CS195-5 : Introduction to Machine Learning
Lecture 25

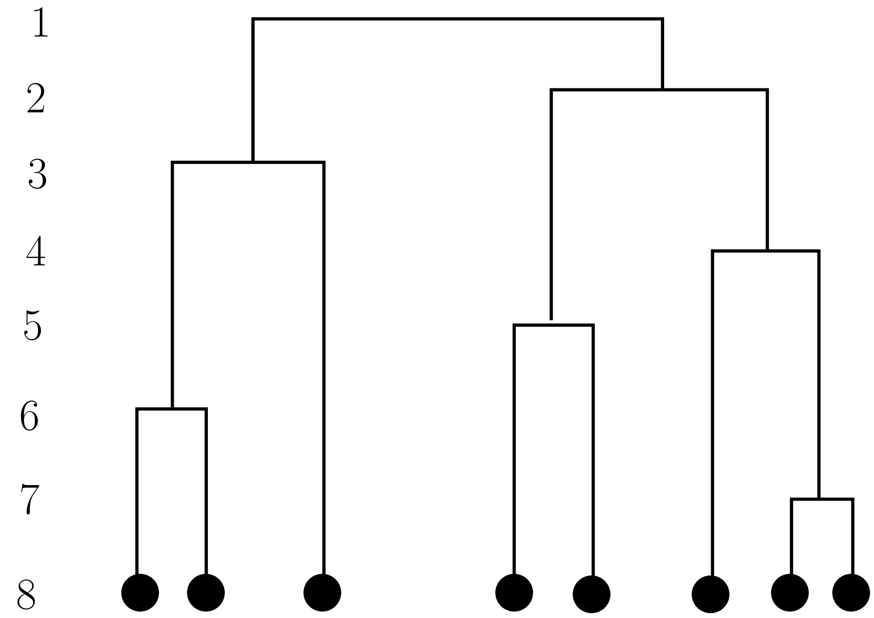
Greg Shakhnarovich

November 13, 2006

Announcements

- Plan for problem sets:
  - PS 5 due 11/22 (before Thanksgiving break)
  - PS 6 due 12/8
  - PS 7: no grade, solutions will be available on 12/13

Review: hiearchical clustering



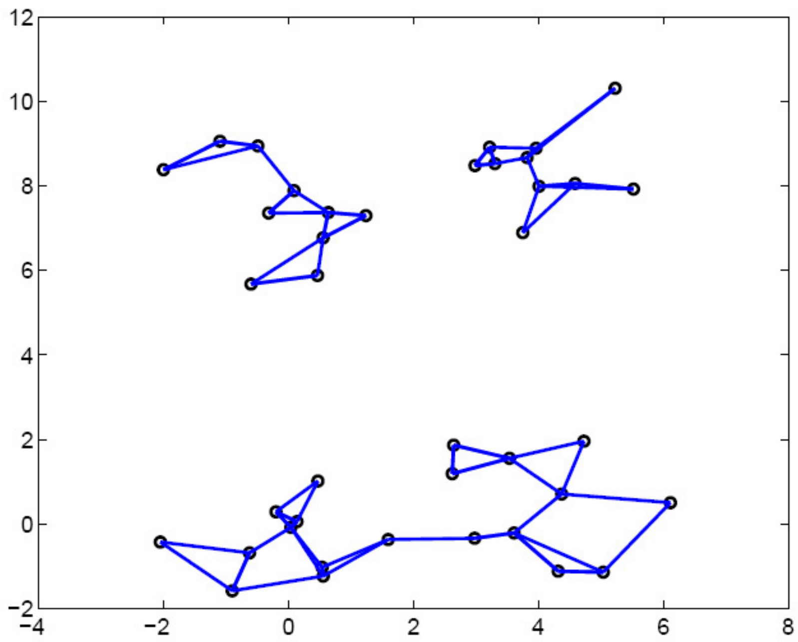
Allgomerative clustering:

