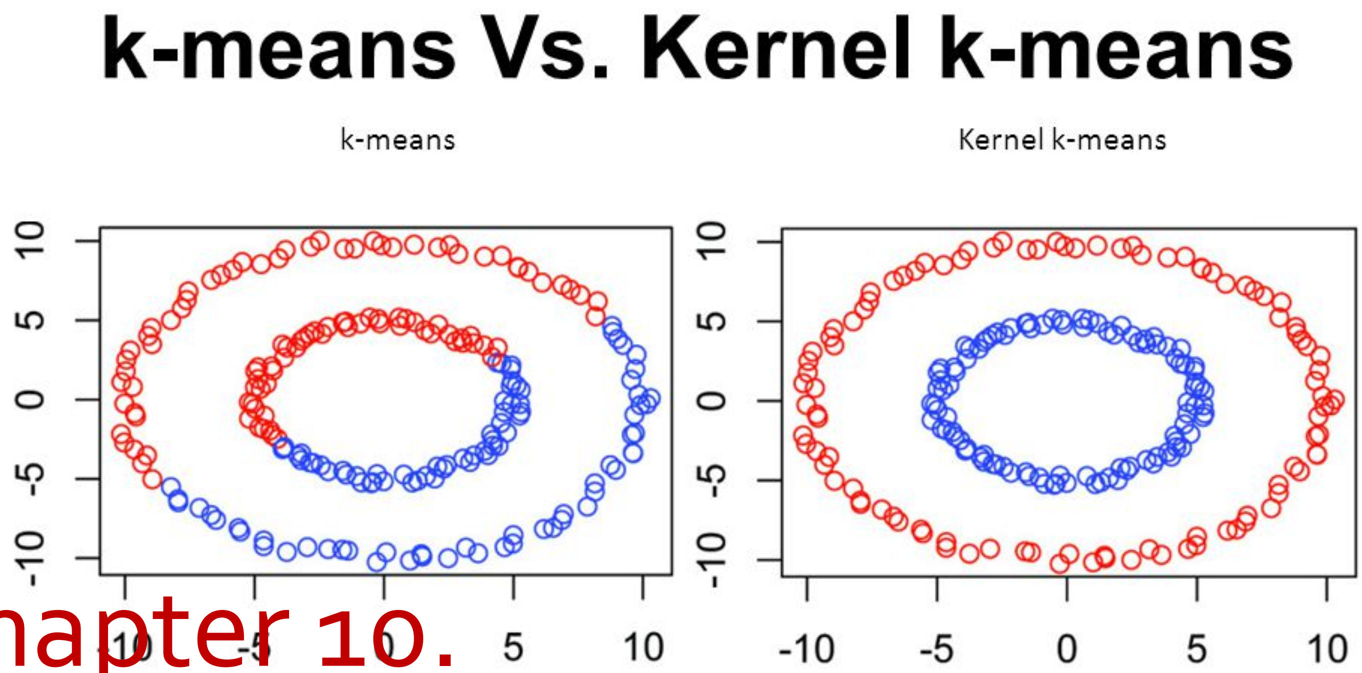


Chapter 10. Cluster Analysis: Kernel K-Means

Meng Jiang

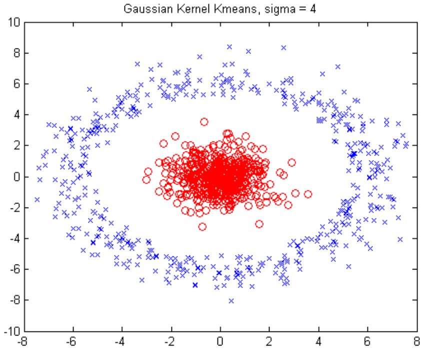
CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Partitioning-Based Clustering Methods

- Basic Concepts of Partitioning Algorithms
- The K-Means Clustering Method
- Initialization of K-Means Clustering
- The K-Medoids Clustering Method
- The K-Medians and K-Modes Clustering Methods
- **The Kernel K-Means Clustering Method**

Kernel K-Means Clustering



- Kernel K-Means can be used to detect non-convex clusters
  - K-Means can only detect clusters that are linearly separable
- Idea: Project data onto the high-dimensional kernel space, and then perform K-Means clustering
  - Map data points in the input space onto a high-dimensional feature space using the kernel function
  - Perform K-Means on the mapped feature space
- Computational complexity is higher than K-Means
  - Need to compute and store n x n kernel matrix generated from the kernel function on the original data
- The widely studied spectral clustering can be considered as a variant of Kernel K-Means clustering

