf{w})$$

$$\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(\mathbf{V}_i, \mathbf{D}\mathbf{w}_i)$$

Learning the Dictionary

- $\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$

For $\operatorname{div}(\cdot) = \|V_i - \mathbf{D}\mathbf{w}_i\|^2$ this gives us the well-known K-means algorithm

$$\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$$

- Grouping V_i by the dictionary entries they are assigned to (\mathbf{w}_i) results in *clustering*
error in each step

$$\mathbf{w}_i = \operatorname{argmin}_{\mathbf{w}} \operatorname{div}(V_i, \mathbf{D}\mathbf{w})$$

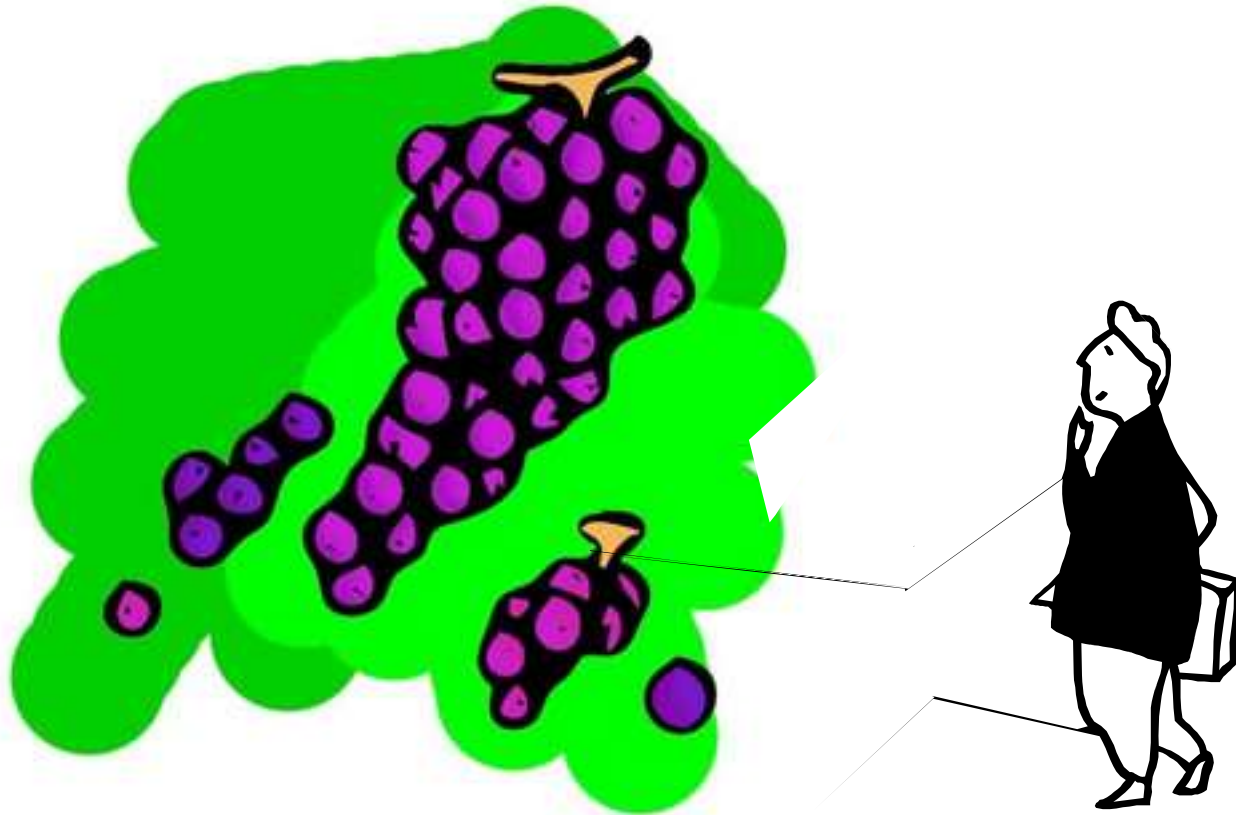
$$\hat{\mathbf{D}} = \operatorname{argmin}_{\mathbf{D}} \sum_i \operatorname{div}(V_i, \mathbf{D}\mathbf{w}_i)$$

Poll 1

So let's look at clustering

- From a more naïve, procedural perspective..

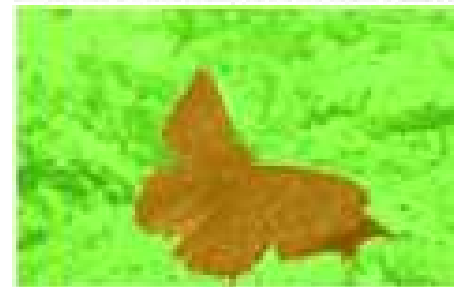
Clustering



Statistical Modelling and Latent Structure

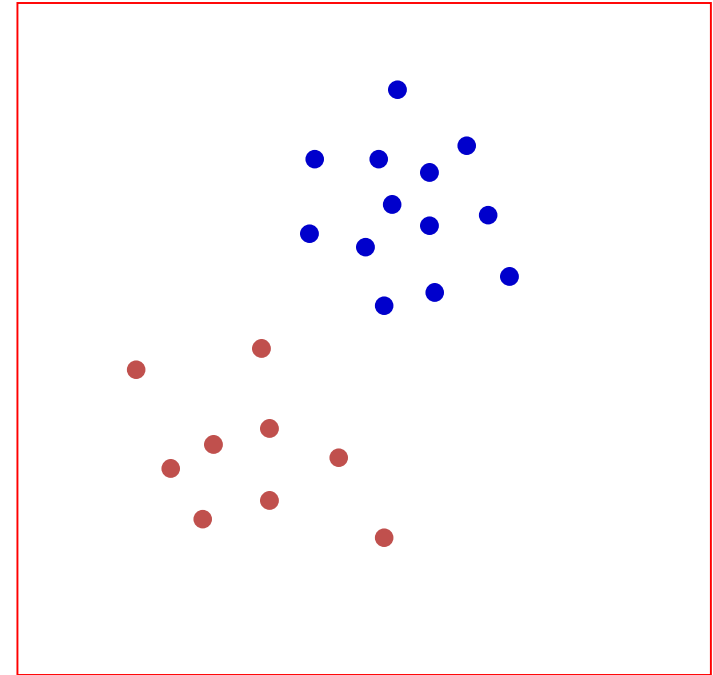
- Much of statistical modelling attempts to identify *latent* structure in the data
 - Structure that is not immediately apparent from the observed data
 - But which, if known, helps us explain it better, and make predictions from or about it
- Clustering methods attempt to extract such structure from *proximity*
 - *First-level* structure (as opposed to deep structure)
- We will see still other forms of latent structure discovery later in the course

How



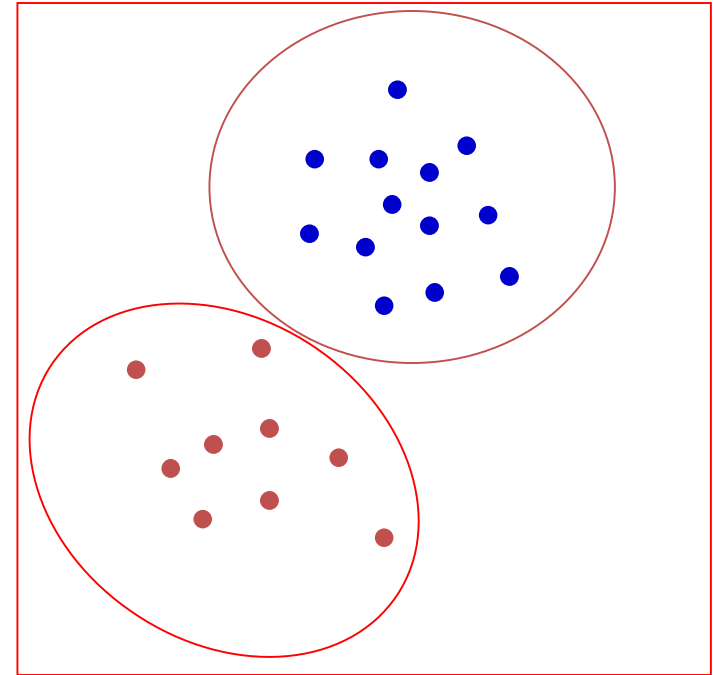
Clustering

- What is clustering
 - Clustering is the determination of naturally occurring grouping of data/instances (with low within-group variability and high between-group variability)



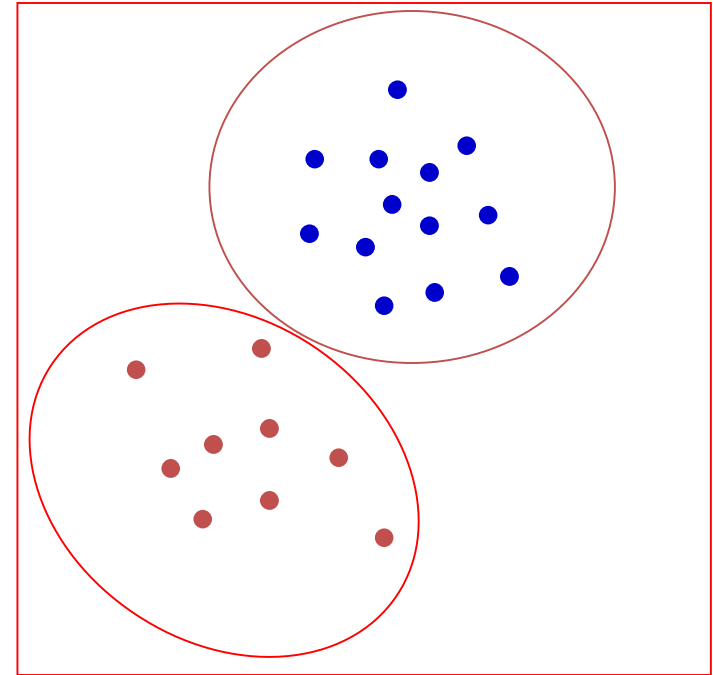
Clustering

- What is clustering
 - Clustering is the determination of naturally occurring grouping of data/instances (with low within-group variability and high between-group variability)



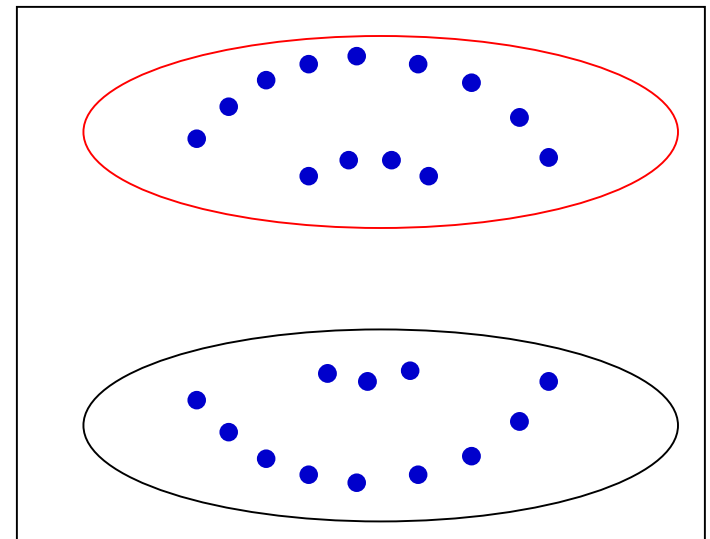
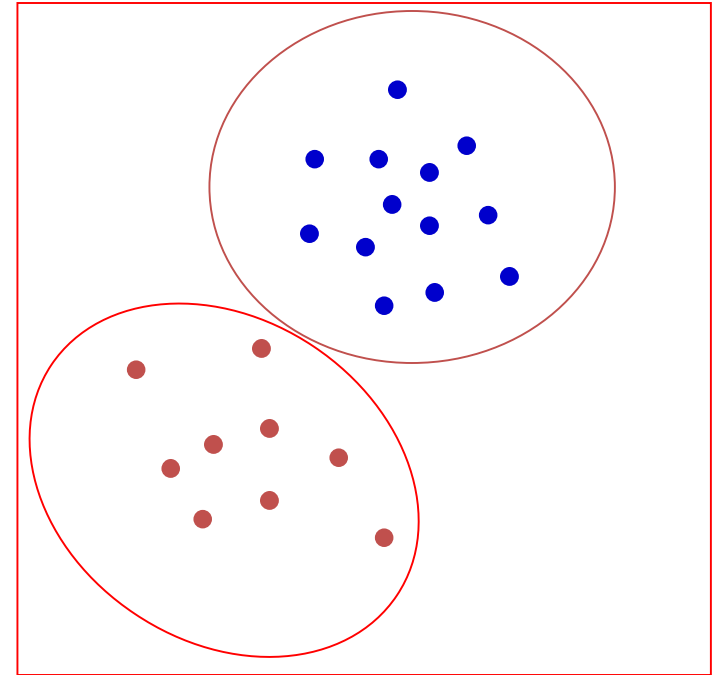
Clustering

- What is clustering
 - Clustering is the determination of naturally occurring grouping of data/instances (**with low within-group variability and high between-group variability**)
- How is it done
 - Find groupings of data such that the groups optimize a “within-group-variability” objective function of some kind



