Today:    - Hierarchical clustering
          - Spectral clustering

Next time: - Bayes classifiers
Hierarchical Clustering

- Greedily join nearest cluster pair [Eisen 1998]
Hierarchical Clustering

- Greedily join nearest cluster pair [Eisen 1998]
Importance of the Ordering

• Genes that are adjacent in the linear ordering are often hypothesized to share a common function.

• Ordering can help determine relationships between genes and clusters in time series data analysis.
The Problem

“There are $2^{n-1}$ linear orderings consistent with the structure of the tree. …

An optimal linear ordering, one that maximizes the similarity of adjacent elements in the ordering, is impractical to compute.”

[Eisen et al, PNAS 1998]
• For $n$ leaves there are $n-1$ internal nodes

• Each flip in an internal node creates a new linear ordering

• There are $2^{n-1}$ possible linear ordering of the leafs of the tree
Optimal leaf ordering

Denote by $\Phi$ the space of the possible linear orderings consistent with the tree.

Denote by $v_1 \ldots v_n$ the tree leaves.

Our goal is to find an ordering that maximizes the similarity of adjacent elements:

$$\max_{\phi \in \Phi} \sum_{i=1}^{n-1} S(v_i^\phi, v_{i+1}^\phi)$$

where $S$ is the similarity matrix.
Computing the Optimal Similarity

Recursively compute the optimal similarity $L_T(u,w)$ for any pair of leaves $(u,w)$ which could be on different corners (leftmost and rightmost) of $T$.

For a leaf $u \in T$, $C_T(u)$ is the set of all possible corner leaves of $T$ when $u$ is on one corner of $T$.

$$L_T(u,w) = \max_{m \in C_{T_1}(u), k \in C_{T_2}(w)} L_{T_1}(u,m) + L_{T_2}(k,w) + S(m,k)$$
For all \( u \in T_1 \)

For all \( w \in T_2 \)

\[
L_T(u, w) = \max_{m \in C_{T_1}(u), k \in C_{T_2}(w)} L_{T_1}(u, m) + L_{T_2}(k, w) + S(m, k)
\]

For all \( u \in T_1 \)

For all \( k \in T_2 \)

\[
LL(u, k) = \max_{m \in C_{T_1}(u)} L_{T_1}(u, m) + S(m, k)
\]

For all \( w \in T_2 \)

\[
L_T(u, w) = \max_{k \in C_{T_2}(w)} LL(u, k) + L_{T_2}(w, k)
\]
Results – Synthetic Data

Hierarchical clustering  Input  Optimal ordering

Hierarchical clustering  Input  Optimal ordering
Biological Results

• Spellman identified 800 genes as cell cycle regulated in *Saccharomyces cerevisiae*.

• Genes were assigned to five groups termed *G1,S,S/G2,G2/M* and *M/G1* which approximate the commonly used cell cycle groups in the literature.

• This assignment was performed using a ‘phasing’ method which is a supervised classification algorithm.

• In addition to the phasing method, the authors clustered these genes using hierarchical clustering.
Cell Cycle – 24 experiments of cdc15 temperature sensitive mutant

Hierarchical clustering

Optimal ordering
24 experiments of cdc15 temperature sensitive mutant
Spectral clustering
Spectral clustering - motivation

Main idea – rely on the graph structure to define the clusters
Spectral clustering

• Spectral clustering (as described here) relies on a random walk over the points.

• We find the random walk via the following steps
  1. construct a neighborhood graph
  2. assign weights to the edges in the graph
  3. define a transition probability matrix based on the weights

• Clustering is performed using the eigenvectors of the resulting transition probability matrix
1. Generating the graph

- We can connect each point to its $k$ nearest neighbors, or connect each point to all neighbors within distance $\varepsilon$. 
2. Edge weights

- We assign symmetric nonnegative edge weights $W_{ij}$:

$$
\begin{cases}
    w_{ij} = e^{-\beta \|x_i - x_j\|} & \text{if } i \text{ and } j \text{ are connected} \\
    w_{ij} = 0 & \text{otherwise}
\end{cases}
$$
3. Transition probability matrix

• Finally, we define a Markov random walk over the neighborhood graph by constructing a transition probability matrix from the edge weights

\[ p_{ij} = \frac{w_{ij}}{w_i} \quad \text{where} \quad w_i = \sum_j w_{ij} \]

• The random walk proceeds by successively selecting points according to \( j \sim P_{ij} \), where \( i \) specifies the current location.
Random walk

- If we start from $i_0$, the distribution of points it that we end up in after $t$ steps is given by

$$i_1 \sim p_{i_0 i_1}$$
Random walk

- If we start from $i_0$, the distribution of points it that we end up in after $t$ steps is given by

$$i_1 \sim p_{i_0 i_1}$$

$$i_2 \sim \sum_{i_1} p_{i_0 i_1} p_{i_1 i_2} = [P^2]_{i_0 i_2}$$

- where $P^t = PP \ldots P$ (t matrix products) and $[\cdot]_{ij}$ denotes the $i,j$ component of the matrix.
Random walk