Kernel Functions and Kernel K-Means Clustering

Kernel matrix:
inner-product(i, j) matrix √
similarity
distance(i, j) matrix ×

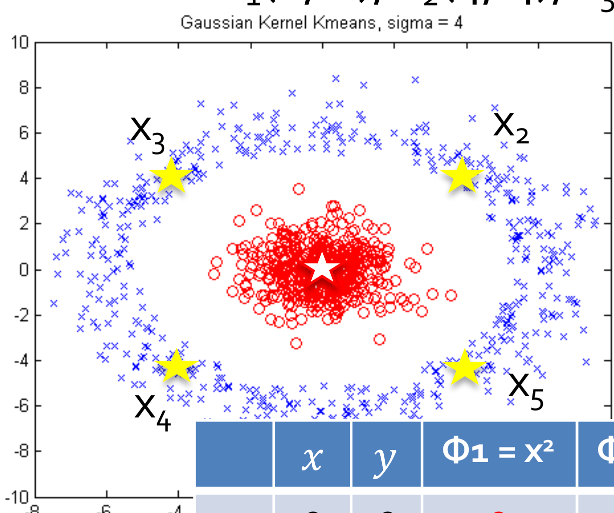
- Typical kernel functions:
  - Polynomial kernel of degree h: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i \cdot \boldsymbol{X}_j + c)^h$
  - Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$
  - Sigmoid kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \tanh(\kappa \boldsymbol{X}_i \cdot \boldsymbol{X}_j - \delta)$
- The formula for **kernel matrix** K for any two points $x_i, x_j \in C_k$ is $K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$
- The SSE criterion of *kernel K-means*: $SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| \phi(x_i) - c_k \|^2$
  - The formula for the cluster centroid: $$c_k = \frac{\sum_{x_i \in C_k} \phi(x_i)}{|C_k|}$$

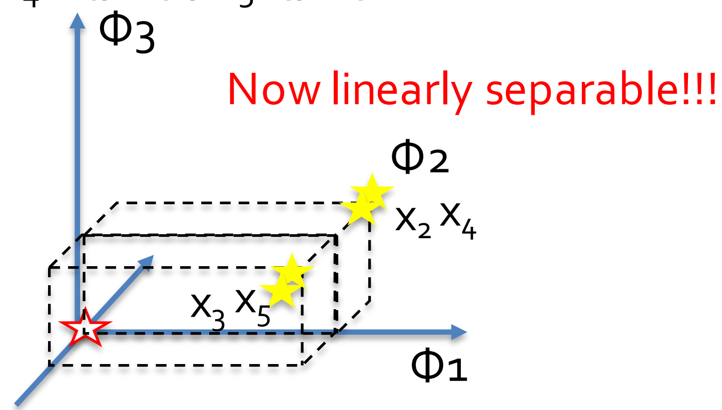
Inderjit S. Dhillon, Yiqiang Guan, Brian Kulis (Univ. of Texas at Austin). "Kernel K-means, Spectral Clustering and Normalized Cuts", KDD 04.

http://www.cs.utexas.edu/users/inderjit/public_papers/kdd_spectral_kernelkmeans.pdf

Example: Kernel Functions and Kernel K-Means Clustering

- Polynomial kernel of degree h=2: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \boldsymbol{X}_i \cdot \boldsymbol{X}_j{}^2 \rightarrow \phi(x, y) = (x^2, \sqrt{2}xy, y^2)$
- Suppose there are 5 original 2-dimensional points:
  - $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$





	x	y	Φ1 = x²	Φ2 = xy	ϕ3 = y²	$K(x_i,x_1)$	$K(x_i,x_2)$	$K(x_i,x_3)$	$K(x_i,x_4)$	$K(x_i,x_5)$
x_1	0	0	0	0	0	0	0	0	0	0
x_2	4	4	16	$16\sqrt{2}$	16	0	32^2	0	32^2	0
x_3	-4	4	16	$-16\sqrt{2}$	16	0	0	32^2	0	32^2
x_4	-4	-4	16	$16\sqrt{2}$	16	0	32^2	0	32^2	0
x_5	4	-4	16	$-16\sqrt{2}$	16	0	0	32^2	0	32^2