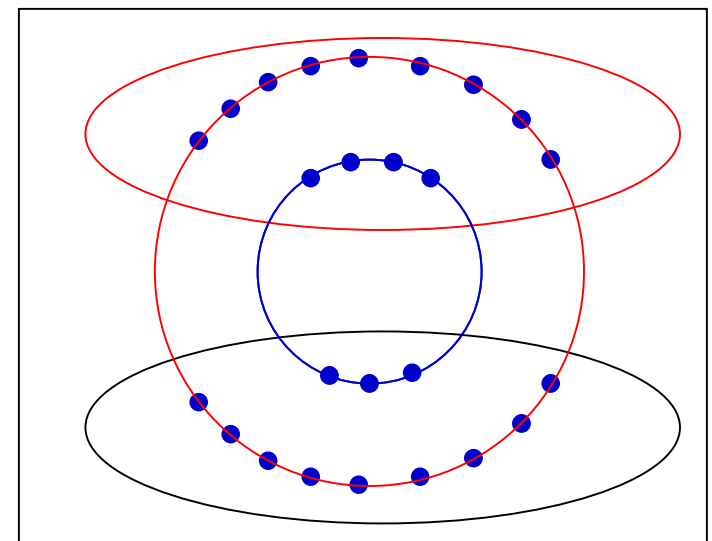
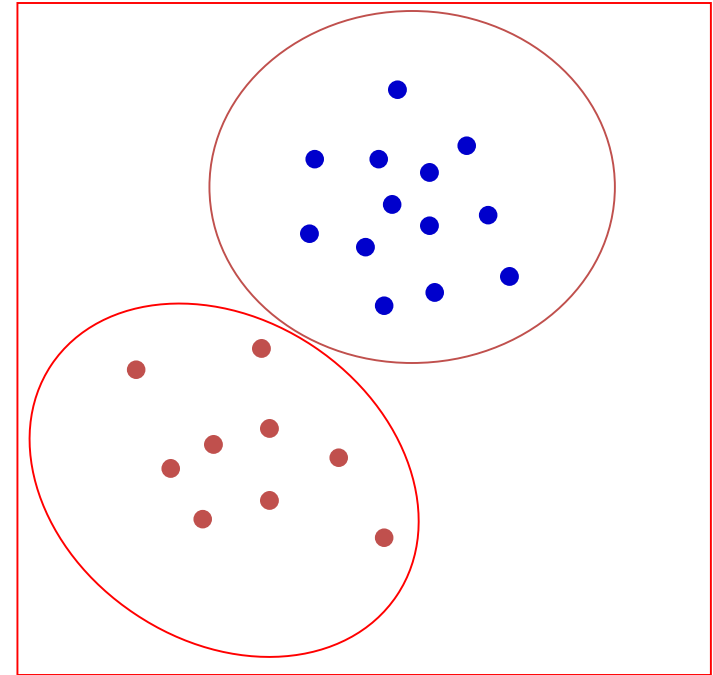
Clustering

- What is clustering
 - Clustering is the determination of naturally occurring grouping of data/instances (**with low within-group variability and high between-group variability**)
- How is it done
 - Find groupings of data such that the groups optimize a “within-group-variability” objective function of some kind
 - The objective function used affects the nature of the discovered clusters
 - E.g. Euclidean distance vs.



Clustering

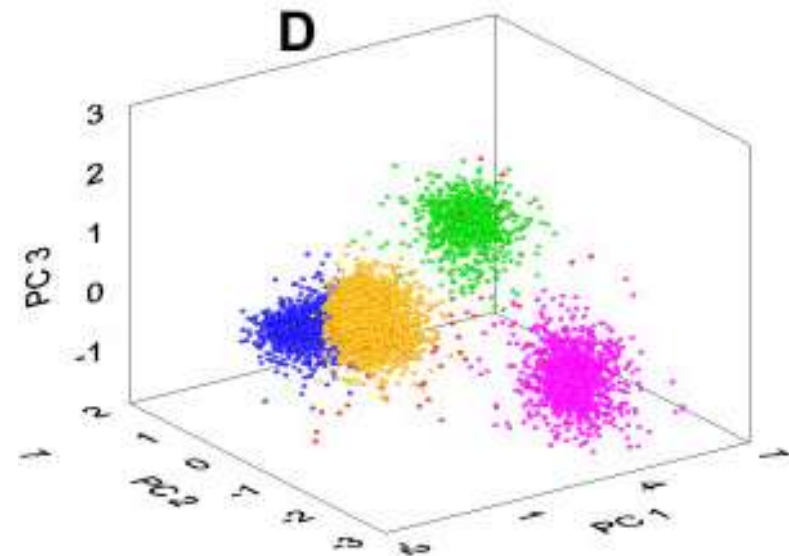
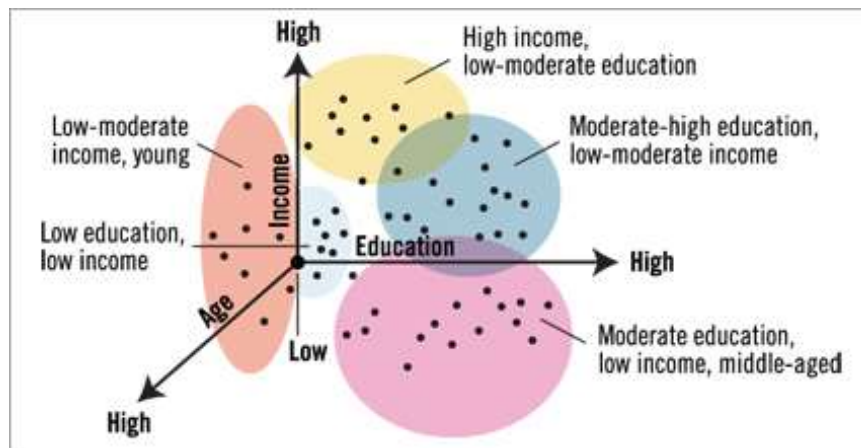
- What is clustering
 - Clustering is the determination of naturally occurring grouping of data/instances (**with low within-group variability and high between-group variability**)
- How is it done
 - Find groupings of data such that the groups optimize a “within-group-variability” objective function of some kind
 - The objective function used affects the nature of the discovered clusters
 - E.g. Euclidean distance vs.
 - Distance from center



Why Clustering

- **Find structure:** Automatic grouping into “Classes”
 - Different clusters may show different behavior
- **Representation:** Quantization
 - All data within a cluster are represented by a single point
- Preprocessing step for other algorithms
 - Indexing, categorization, etc.

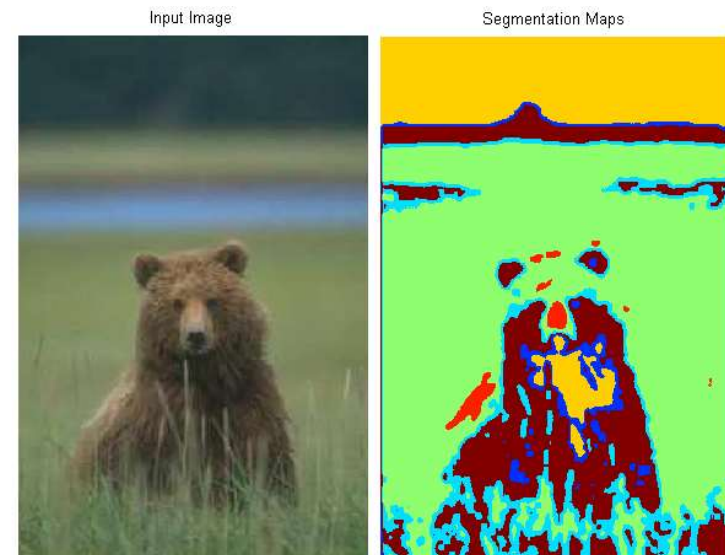
Finding natural structure in data



- Find natural groupings in data for further analysis
- Discover *latent* structure in data

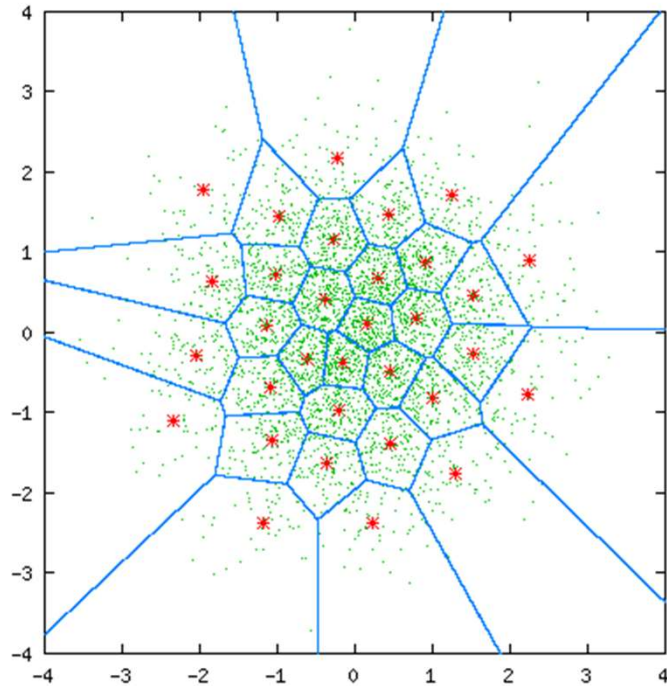
Some Applications of Clustering

- Image segmentation

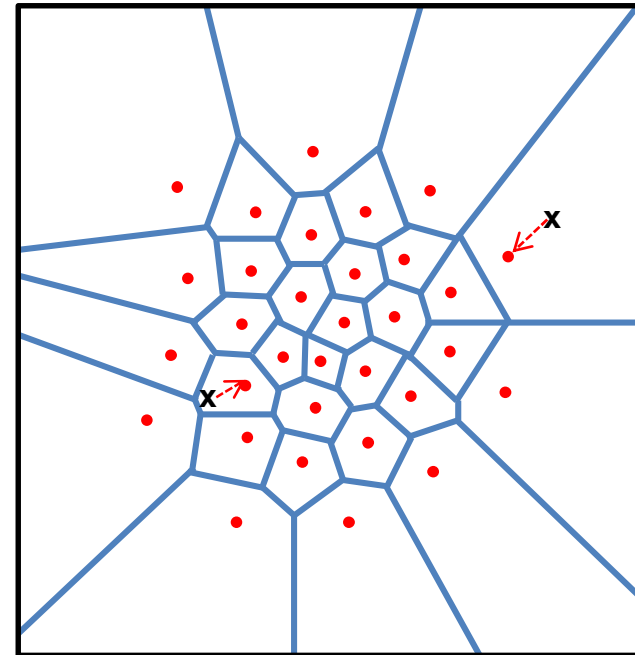


Representation: Quantization

TRAINING



QUANTIZATION

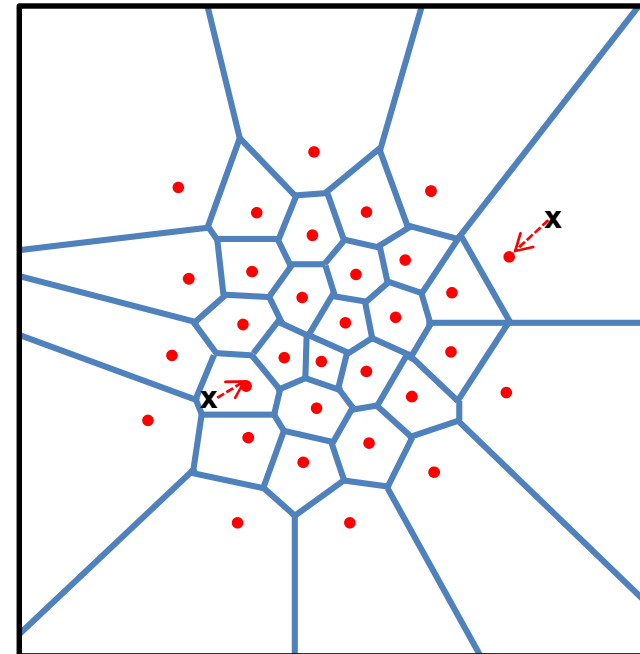


- *Quantize every vector* to one of K (vector) values
- What are the optimal K vectors? How do we find them? How do we perform the quantization?
- **LBG algorithm**

Quantization: Formally

$$V = \sum_i w_i d_i$$

$$V = \mathbf{D}\mathbf{w} \quad \begin{array}{l} |\mathbf{w}| = 1 \\ |\mathbf{w}|_0 = 1 \end{array}$$



- d_i are the “representative” vectors of each cluster
- Restriction: only one of the w_i is 1, the rest are 0
 - $\sum_i w_i = 1$
 - \mathbf{w} is unit length and one-sparse

Representation: BOW

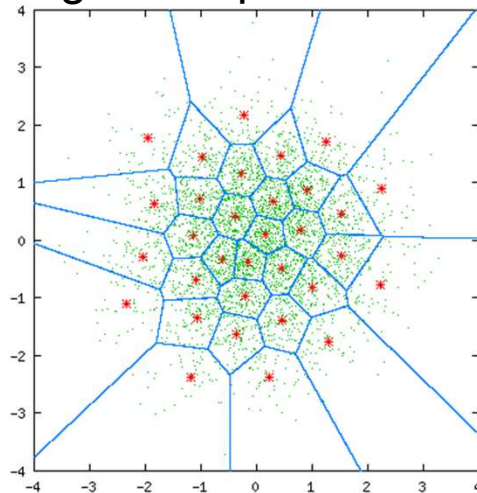


- How to retrieve all music videos by this guy?
- Build a classifier
 - But how do you *represent* the video?