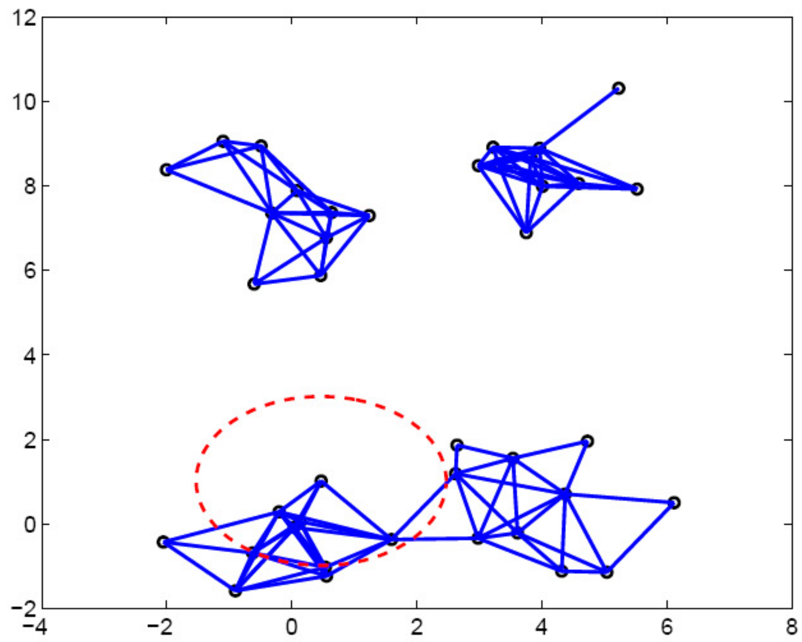
- Start with N singleton clusters
- At each level merge two clusters

- Single linkage: $D(A, B) = \min_{\mathbf{a} \in A, \mathbf{b} \in B} D(\mathbf{a}, \mathbf{b})$
- Average linkage: $D(A, B) = \frac{1}{|A||B|} \sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} D(\mathbf{a}, \mathbf{b})$
- Complete linkage: $D(A, B) = \max_{\mathbf{a} \in A, \mathbf{b} \in B} D(\mathbf{a}, \mathbf{b})$

Spectral clustering

- Suppose we have a $N \times N$ *distance matrix*
- We can represent the data as a graph:
  - N vertices,
  - edges corresponding to nearest neighbors.





Random walk model

- Assign weights to edges:

$$W_{ij} = \begin{cases} \exp(-\beta\|\mathbf{x}_i - \mathbf{x}_j\|) & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ connected,} \\ 0 & \text{otherwise} \end{cases}$$

- The weight of a path $\mathbf{x}_1 \to \mathbf{x}_2 \to \ldots \to \mathbf{x}_n$ is

$$W_{12} \cdot W_{23} \cdots W_{n-1,n} \;=\; exp\left(-\beta \sum_{i=1}^{n-1} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|\right)$$

Spectral clustering: intuition

- The idea behind spectral clustering: imagine a *random walk* with probability of step $i \to j$ given by the *transition matrix* $\mathbf{P}$

$$P_{ij} = \frac{W_{ij}}{\sum_l W_{il}}.$$

  – If we start within a cluster, we will likely remain within that cluster for a long time.

Properties of the random walk

- If we start at i_0, where will we end up after t steps?

$$
\begin{aligned}
i_1 &\sim P_{i_0 i_1}\,, \\
i_2 &\sim \sum_{i_1} P_{i_0 i_1} P_{i_1 i_2}
\end{aligned}
$$

Properties of the random walk

- If we start at i_0, where will we end up after t steps?

$$
\begin{aligned}
i_1 &\sim P_{i_0 i_1}\,, \\
i_2 &\sim \sum_{i_1} P_{i_0 i_1} P_{i_1 i_2} = \left(\mathbf{P}^2\right)_{i_0 i_2}\,,
\end{aligned}
$$

Properties of the random walk

- If we start at i_0, where will we end up after t steps?

$$\begin{aligned} i_1 &\sim P_{i_0 i_1}\,, \\ i_2 &\sim \sum_{i_1} P_{i_0 i_1} P_{i_1 i_2} = \left(\mathbf{P}^2\right)_{i_0 i_2}\,, \\ i_3 &\sim \sum_{i_2} \left(\mathbf{P}^2\right)_{i_0 i_2} P_{i_2 i_3} = \left(\mathbf{P}^3\right)_{i_0 i_3}\,, \end{aligned}$$

Properties of the random walk

- If we start at i_0, where will we end up after t steps?