Example: Kernel Functions and Kernel K-Means Clustering

- Suppose there are 5 original 2-dimensional points:
  - $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$

Original Space

	x	y	$(\boldsymbol{x_i}, \boldsymbol{x_1})$	$(\boldsymbol{x_i}, \boldsymbol{x_2})$	$(\boldsymbol{x_i}, \boldsymbol{x_3})$	$(\boldsymbol{x_i}, \boldsymbol{x_4})$	$(\boldsymbol{x_i}, \boldsymbol{x_5})$
x_1	0	0	0	0	0	0	0
x_2	4	4	0	32	0	-32	0
x_3	-4	4	0	0	32	0	-32
x_4	-4	-4	0	-32	0	32	0
x_5	4	-4	0	0	-32	0	32

Example: Kernel Functions and Kernel K-Means Clustering

- Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X_i}, \boldsymbol{X_j}) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$
- Suppose there are 5 original 2-dimensional points:
  - $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$

$$K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$$

- If we set σ to 4, we will have the following points in the kernel space
  - E.g., $||x_1 - x_2||^2 = (0-4)^2 + (0-4)^2 = 32$, therefore, $K(x_1, x_2) = e^{-\frac{32}{2\cdot 4^2}} = e^{-1}$

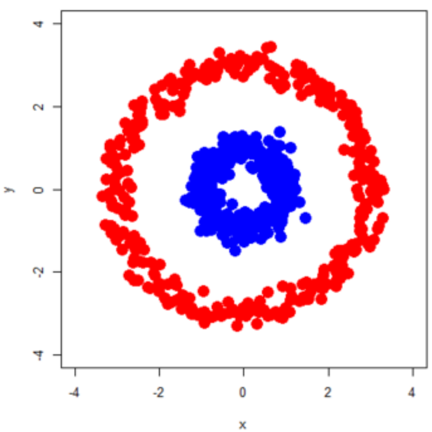
Original Space

RBF Kernel Space ($\sigma = 4$)

	x	y	$K(x_i, x_1)$	$K(x_i, x_2)$	$K(x_i, x_3)$	$K(x_i, x_4)$	$K(x_i, x_5)$
x_1	0	0	1	$e^{-\frac{4^2+4^2}{2\cdot 4^2}} = e^{-1}$	e^{-1}	e^{-1}	e^{-1}
x_2	4	4	e^{-1}	1	e^{-2}	e^{-4}	e^{-2}
x_3	-4	4	e^{-1}	e^{-2}	1	e^{-2}	e^{-4}
x_4	-4	-4	e^{-1}	e^{-4}	e^{-2}	1	e^{-2}
x_5	4	-4	e^{-1}	e^{-2}	e^{-4}	e^{-2}	1

Example: Kernel Functions and Kernel K-Means Clustering

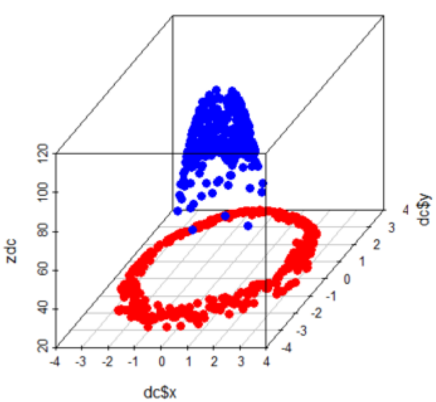
- Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X_i}, \boldsymbol{X_j}) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$

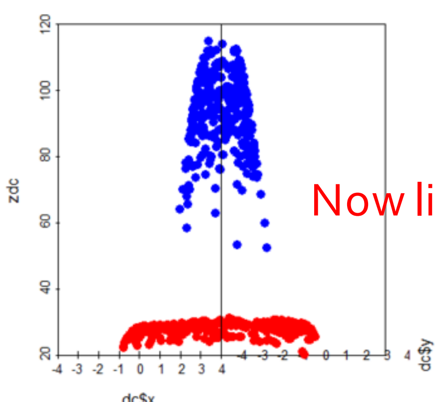


$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \mathbf{a}_l \right\|_2^2$$