Representation: BOW

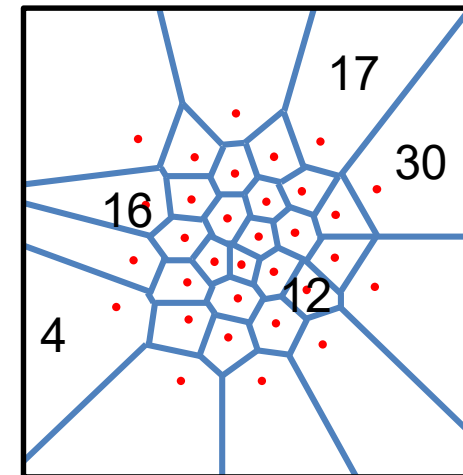


Training: Each point is a video frame



$$V_k = \mathbf{D}\mathbf{w}_k \quad f = \sum_k \mathbf{w}_k$$

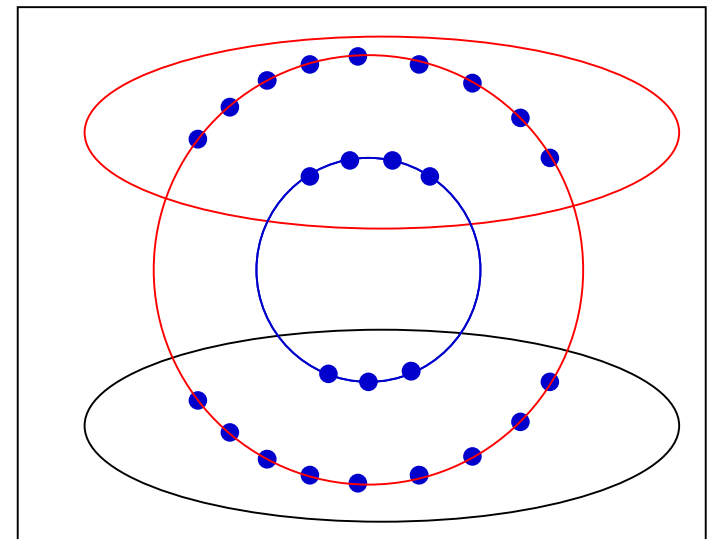
Representation: Each number is the #frames assigned to the codeword



- Bag of words representations of video/audio/data

Obtaining “Meaningful” Clusters

- Two key aspects:
 - 1. The feature representation used to characterize your data
 - 2. The “clustering criteria” employed

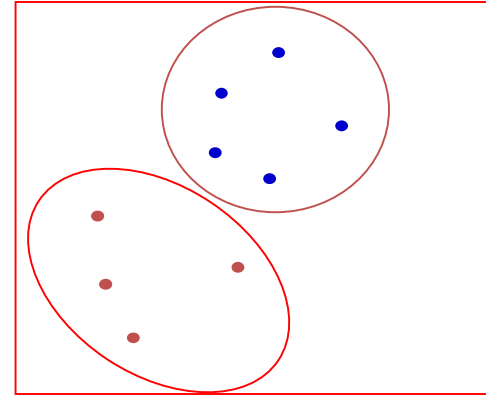


Clustering Criterion

- The “Clustering criterion” actually has two aspects
- Cluster compactness criterion
 - Measure that shows how “good” clusters are
 - The objective function
- Distance of a point from a cluster
 - To determine the cluster a data vector belongs to

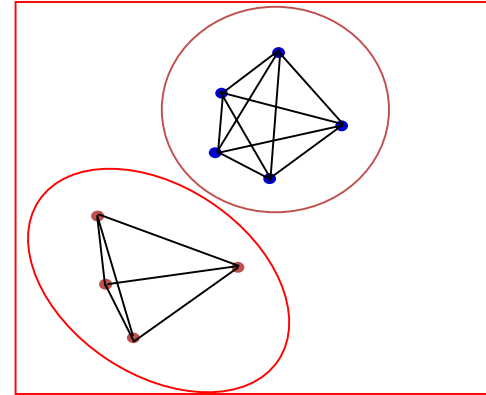
“Compactness” criteria for clustering

- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster



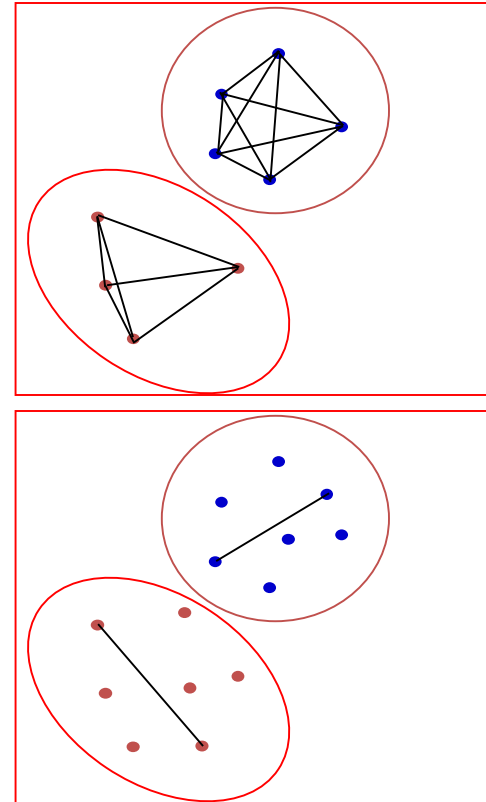
“Compactness” criteria for clustering

- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster



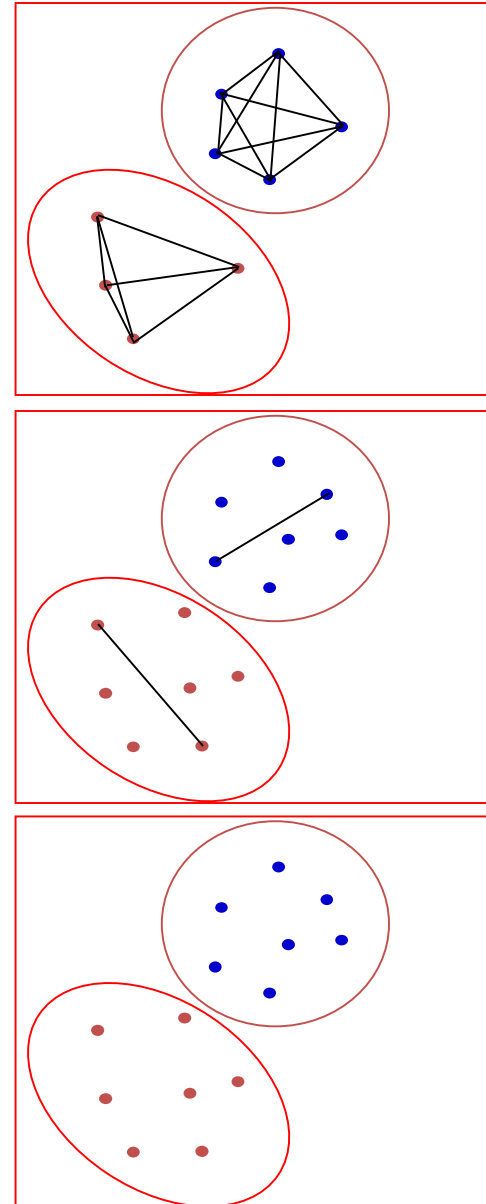
“Compactness” criteria for clustering

- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster
 - Distance between the two farthest points in the cluster



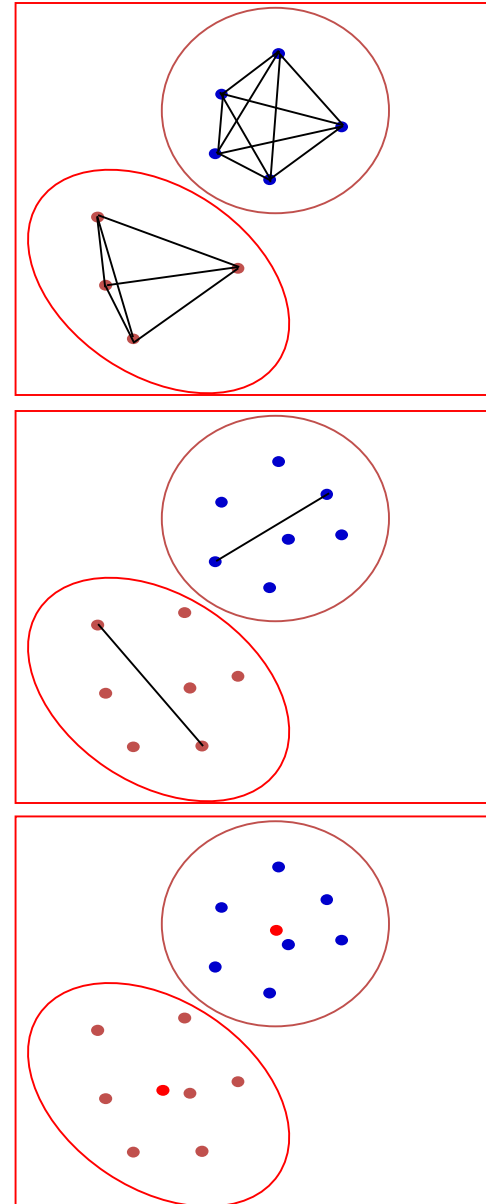
“Compactness” criteria for clustering

- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster
 - Distance between the two farthest points in the cluster
 - Total distance of every element in the cluster from the centroid of the cluster



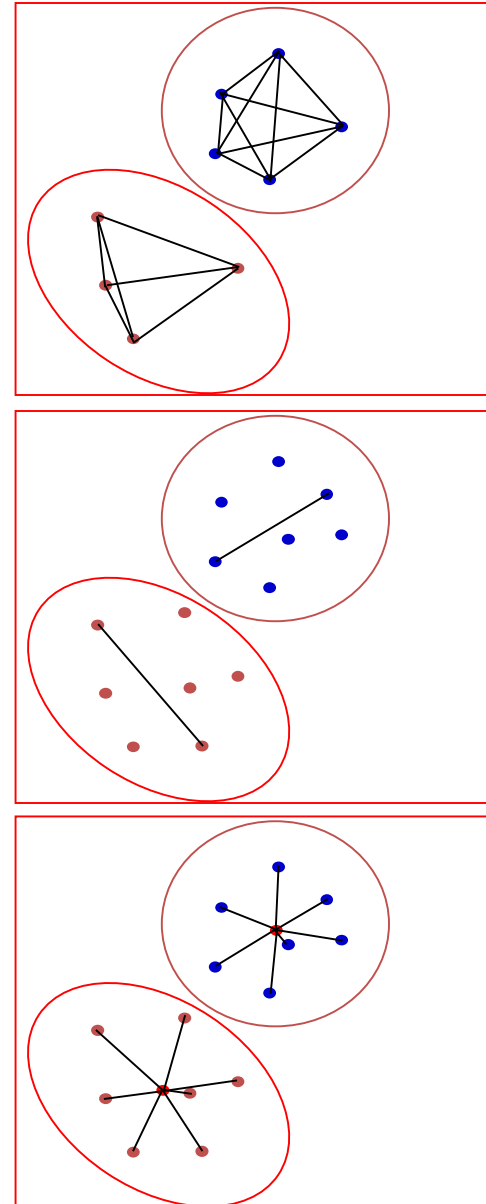
“Compactness” criteria for clustering

- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster
 - Distance between the two farthest points in the cluster
 - Total distance of every element in the cluster from the centroid of the cluster



“Compactness” criteria for clustering

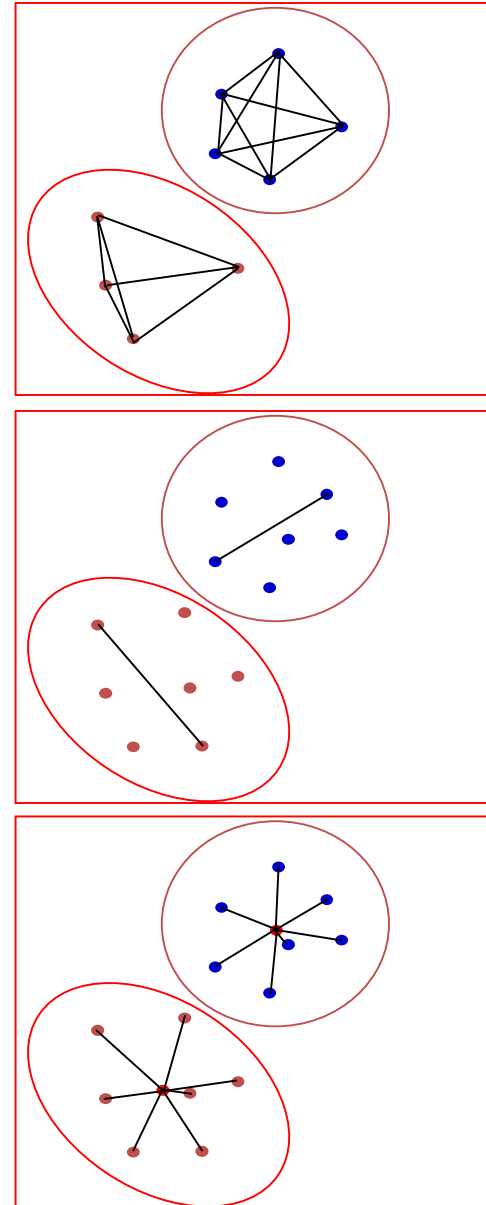
- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster
 - Distance between the two farthest points in the cluster
 - Total distance of every element in the cluster from the centroid of the cluster