$$
\begin{aligned}
i_1 &\sim P_{i_0 i_1}\,, \\
i_2 &\sim \sum_{i_1} P_{i_0 i_1} P_{i_1 i_2} \;=\; \left(\mathbf{P}^2\right)_{i_0 i_2}\,, \\
i_3 &\sim \sum_{i_2} \left(\mathbf{P}^2\right)_{i_0 i_2} P_{i_2 i_3} \;=\; \left(\mathbf{P}^3\right)_{i_0 i_3}\,, \\
&\ldots \\
i_t &\sim \left(\mathbf{P}^t\right)_{i_0 i_t}\,.
\end{aligned}
$$

Transition matrix decomposition

- Recall that $P_{ij} = W_{ij} / \sum_j Wij$.
- Let $\mathbf{W}$ be the weight matrix, and $\mathbf{D}$ be the diagonal matrix, $D_{ij} = \sum_j W_{ij}$. We have

$$\mathbf{P} \;=\; \mathbf{D}^{-1}\mathbf{W}$$

- We will focus on a symmetric matrix

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$$

It can be decomposed using its eigenvectors $\mathbf{z}_1, \ldots, \mathbf{z}_N$ corresponding to eigenvalues $|\lambda_1| \geq \ldots \geq |\lambda_N|$

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} \;=\; \lambda_1 \mathbf{z}_1 \mathbf{z}_1^T + \ldots + \lambda_N \mathbf{z}_N \mathbf{z}_N^T$$

Eigendecomposition

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} \;=\; \lambda_1 \mathbf{z}_1 \mathbf{z}_1^T + \ldots + \lambda_N \mathbf{z}_N \mathbf{z}_N^T$$

Eigenvector/value: $\mathbf{A}\mathbf{z} = \lambda \mathbf{z}$

- The eigenvectors are orthogonal, i.e., $z_i^T z_j = 0$ for $i \neq j$.
- Assume the graph is connected; the random walk then is *ergodic*–there is non-zero probability of getting from any $\mathbf{x}_i$ to any $\mathbf{x}_j$ (in some number of steps).
- Spectral graph theory: the largest eigenvalue is always $\lambda_1 = 1$, and $|\lambda_n| < 1$ for $n = 2, \ldots, N$.

Random walk distribution

$$(\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})^t = (\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})\cdots(\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}) = \mathbf{D}^{\frac{1}{2}}\mathbf{P}^t\mathbf{D}^{-\frac{1}{2}}$$

- Thus,

$$\begin{aligned}\mathbf{P}^t &= \mathbf{D}^{-\frac{1}{2}}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\right)^t\mathbf{D}^{\frac{1}{2}} \\ &= \mathbf{D}^{-\frac{1}{2}}\left(\lambda_1\mathbf{z}_1\mathbf{z}_1^T + \ldots + \lambda_N\mathbf{z}_N\mathbf{z}_N^T\right)^t\mathbf{D}^{\frac{1}{2}} \\ &= \mathbf{D}^{-\frac{1}{2}}\left(\lambda_1^t\mathbf{z}_1\mathbf{z}_1^T + \ldots + \lambda_N^t\mathbf{z}_N\mathbf{z}_N^T\right)\mathbf{D}^{\frac{1}{2}}\end{aligned}$$

- Since $\lambda_1 = 1$, and $|\lambda_i| \leq 1$, when $t \to \infty$ we get

$$\mathbf{P}^\infty = \mathbf{D}^{-\frac{1}{2}}\left(\mathbf{z}_1\mathbf{z}_1^T\right)\mathbf{D}^{\frac{1}{2}}$$

Finite number of steps

$$\mathbf{P}^{\infty} \;=\; \mathbf{D}^{-\frac{1}{2}}\left(\mathbf{z}_1\mathbf{z}_1^T\right)\mathbf{D}^{\frac{1}{2}}$$

- Assuming the graph is ergodic, in the limit the distribution does not depend on the starting point!

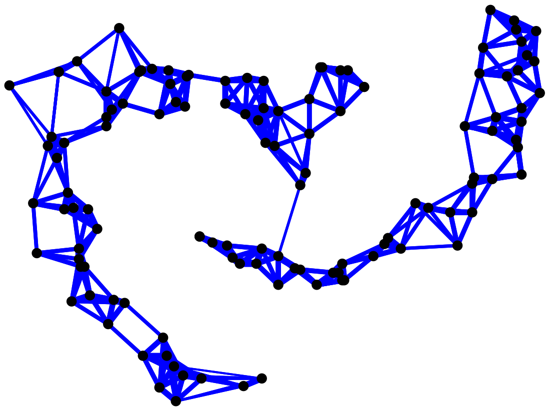
- When t is very large (but finite), we can focus on the *largest* correction:

$$\mathbf{P}^{t} \;\approx\; \mathbf{P}^{\infty} + \mathbf{D}^{-\frac{1}{2}}\left(\lambda_2^2\mathbf{z}_2\mathbf{z}_2^T\right)\mathbf{D}^{\frac{1}{2}}$$