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \phi(\mathbf{a}_j) - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \phi(\mathbf{a}_l) \right\|_2^2$$

$$\kappa(\mathbf{a}_i, \mathbf{a}_j) = \langle \phi(\mathbf{a}_i), \phi(\mathbf{a}_j) \rangle.$$



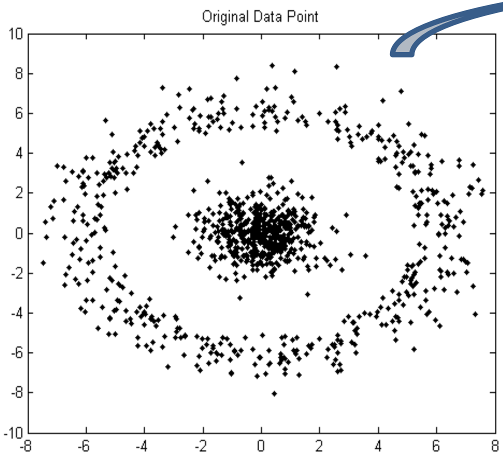


Now linearly separable!!!

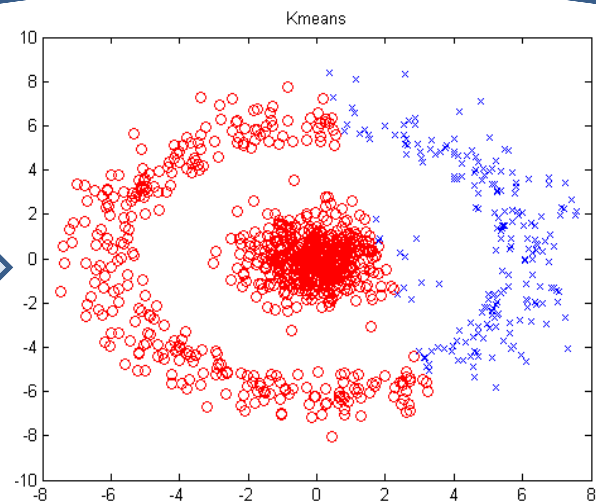
$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \cdots$$

Countless new features in RBF kernel space…

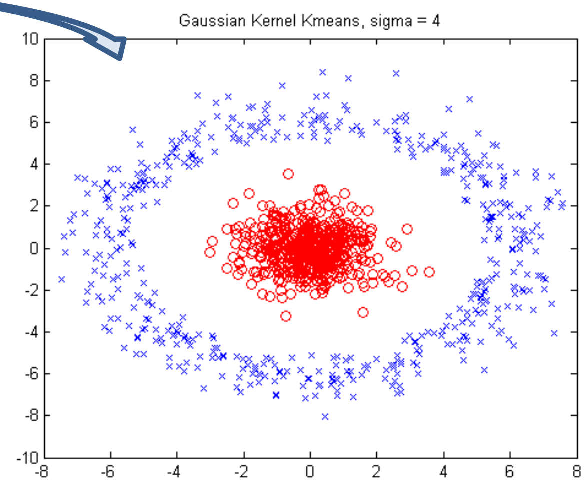
Example: Kernel K-Means Clustering



The original data set



The result of K-Means clustering



The result of Gaussian Kernel K-Means clustering

- The above data set cannot generate quality clusters by K-Means since it contains non-convex clusters
- Gaussian RBF Kernel transformation maps data to a kernel matrix K for any two points x_i, x_j: $K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$ and Gaussian kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$
- K-Means clustering is conducted on the mapped data, generating quality clusters

References: (II) Partitioning Methods

- J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, 1967
- S. Lloyd. Least Squares Quantization in PCM. IEEE Trans. on Information Theory, 28(2), 1982
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988
- R. Ng and J. Han. Efficient and Effective Clustering Method for Spatial Data Mining. VLDB'94
- B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural computation, 10(5):1299–1319, 1998
- I. S. Dhillon, Y. Guan, and B. Kulis. Kernel K-Means: Spectral Clustering and Normalized Cuts. KDD'04
- D. Arthur and S. Vassilvitskii. K-means++: The Advantages of Careful Seeding. SODA'07
- C. K. Reddy and B. Vinzamuri. A Survey of Partitional and Hierarchical Clustering Algorithms, in (Chap. 4) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014