“Compactness” criteria for clustering

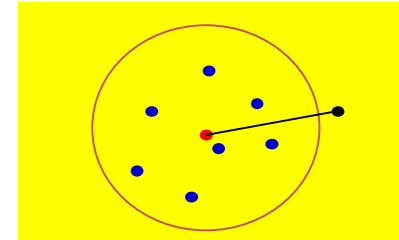
- Distance based measures
 - Total distance between each element in the cluster and every other element in the cluster
 - Distance between the two farthest points in the cluster
 - Total distance of every element in the cluster from the centroid of the cluster
 - Distance measures are often weighted Minkowski metrics

$$dist = \sqrt[n]{w_1|a_1 - b_1|^n + w_2|a_2 - b_2|^n + \dots + w_M|a_M - b_M|^n}$$



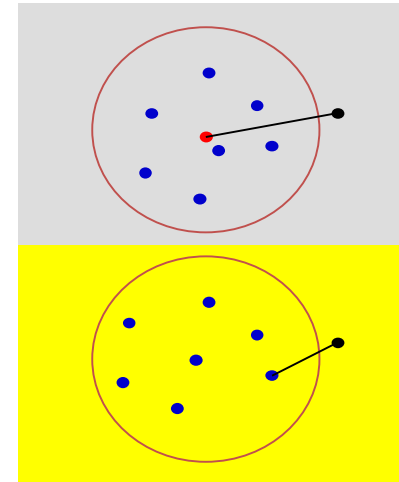
Clustering: Distance from cluster

- How far is a data point from a cluster?
 - Euclidean or Minkowski distance from the centroid of the cluster



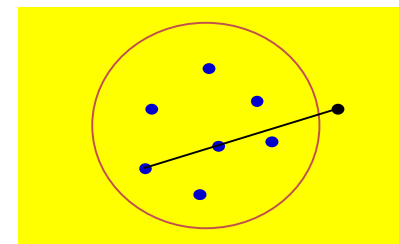
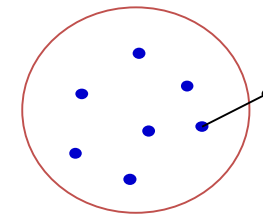
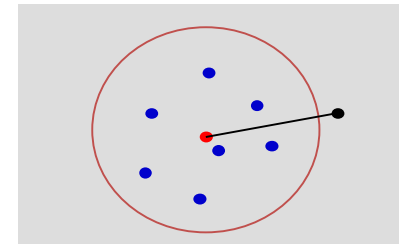
Clustering: Distance from cluster

- How far is a data point from a cluster?
 - Euclidean or Minkowski distance from the centroid of the cluster
 - Distance from the closest point in the cluster



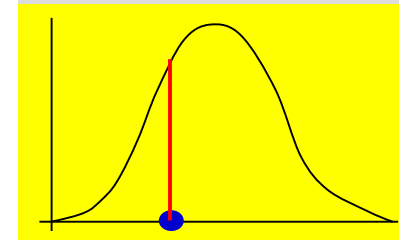
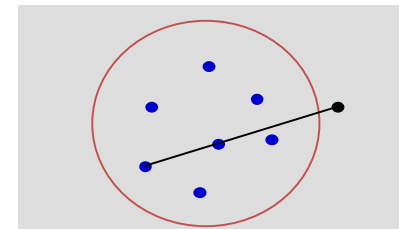
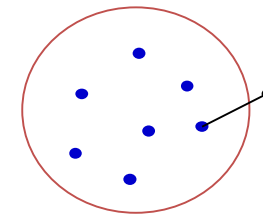
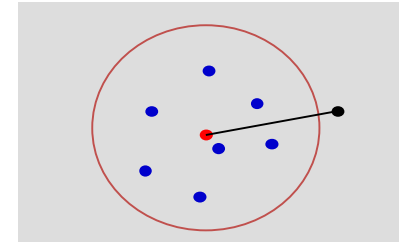
Clustering: Distance from cluster

- How far is a data point from a cluster?
 - Euclidean or Minkowski distance from the centroid of the cluster
 - Distance from the closest point in the cluster
 - Distance from the farthest point in the cluster



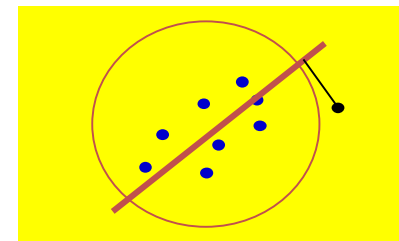
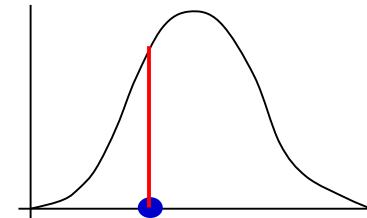
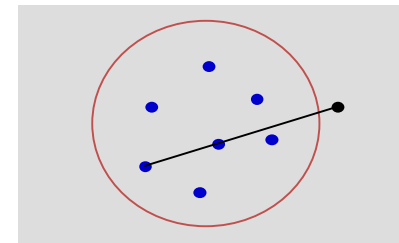
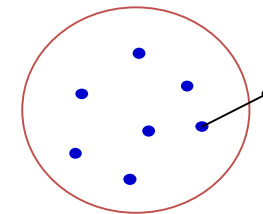
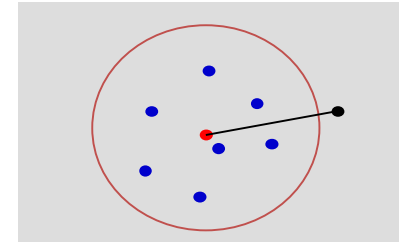
Clustering: Distance from cluster

- How far is a data point from a cluster?
 - Euclidean or Minkowski distance from the centroid of the cluster
 - Distance from the closest point in the cluster
 - Distance from the farthest point in the cluster
 - Probability of data measured on cluster distribution



Clustering: Distance from cluster

- How far is a data point from a cluster?
 - Euclidean or Minkowski distance from the centroid of the cluster
 - Distance from the closest point in the cluster
 - Distance from the farthest point in the cluster
 - Probability of data measured on cluster distribution
 - Fit of data to cluster-based regression



Optimal clustering: Exhaustive enumeration

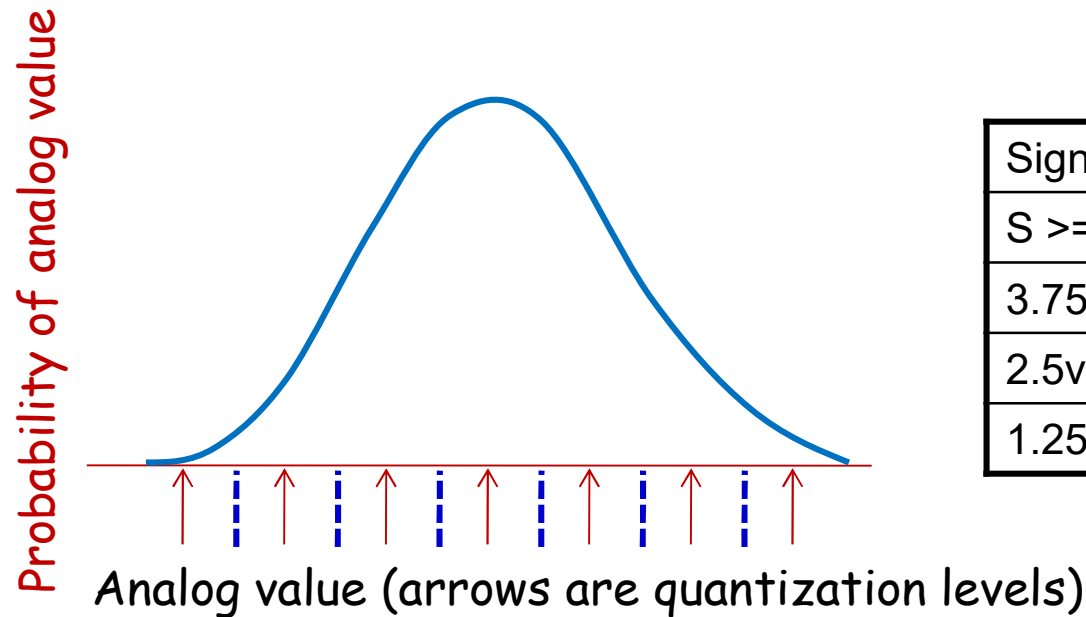
- Find the clusters such that cluster compactness measure is optimized
- All possible combinations of data must be evaluated
 - If there are M data points, and we desire N clusters, the number of ways of separating M instances into N clusters is

$$\frac{1}{M!} \sum_{i=0}^N (-1)^i \binom{N}{i} (N-i)^M$$

- **Exhaustive enumeration** based clustering requires that the objective function (the “Goodness measure”) be evaluated for every one of these, and the best one chosen
- This is the only correct way of optimal clustering
 - Unfortunately, it is also computationally unrealistic

Poll 2

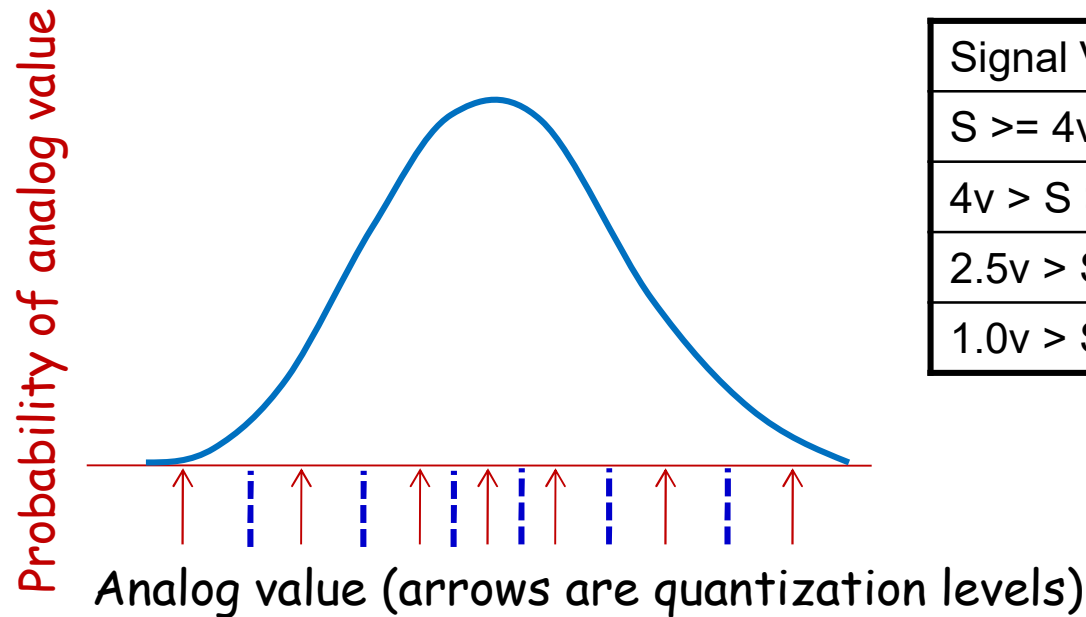
Not-quite non sequitur: Quantization



Signal Value	Bits	Mapped to
$S \geq 3.75v$	11	$3 * \text{const}$
$3.75v > S \geq 2.5v$	10	$2 * \text{const}$
$2.5v > S \geq 1.25v$	01	$1 * \text{const}$
$1.25v > S \geq 0v$	00	0

- Linear quantization (uniform quantization):
 - Each digital value represents an equally wide range of analog values
 - Regardless of distribution of data
 - Digital-to-analog conversion represented by a “uniform” table

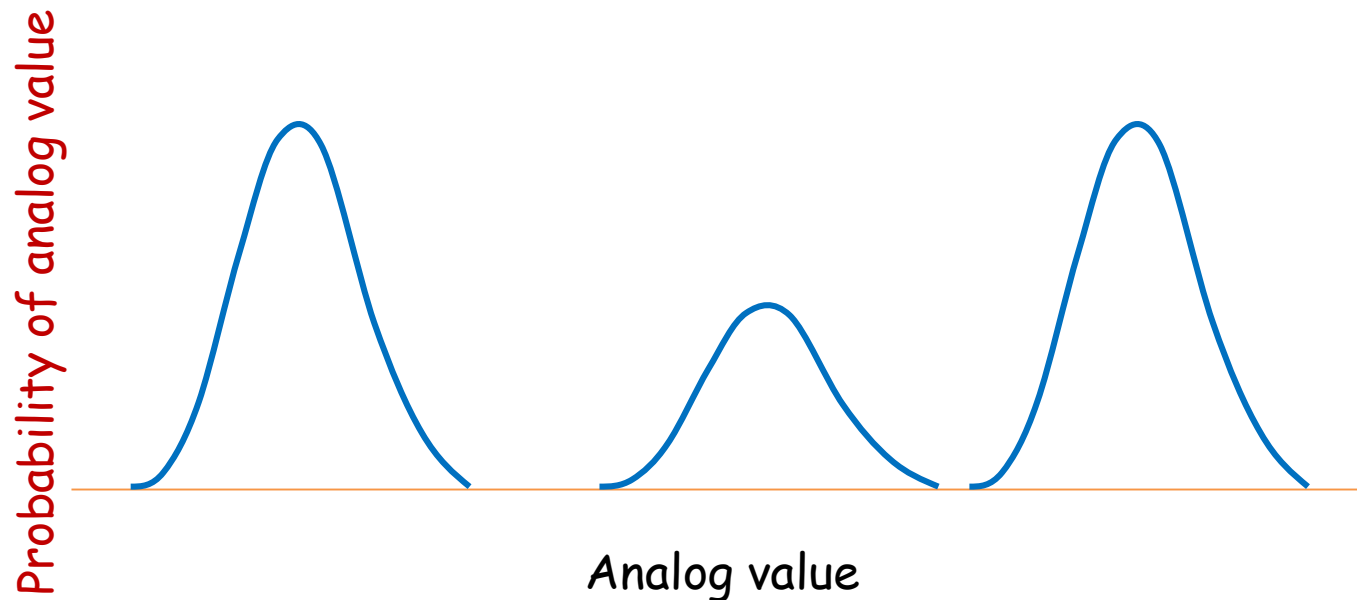
Not-quite non sequitur: Quantization



Signal Value	Bits	Mapped to
$S \geq 4v$	11	4.5
$4v > S \geq 2.5v$	10	3.25
$2.5v > S \geq 1v$	01	1.25
$1.0v > S \geq 0v$	00	0.5