- $(\mathbf{z}_2\mathbf{z}_2^T)_{ij} \;=\; z_{2i}z_{2j}$, so the probability of starting in $\mathbf{x}_i$ and ending in $\mathbf{x}_j$ is a little bit *increased* if $\mathrm{sign}(z_{2i}) = \mathrm{sign}(z_{2j})$, and decreased otherwise.
  $\Rightarrow$ Cluster based on the sign of z_{2i}

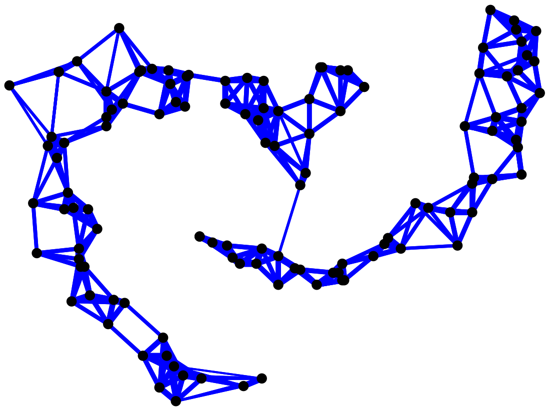
Example

Data & Graph, 5-NN

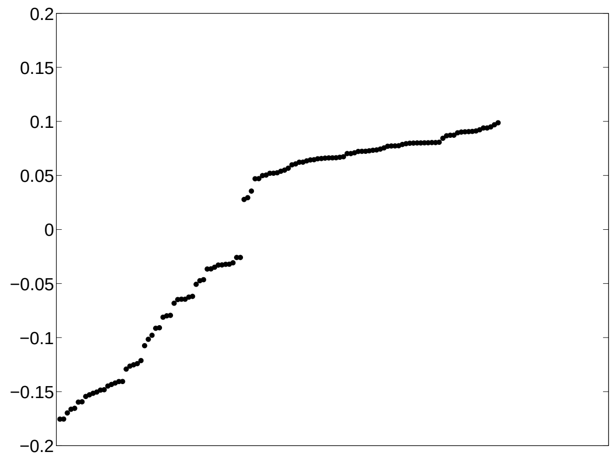


Example

Data & Graph, 5-NN

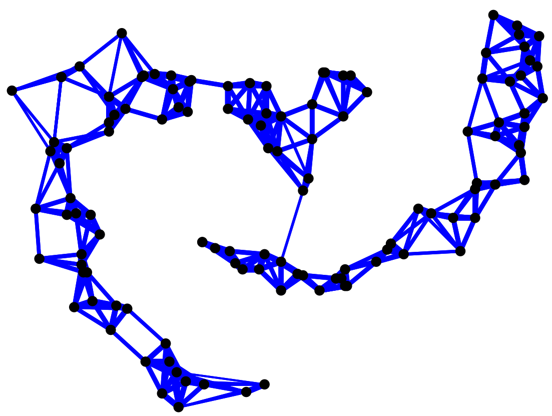


2nd eigenvalue (sorted)