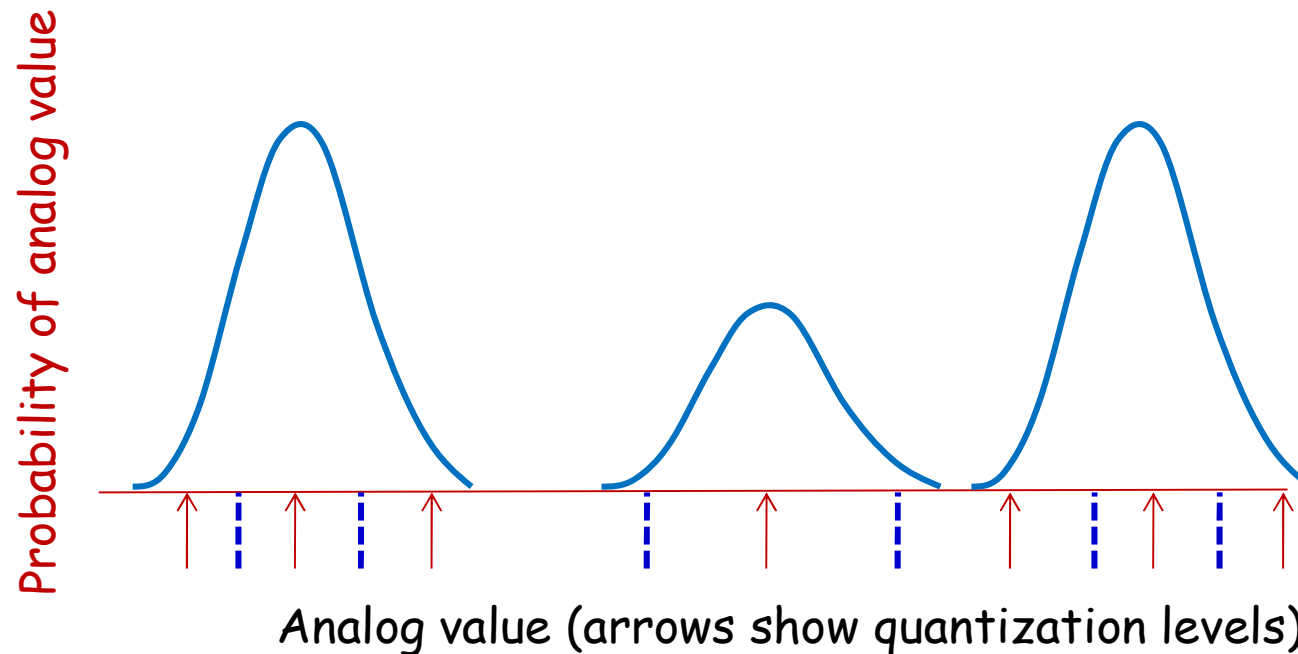
- Non-Linear quantization:
 - Each digital value represents a different range of analog values
 - Finer resolution in high-density areas
 - Mu-law / A-law assumes a Gaussian-like distribution of data
 - Digital-to-analog conversion represented by a “non-uniform” table

Non-uniform quantization



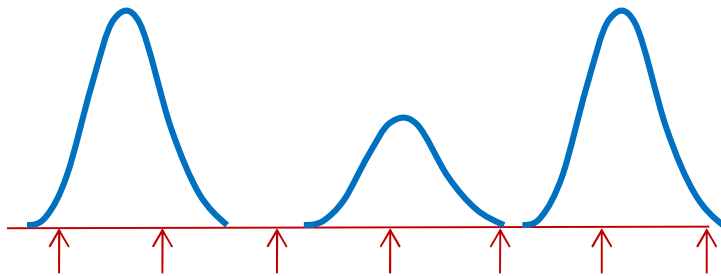
- If data distribution is not Gaussian-ish?
 - Mu-law / A-law are not optimal
 - How to compute the optimal ranges for quantization?
 - Or the optimal table

The Lloyd Quantizer



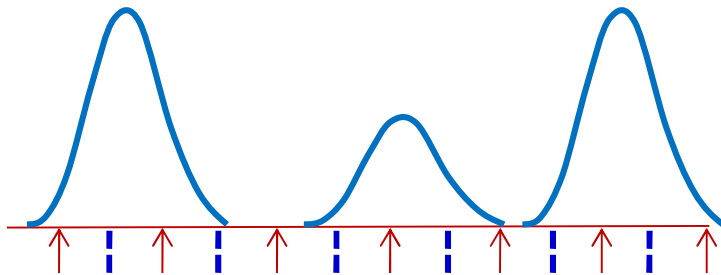
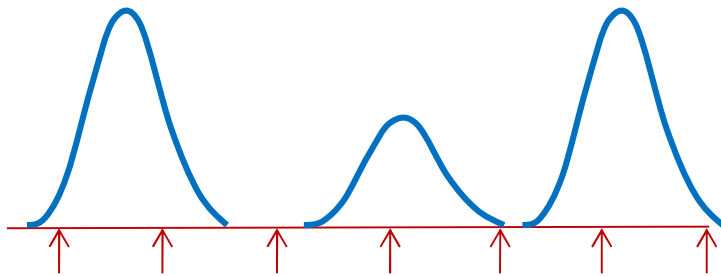
- Lloyd quantizer: An iterative algorithm for computing optimal quantization tables for non-uniformly distributed data
- **Learned from “training” data**

Lloyd Quantizer



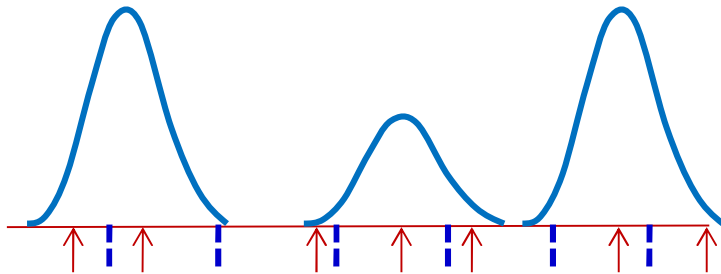
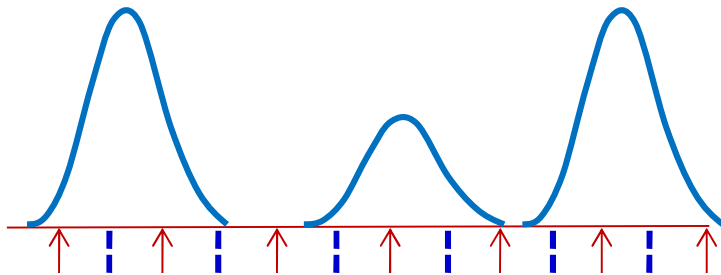
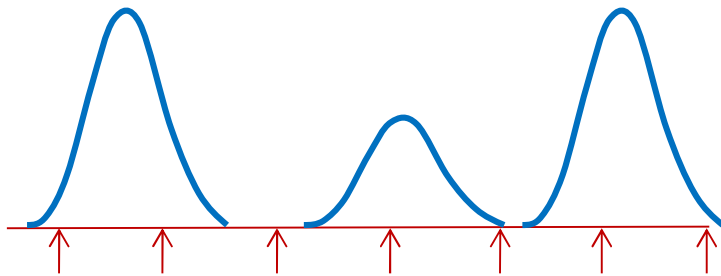
- Randomly initialize quantization points
 - Right column entries of quantization table

Lloyd Quantizer



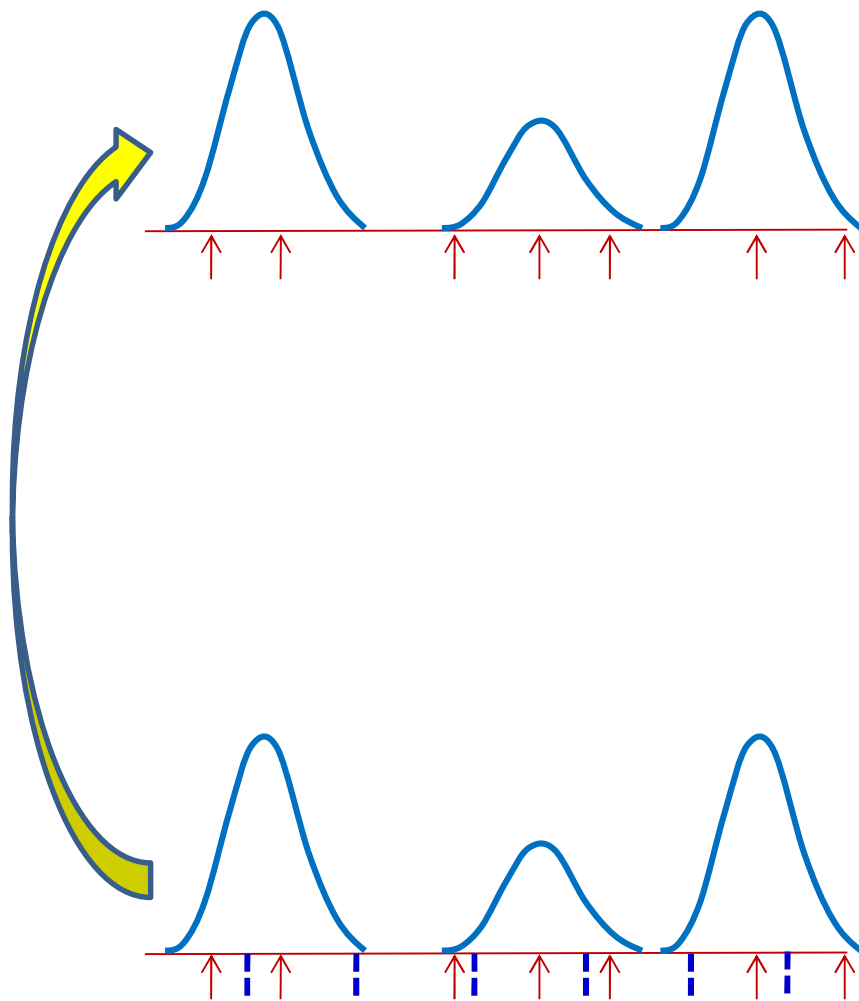
- Randomly initialize quantization points
 - Right column entries of quantization table
- Assign all training points to the nearest quantization point
 - Draw boundaries

Lloyd Quantizer



- Randomly initialize quantization points
 - Right column entries of quantization table
- Assign all training points to the nearest quantization point
 - Draw boundaries
- Reestimate quantization points

Lloyd Quantizer



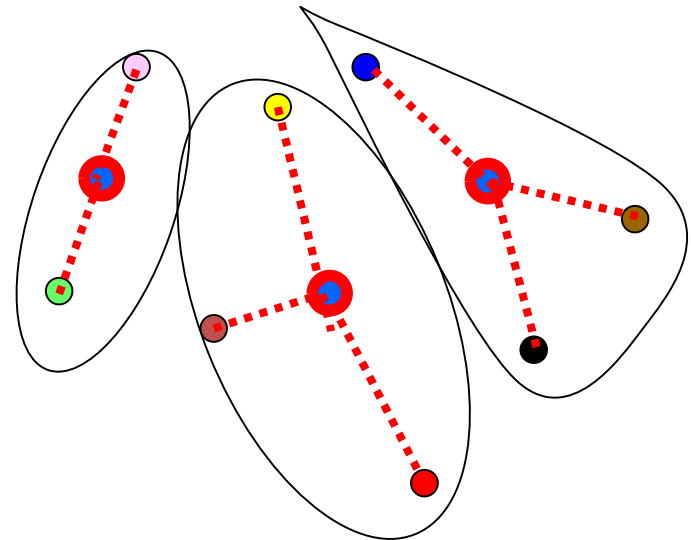
- Randomly initialize quantization points
 - Right column entries of quantization table
- Assign all training points to the nearest quantization point
 - Draw boundaries
- Reestimate quantization points
- Iterate until convergence

Generalized Lloyd Algorithm: K-means clustering

- K means is an iterative algorithm for clustering **vector** data
 - McQueen, J. 1967. “Some methods for classification and analysis of multivariate observations.” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297
- General procedure:
 - Initially group data into the required number of clusters somehow (initialization)
 - Assign each data point to the closest cluster
 - Once all data points are assigned to clusters, redefine clusters
 - Iterate

K-means

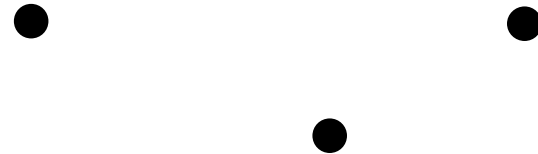
- Problem: Given a set of data vectors, find natural clusters
- Clustering criterion is **scatter**: distance from the centroid
 - Every cluster has a centroid
 - The centroid represents the cluster
- **Definition:** The **centroid** is the weighted mean of the cluster
 - Weight = 1 for basic scheme



$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$

K-means

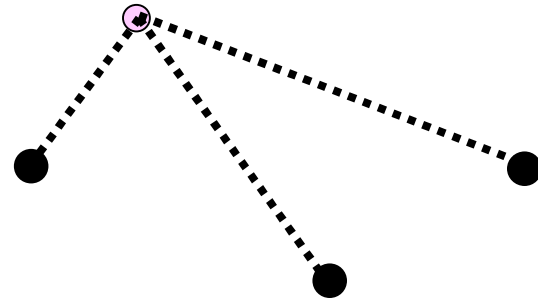
1. Initialize a set of centroids randomly



K-means

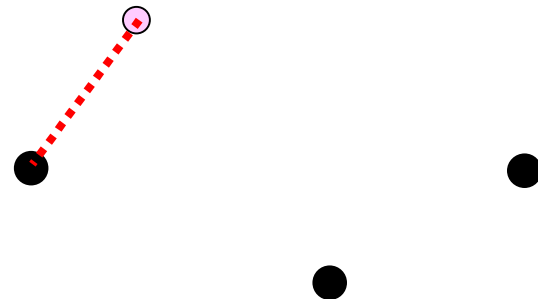
1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster

- $d_{cluster} = \text{distance}(x, m_{cluster})$



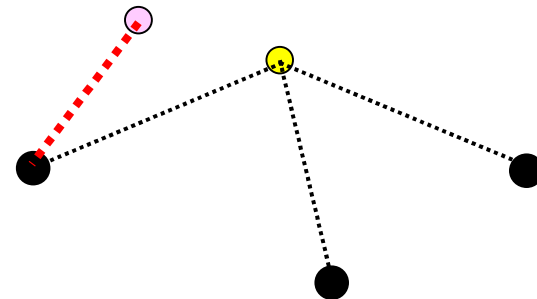
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



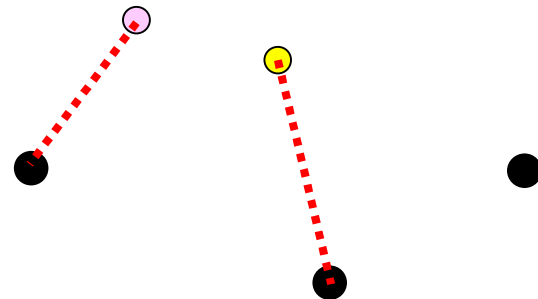
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



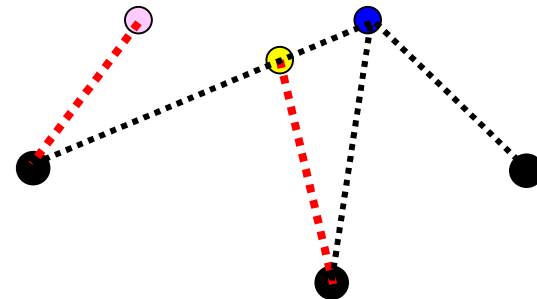
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



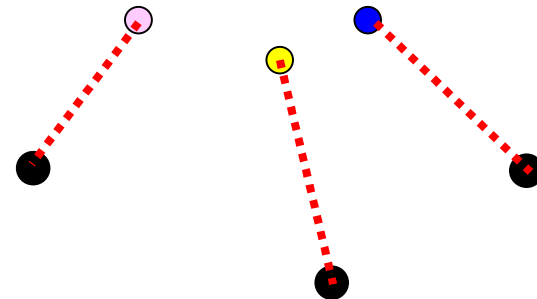
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



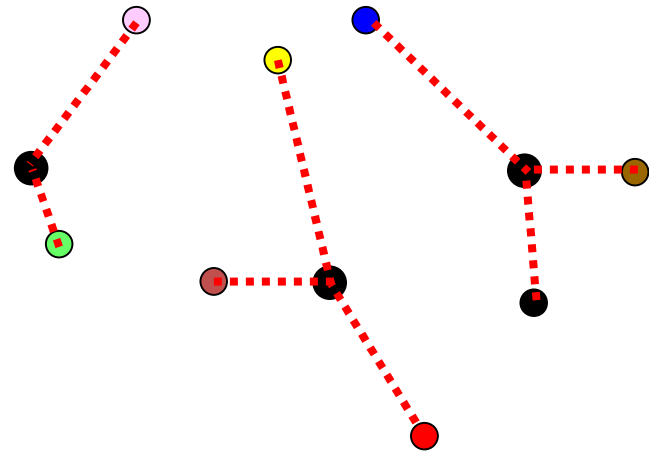
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



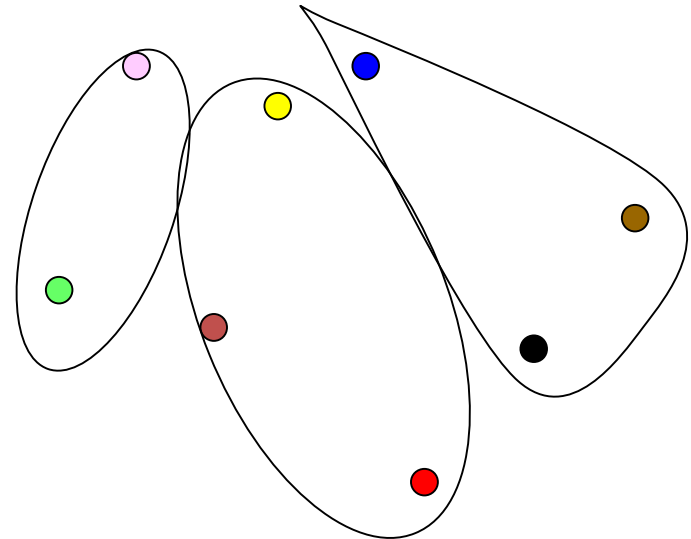
K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



K-means

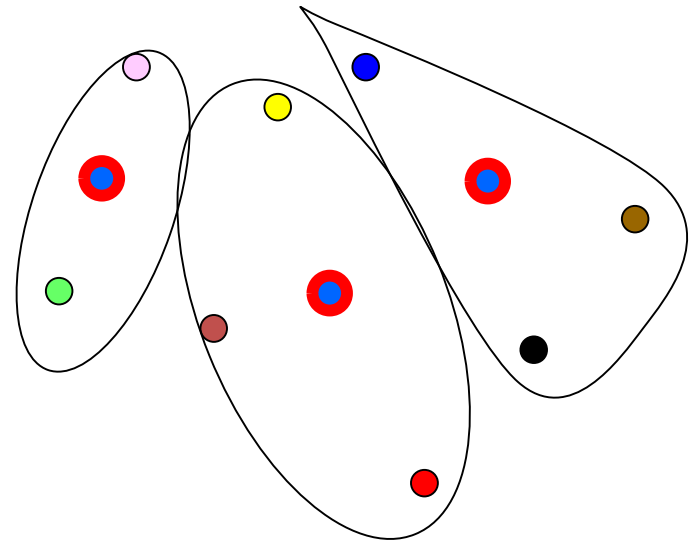
1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum
4. When all data points are clustered, recompute centroids

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$

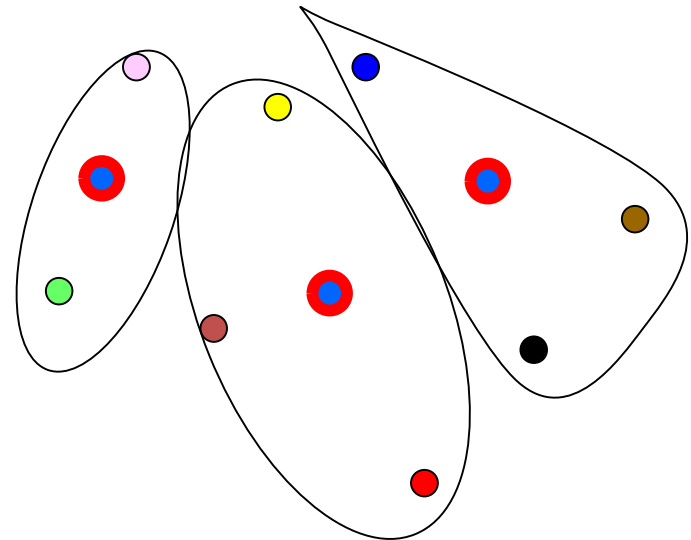


K-means

1. Initialize a set of centroids randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum
4. When all data points are clustered, recompute centroids

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$

5. If not converged, go back to 2



K-Means comments

- The distance metric determines the clusters
 - In the original formulation, the distance is L_2 distance
 - Euclidean norm, $w_i = 1$

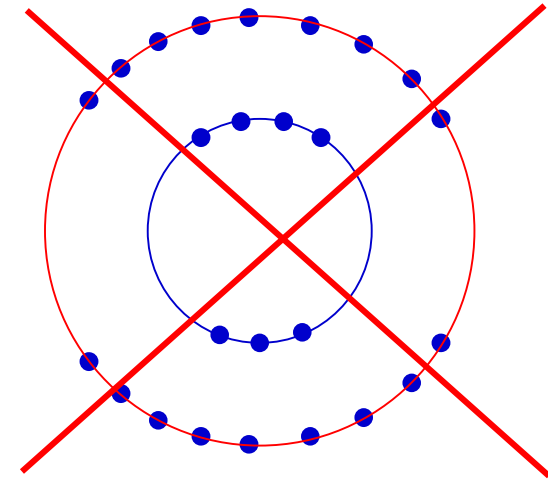
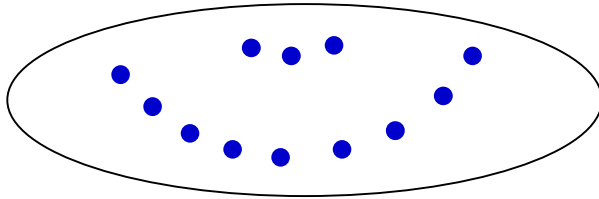
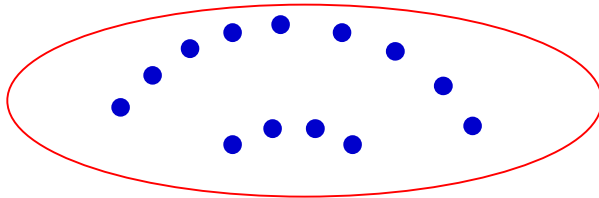
$$\text{distance}_{cluster}(x, m_{cluster}) = \|x - m_{cluster}\|_2$$

$$m_{cluster} = \frac{1}{N_{cluster}} \sum_{i \in cluster} x_i$$

- If we replace every x by $m_{cluster}(x)$, we get *Vector Quantization*
- K-means is an instance of *generalized* EM
- Not guaranteed to converge for all distance metrics

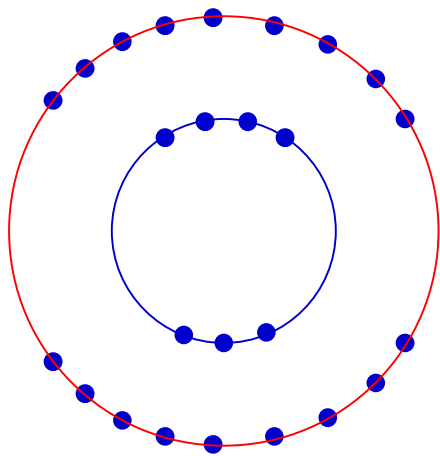
Poll 3

Non-Euclidean clusters

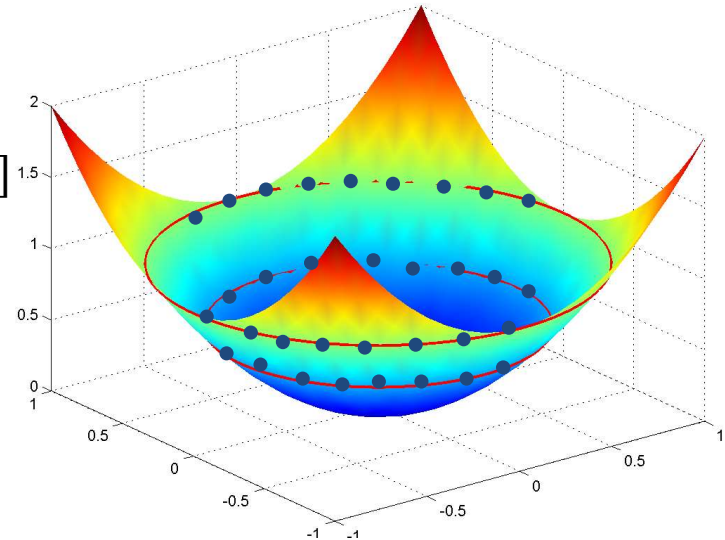


- Basic K-means results in good clusters in Euclidean spaces
 - Alternately stated, will only find clusters that are “good” in terms of Euclidean distances
- Will not find other types of clusters

Non-Euclidean clusters

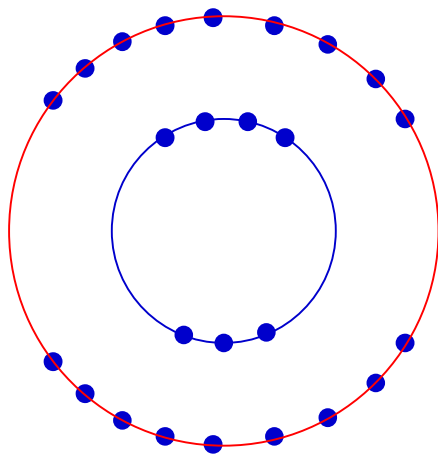


$$f([x,y]) \rightarrow [x,y,z]$$
$$x = x$$
$$y = y$$
$$z = \alpha(x^2 + y^2)$$

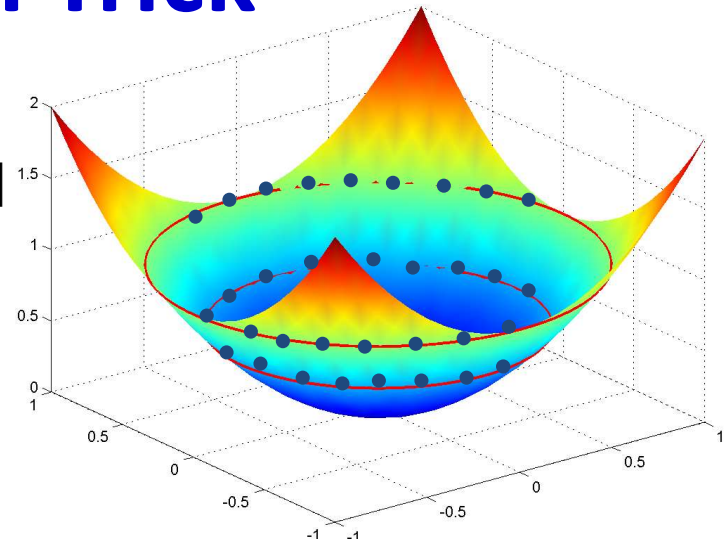


- For other forms of clusters we must modify the distance measure
 - E.g. distance from a circle
 - But finding nearest values and representative vectors (centroids) becomes hard
- May be viewed as a distance in a higher dimensional space
 - I.e. *Kernel* distances
 - *Kernel* K-means
- Other related clustering mechanisms:
 - Spectral clustering
 - Non-linear weighting of adjacency
 - Normalized cuts..