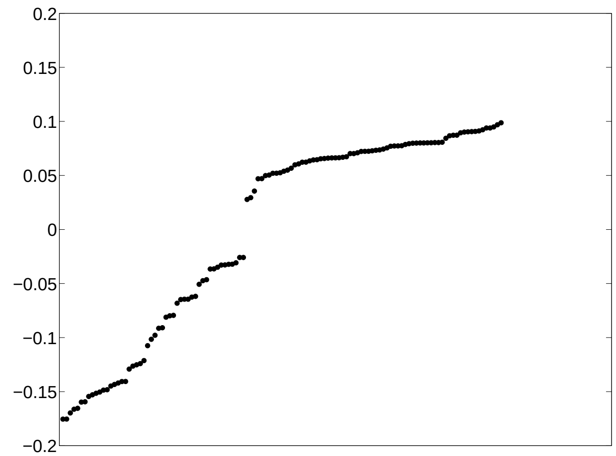


Example

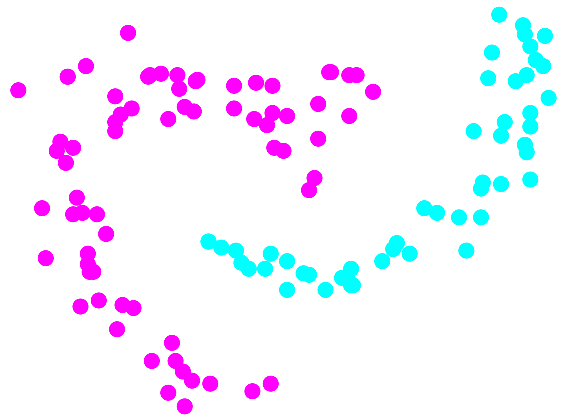
Data & Graph, 5-NN



2nd eigenvalue (sorted)



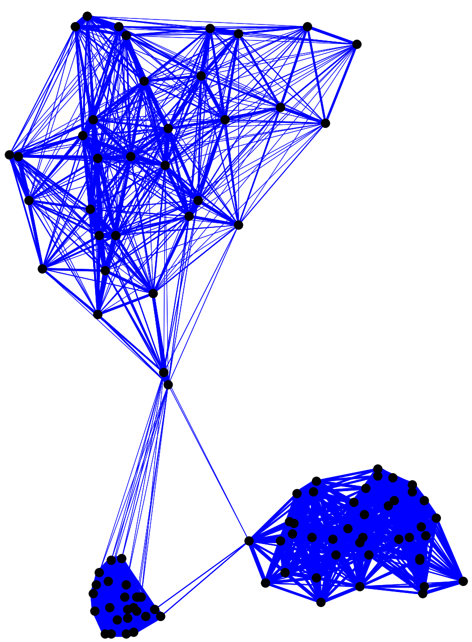
Clustering



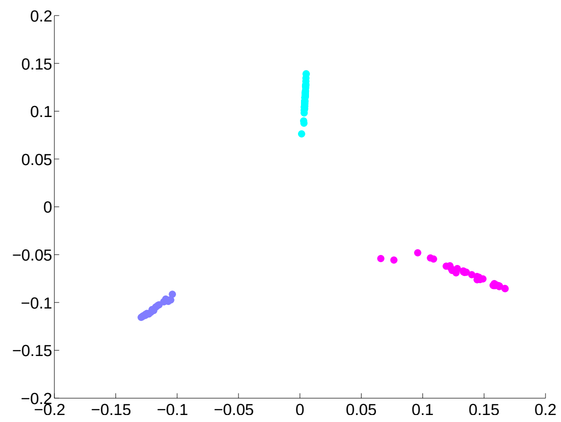
Beyond binary clustering

- When $k > 2$:
  - Let $\mathbf{Z}_i = [z_{1i}, \ldots, z_{ki}]^T$.
  - Apply k-means clustering on $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$.