The Kernel Trick



$$\begin{aligned} f([x,y]) &\rightarrow [x,y,z] \\ x &= x \\ y &= y \\ z &= \alpha(x^2 + y^2) \end{aligned}$$



- Transform the data into a synthetic higher-dimensional space where the desired patterns become natural clusters based on *Euclidean* distance
 - E.g. the quadratic transform above
- Problem: What is the function/space?
- Problem: Distances in higher dimensional-space are more expensive to compute
 - Yet only carry the same information in the lower-dimensional space

Distance in higher-dimensional space

- Transform data \mathbf{x} through a *possibly unknown* function $\Phi(\mathbf{x})$ into a higher (potentially infinite) dimensional space
 - $\mathbf{z} = \Phi(\mathbf{x})$
- The distance between two points is computed in the higher-dimensional space
 - $d(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{z}_1 - \mathbf{z}_2||^2 = ||\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)||^2$
- $d(\mathbf{x}_1, \mathbf{x}_2)$ can be computed without computing \mathbf{z}
 - Since it is a direct function of \mathbf{x}_1 and \mathbf{x}_2

Distance in higher-dimensional space

- Distance in lower-dimensional space: A combination of dot products
 - $||\mathbf{z}_1 - \mathbf{z}_2||^2 = (\mathbf{z}_1 - \mathbf{z}_2)^T (\mathbf{z}_1 - \mathbf{z}_2) = \mathbf{z}_1 \cdot \mathbf{z}_1 + \mathbf{z}_2 \cdot \mathbf{z}_2 - 2 \mathbf{z}_1 \cdot \mathbf{z}_2$
- Distance in higher-dimensional space
 - $d(\mathbf{x}_1, \mathbf{x}_2) = ||\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)||^2$
 $= \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_1) + \Phi(\mathbf{x}_2) \cdot \Phi(\mathbf{x}_2) - 2 \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$
- $d(\mathbf{x}_1, \mathbf{x}_2)$ can be computed without knowing $\Phi(\mathbf{x})$ if:
 - $\Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$ can be computed for any \mathbf{x}_1 and \mathbf{x}_2 without knowing $\Phi(\cdot)$

The Kernel function

- A kernel function $K(\mathbf{x}_1, \mathbf{x}_2)$ is a function such that:
 - $K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$
- Once such a kernel function is found, the distance in higher-dimensional space can be found in terms of the kernels
 - $d(\mathbf{x}_1, \mathbf{x}_2) = ||\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)||^2$
$$= \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_1) + \Phi(\mathbf{x}_2) \cdot \Phi(\mathbf{x}_2) - 2 \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$$
$$= K(\mathbf{x}_1, \mathbf{x}_1) + K(\mathbf{x}_2, \mathbf{x}_2) - 2K(\mathbf{x}_1, \mathbf{x}_2)$$
- But what is $K(\mathbf{x}_1, \mathbf{x}_2)$?

A property of the dot product

- For any vector \mathbf{v} , $\mathbf{v}^T \mathbf{v} = ||\mathbf{v}||^2 \geq 0$
 - This is just the length of \mathbf{v} and is therefore non-negative
- For any vector $\mathbf{u} = \sum_i a_i \mathbf{v}_i$, $||\mathbf{u}||^2 \geq 0$
 - $\Rightarrow (\sum_i a_i \mathbf{v}_i)^T (\sum_i a_i \mathbf{v}_i) \geq 0$
 - $\Rightarrow \sum_i \sum_j a_i a_j \mathbf{v}_i \cdot \mathbf{v}_j \geq 0$
- This holds for ANY real $\{a_1, a_2, \dots\}$

The Mercer Condition

- **If** $\mathbf{z} = \Phi(\mathbf{x})$ is a high-dimensional vector derived from \mathbf{x} **then** for all real $\{a_1, a_2, \dots\}$ and any set $\{\mathbf{z}_1, \mathbf{z}_2, \dots\} = \{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \dots\}$
 - $\sum_i \sum_j a_i a_j \mathbf{z}_i \cdot \mathbf{z}_j \geq 0$
 - $\sum_i \sum_j a_i a_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \geq 0$
- If $K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$
 $\Rightarrow \sum_i \sum_j a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- Any function $K()$ that satisfies the above condition is a valid kernel function

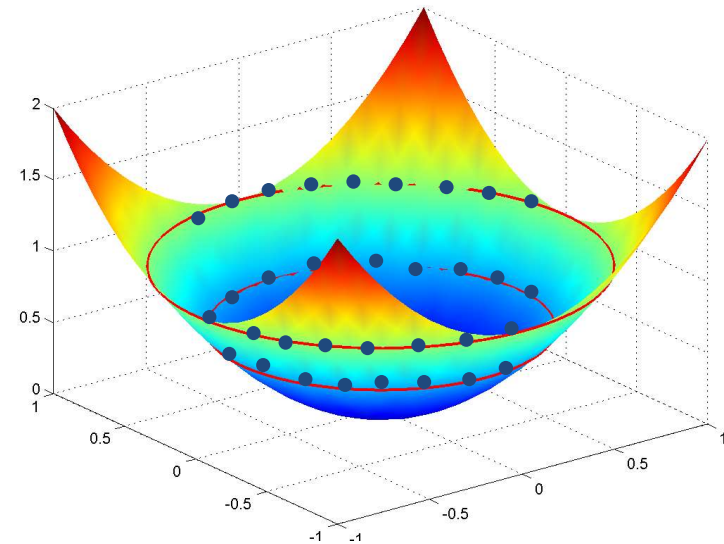
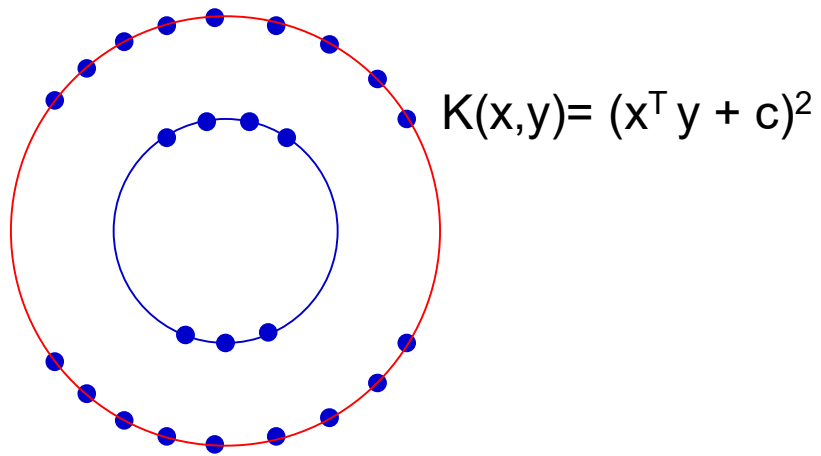
The Mercer Condition

- $K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$
 $\Rightarrow \sum_i \sum_j a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- **A corollary:** If any kernel $K(\cdot)$ satisfies the Mercer condition
 $d(\mathbf{x}_1, \mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_1) + K(\mathbf{x}_2, \mathbf{x}_2) - 2K(\mathbf{x}_1, \mathbf{x}_2)$
satisfies the following requirements for a “distance”
 - $d(\mathbf{x}, \mathbf{x}) = 0$
 - $d(\mathbf{x}, \mathbf{y}) \geq 0$
 - $d(\mathbf{x}, \mathbf{w}) + d(\mathbf{w}, \mathbf{y}) \geq d(\mathbf{x}, \mathbf{y})$

Typical Kernel Functions

- Linear: $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y} + c$
- Polynomial $K(\mathbf{x}, \mathbf{y}) = (a\mathbf{x}^T \mathbf{y} + c)^n$
- Gaussian: $K(\mathbf{x}, \mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}||^2 / \sigma^2)$
- Exponential: $K(\mathbf{x}, \mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}|| / \lambda)$
- Several others
 - Choosing the right Kernel with the right parameters for your problem is an artform

Kernel K-means



- Perform the K-mean in the Kernel space
 - The space of $\mathbf{z} = \Phi(\mathbf{x})$
- The algorithm..

The mean of a cluster

- The average value of the points in the cluster *computed in the high-dimensional space*

$$m_{cluster} = \frac{1}{N_{cluster}} \sum_{i \in cluster} \Phi(x_i)$$

- Alternately the weighted average

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i \Phi(x_i) = C \sum_{i \in cluster} w_i \Phi(x_i)$$

The mean of a cluster

- The average value of the points in the cluster *computed in the high-dimensional space*

$$m_{cluster} = \frac{1}{N_{cluster}} \sum_{i \in cluster} \Phi(x_i)$$

RECALL: We may never actually be able to compute this mean because $\Phi(x)$ is not known

- Alternately the weighted average

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i \Phi(x_i) = C \sum_{i \in cluster} w_i \Phi(x_i)$$

K-means

- Initialize the clusters with a random set of K points

– $N_{cluster}$ is no. of points in cluster

$$m_{cluster} = \frac{1}{N_{cluster}} \sum_{i \in cluster} \Phi(x_i)$$

- For each data point x , find the closest cluster

$$cluster(x) = \min_{cluster} d(x, cluster) = \min_{cluster} \| \Phi(x) - m_{cluster} \|^2$$

$$d(x, cluster) = \| \Phi(x) - m_{cluster} \|^2 = \left(\Phi(x) - \frac{1}{N_{cluster}} \sum_{i \in cluster} \Phi(x_i) \right)^T \left(\Phi(x) - \frac{1}{N_{cluster}} \sum_{i \in cluster} \Phi(x_i) \right)$$

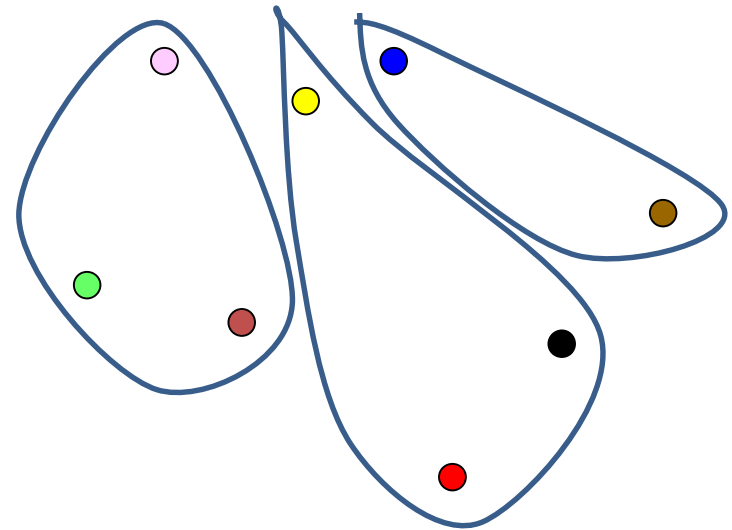
$$= \left(\Phi(x)^T \Phi(x) - \frac{2}{N_{cluster}} \sum_{i \in cluster} \Phi(x)^T \Phi(x_i) + \frac{1}{N_{cluster}^2} \sum_{i \in cluster} \sum_{j \in cluster} \Phi(x_i)^T \Phi(x_j) \right)$$

$$= K(x, x) - \frac{2}{N_{cluster}} \sum_{i \in cluster} K(x, x_i) + \frac{1}{N_{cluster}^2} \sum_{i \in cluster} \sum_{j \in cluster} K(x_i, x_j)$$

Computed entirely using only the kernel function!

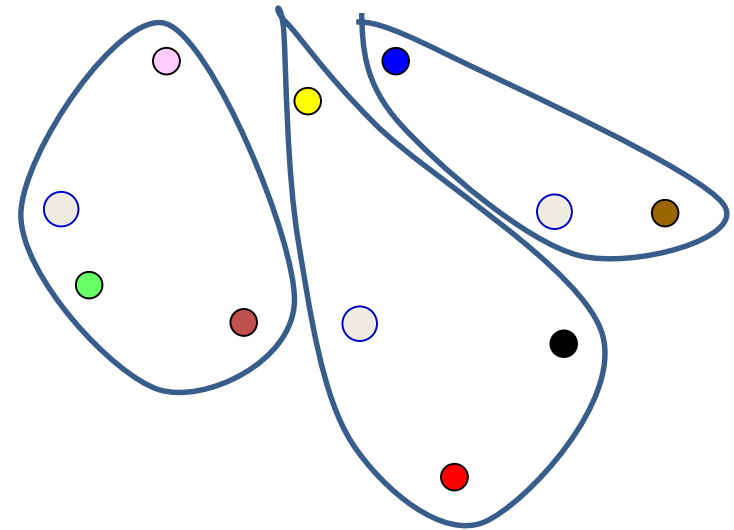
K-means

1. Initialize a set of *clusters* randomly



K-means

1. Initialize a set of *clusters* randomly



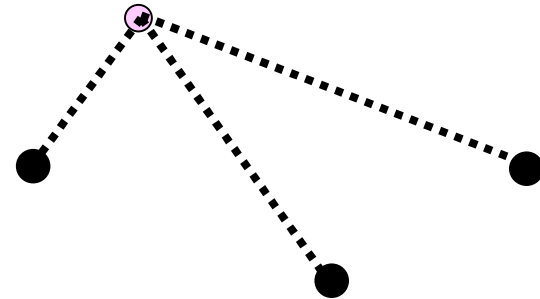
The centroids are *virtual*:
we don't actually compute
them explicitly!

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$

K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster

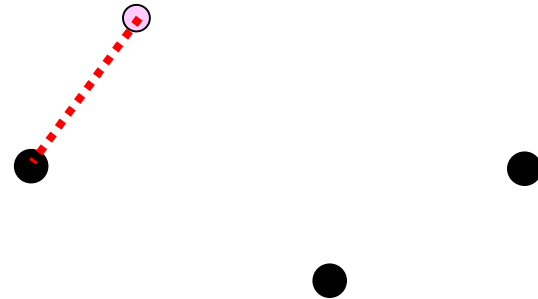
- $d_{cluster} = \mathbf{distance}(x, m_{cluster})$



$$d_{cluster} = K(x, x) - 2C \sum_{i \in cluster} w_i K(x, x_i) + C^2 \sum_{i \in cluster} \sum_{j \in cluster} w_i w_j K(x_i, x_j)$$

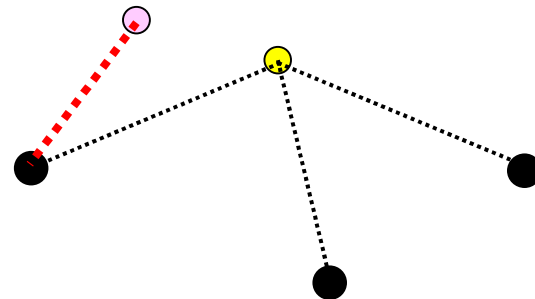
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



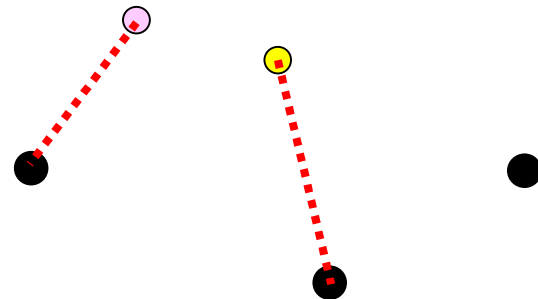
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



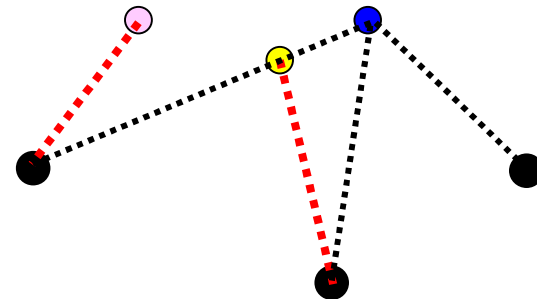
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



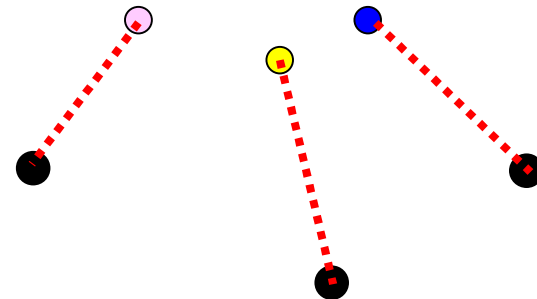
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



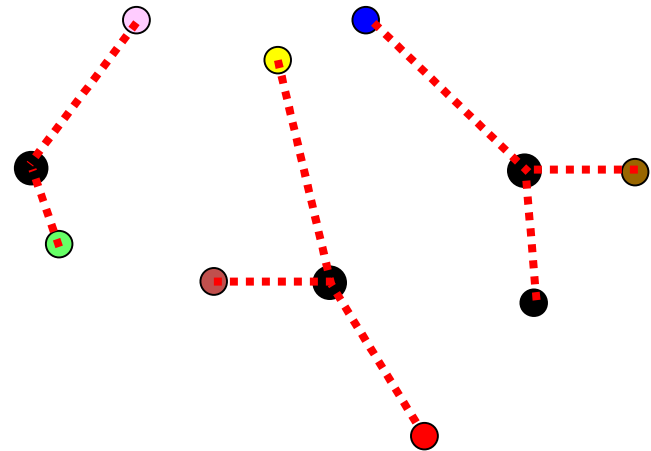
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



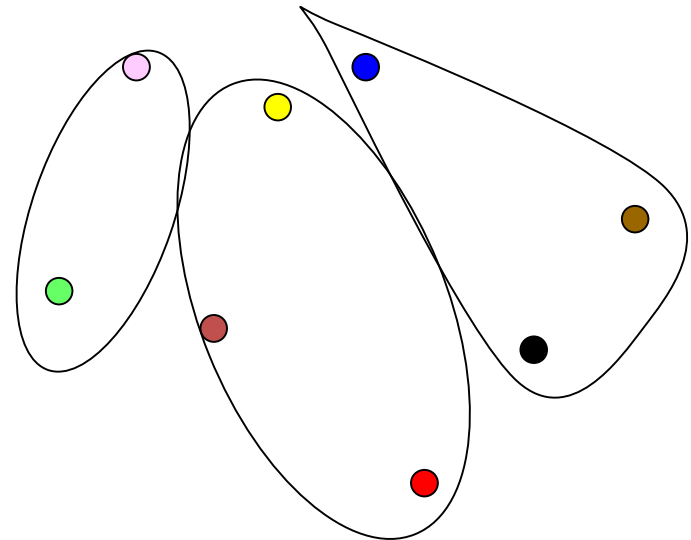
K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



K-means

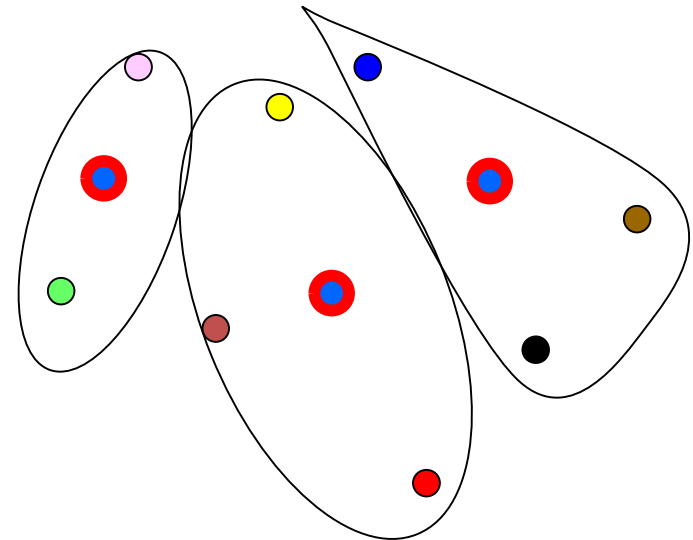
1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum



K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum
4. *When all data points are clustered, recompute centroids*

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$



- We do not explicitly compute the means
- May be impossible – we do not know the high-dimensional space
- We only know how to compute inner products in it

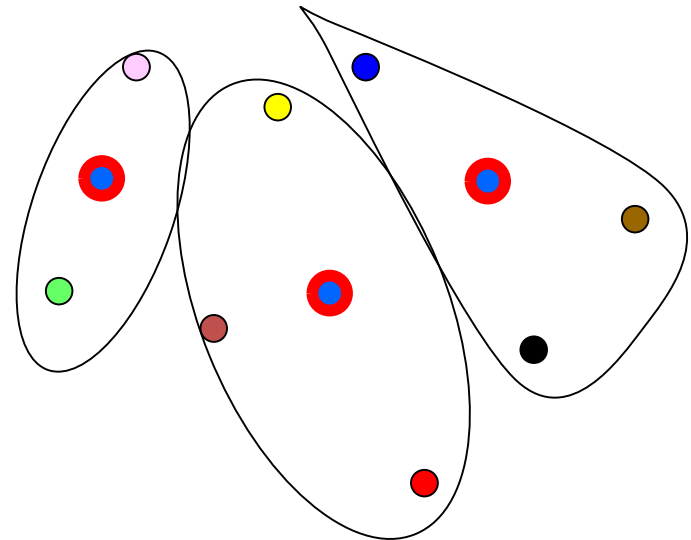
Kernel K-means

1. Initialize a set of clusters randomly
2. For each data point x , find the distance from the centroid for each cluster
 - $d_{cluster} = \text{distance}(x, m_{cluster})$
3. Put data point in the cluster of the closest centroid
 - Cluster for which $d_{cluster}$ is minimum

4. *When all data points are clustered, recompute centroids*

$$m_{cluster} = \frac{1}{\sum_{i \in cluster} w_i} \sum_{i \in cluster} w_i x_i$$

5. If not converged, go back to 2



- We do not explicitly compute the means
- May be impossible – we do not know the high-dimensional space
- We only know how to compute inner products in it

Poll 4

How many clusters?

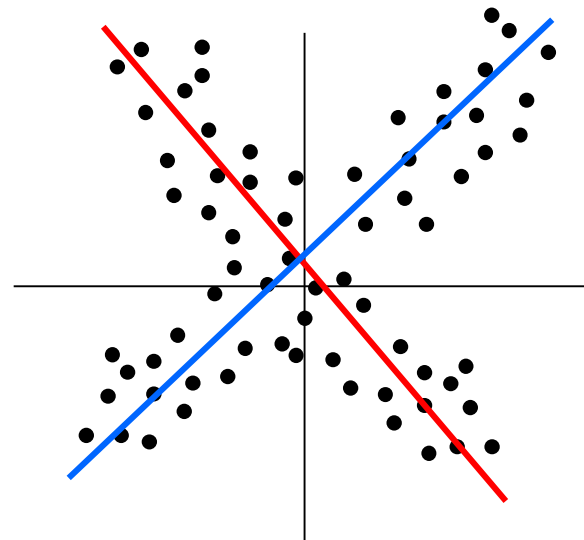
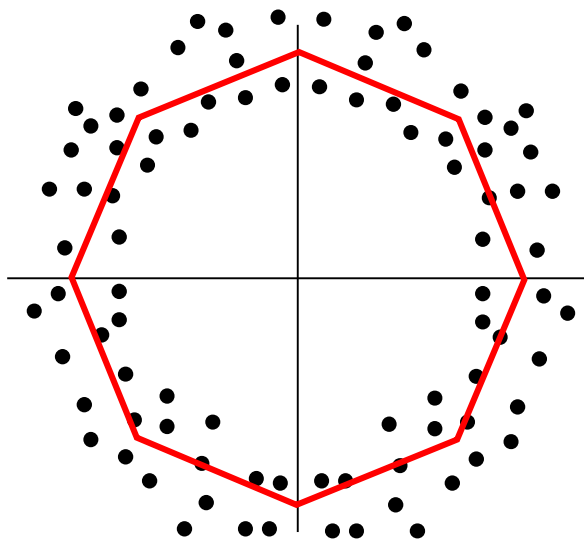
- Assumptions:
 - Dimensionality of kernel space $>$ no. of clusters
 - Clusters represent separate *directions* in Kernel spaces
- Kernel correlation matrix \mathbf{K}
 - $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$
- Find Eigen values Λ and Eigen vectors \mathbf{e} of kernel matrix
 - No. of clusters = no. of dominant $\lambda_i (1^T \mathbf{e}_i)$ terms

Spectral Methods

- “Spectral” methods attempt to find “principal” subspaces of the high-dimensional kernel space
- Clustering is performed in the principal subspaces
 - Normalized cuts
 - Spectral clustering
- Involves finding Eigenvectors and Eigen values of Kernel matrix
- Fortunately, provably analogous to Kernel K-means

Other clustering methods

- Regression based clustering
- Find a regression representing each cluster
- Associate each point to the cluster with the best regression
 - Related to kernel methods



Clustering..

- Many many other variants
 - Many applications..
 - Important: Appropriate choice of feature
 - Appropriate choice of feature may eliminate need for kernel trick..
- Key Features:
 - Identifies latent structure in the distribution of the data
 - Provides an L2-sense optimal quantized representation of the data
 - We will build on this in the next class