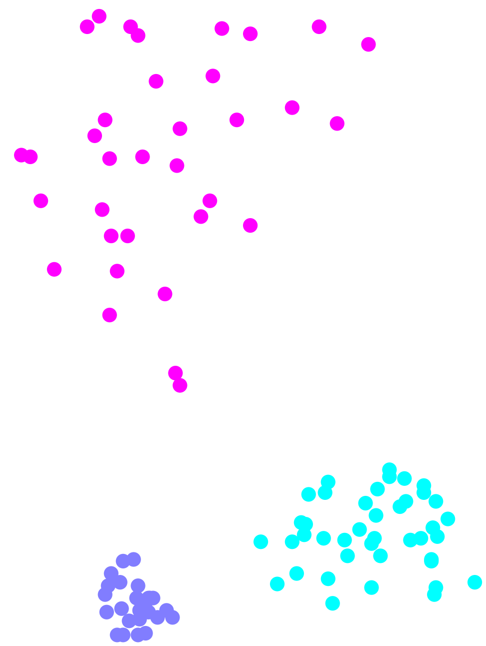
Graph, 20-NN



$\mathbf{Z}$

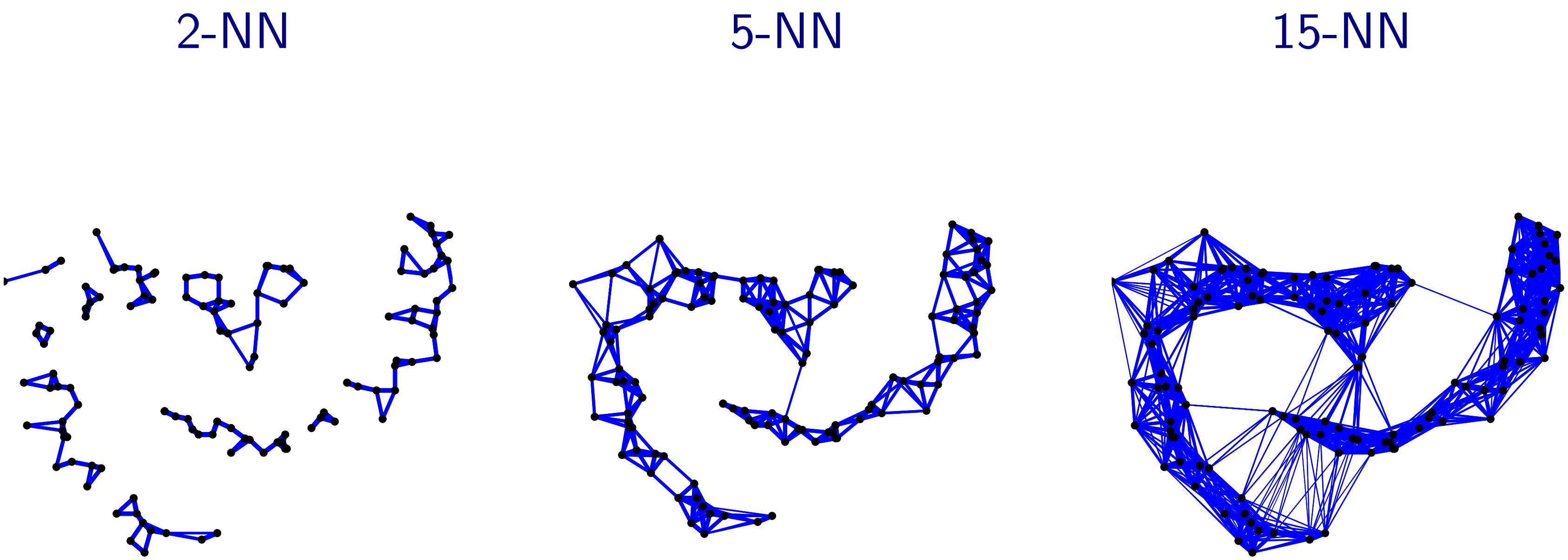


Clustering

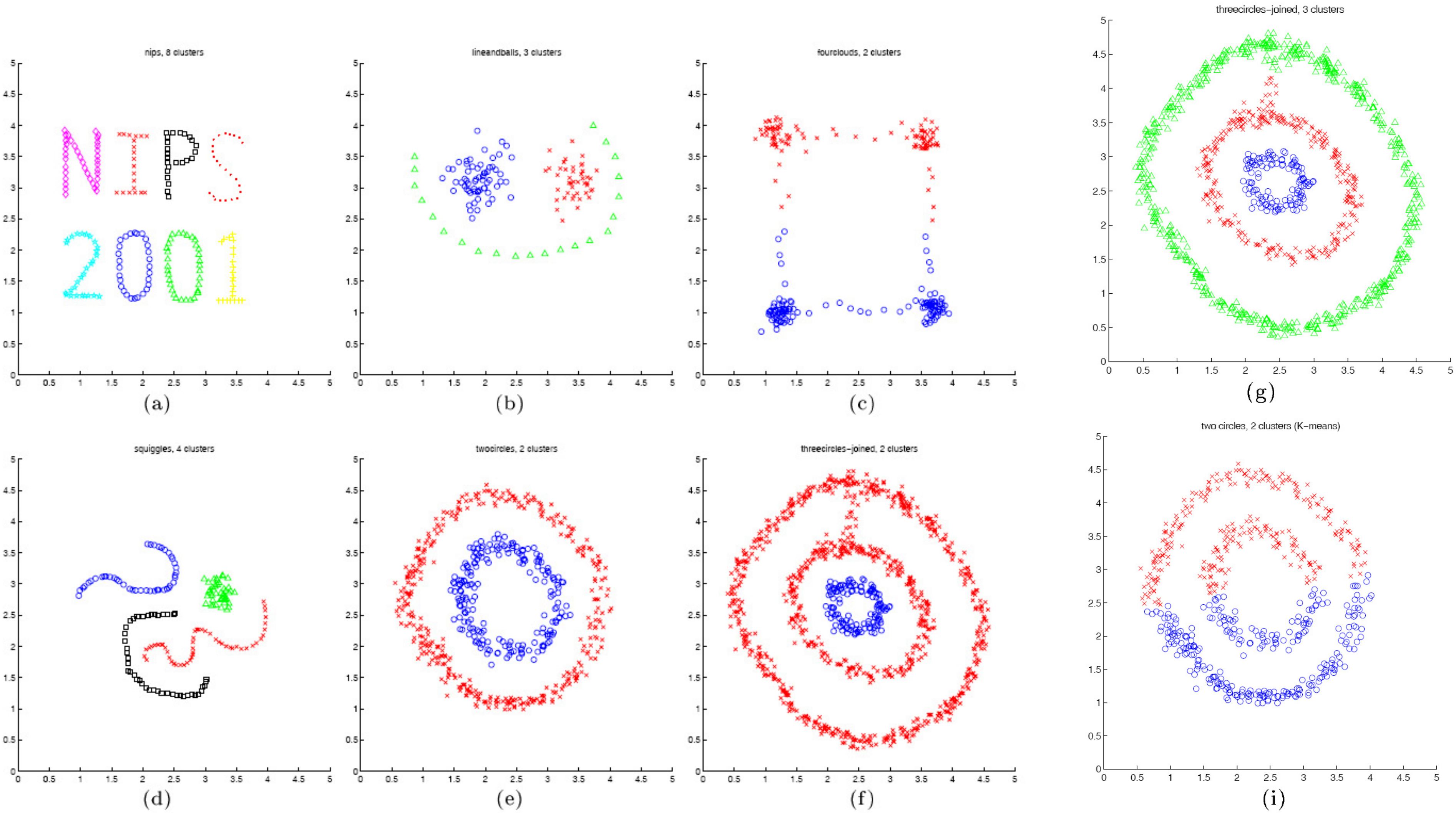


Parameters of spectral clustering

- Two parameters (in addition to k):
  - Neighborhood size (# of nearest neighbors)
  - Distance falloff parameter β.



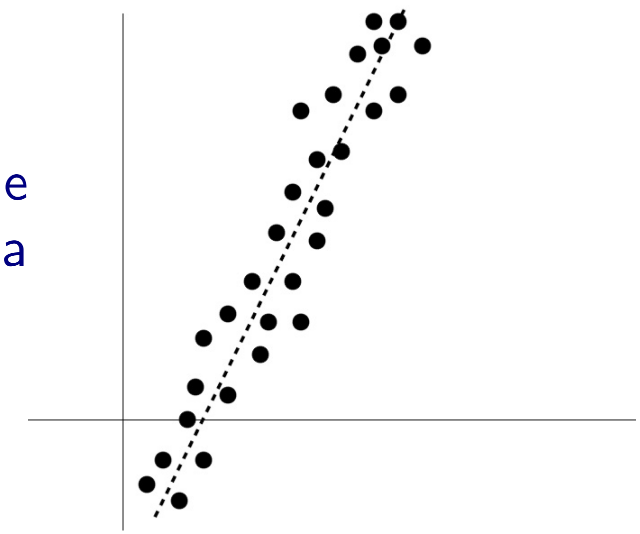
More examples, from [Ng *et al* '01]



Dimensionality reduction

- The dimensionality of observations is dictated by the number/type of sensors, and could be quite arbitrary.

- The *intrinsic* dimensionality is a property of the generating process $\Rightarrow$ assumption that the data lie on (or near) a subspace.



Dimensionality reduction vs. clustering

- Dimensionality reduction and clustering are both about recovering simple structure that "explains" the data.
  - Clustering: discrete explanation (cluster labels)
  - Dimensionality reduction: continuous explanation (underlying subspace).
- In both cases, the structure is represented by hidden variables that need to be recovered.

Criteria

- Recall clustering objective: minimize distortion within clusters.
- Objective in dimensionality reduction: find k-dim. subspace $\mathcal{M}$ in $\mathbb{R}^d$, and define a projection $\mathbf{x} \in \mathbb{R}^d \to \mathbf{x}' \in \mathcal{M}$, such that the *residual* $\|\mathbf{x}' - \mathbf{x}\|$ is minimized.

Next time

PCA;
Feature selection.