• If we start from $i_0$, the distribution of points it that we end up in after $t$ steps is given by

\[ i_1 \sim p_{i_0i_1} \]
\[ i_2 \sim \sum_{i_1} p_{i_0i_1} p_{i_1i_2} = [P^2]_{i_0i_2} \]
\[ i_3 \sim \sum_{i_1} \sum_{i_2} p_{i_0i_1} p_{i_1i_2} p_{i_2i_3} = [P^3]_{i_0i_3} \]
\[ \ldots \]
\[ i_t \sim [P^t]_{i_0i_t} \]

• where $P^t = PP \ldots P$ (t matrix products) and $[\cdot]_{ij}$ denotes the $i,j$ component of the matrix.
Properties of the random walk

- The distributions of points we end up in after $t$ steps converge as $t$ increases. If the graph is connected, the resulting distribution is independent of the starting point.

- However, even for large $t$, the transition probabilities $[P^t]_{ij}$ have a slightly higher probability of transitioning within "clusters" than across; we want to recover this effect from eigenvalues/vectors.
Eigenvalue decomposition

• Let \( W \) be the matrix with components \( W_{ij} \) and \( D \) a diagonal matrix such that \( D_{ii} = \sum_j W_{ij} \). Then

\[
P = D^{-1}W
\]

• To find out how \( P^t \) behaves for large \( t \) it is useful to examine the eigen decomposition of the following symmetric matrix

\[
D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = \lambda_1 z_1 z_1^T + \lambda_2 z_2 z_2^T + \cdots + \lambda_n z_n z_n^T
\]

• where \( z_i \) is the \( i \)th eigenvector and the ordering is such that \( |\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| \)
Eigen decomposition (cont.)

- The symmetric matrix is related to $P^t$ since

\[
(D \begin{pmatrix} -1 & 1 \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \cdots (D \begin{pmatrix} -1 & 1 \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix}) = D^2 (P \cdots P) D^2
\]
Eigen decomposition (cont.)

• The symmetric matrix is related to $P^t$ since

$$
\begin{pmatrix}
\frac{1}{\lambda_1} & \frac{1}{\lambda_1} \\
\frac{1}{\lambda_2} & \frac{1}{\lambda_2} \\
\end{pmatrix} \cdots \begin{pmatrix}
\frac{1}{\lambda_1} & \frac{1}{\lambda_1} \\
\frac{1}{\lambda_2} & \frac{1}{\lambda_2} \\
\end{pmatrix} = D^2 (P \cdots P) D^{-2}
$$

• This allows us to write the $t$ step transition probability matrix in terms of the eigenvalues/vectors of the symmetric matrix

$$
P^t = D^{-2} D^2 (P \cdots P) D^{-2} D^2
$$

$$
= D^{-2} \left( D^{-2} W D^{-2} \right)^t D^2
$$
Eigen decomposition (cont.)

• The symmetric matrix is related to $P^t$ since

\[
(D^{-\frac{1}{2}} W D^{-\frac{1}{2}}) \cdots (D^{-\frac{1}{2}} W D^{-\frac{1}{2}}) = D^2 (P \cdots P) D^{-\frac{1}{2}}
\]

• This allows us to write the $t$ step transition probability matrix in terms of the eigenvalues/vectors of the symmetric matrix

\[
P^t = D^{-\frac{1}{2}} D^2 \left( P \cdots P \right) D^{-\frac{1}{2}} D^2 = D^{-\frac{1}{2}} \left( D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \right)^t D^2 = D^{-\frac{1}{2}} \left( \lambda_1^t z_1 z_1^T + \lambda_2^t z_2 z_2^T \cdots + \lambda_n^t z_n z_n^T \right) D^2
\]
Expressing $P^t$

$$P^t = D^{-rac{1}{2}} \left( \lambda_1^t z_1 z_1^T + \lambda_2^t z_2 z_2^T + \cdots + \lambda_n^t z_n z_n^T \right) D^2$$

Where $\lambda_1 = 1$ and for all other $i$, $\lambda_i < 1$

• Thus:

$$P^\infty = $$
Expressing $P^t$

$$P^t = D^{-\frac{1}{2}} (\lambda_1^t z_1 z_1^T + \lambda_2^t z_2 z_2^T \cdots + \lambda_n^t z_n z_n^T )D^{\frac{1}{2}}$$

Where $\lambda_1 = 1$ and for all other $i$, $\lambda_i < 1$

• Thus:

$$P^\infty = D^{-\frac{1}{2}} (z_1 z_1^T )D^{\frac{1}{2}}$$
Clustering

• We are interested in the largest correction to the asymptotic limit

\[ P^t = P^\infty + D \frac{1}{2} (\lambda z_2 z_2^T) \frac{1}{D^2} \]

• The largest correction term should *increase* the probability of transitions between points in the same cluster and *decrease* the transition probability between points in different clusters

• Thus, points in the same cluster will share the same sign of \( z_2 \) and points in different clusters will differ in their sign
Binary clustering

- We divide the points into two clusters based on the sign of the elements of $z_2$:

\[ z_{2j} > 0 \implies \text{cluster 1, otherwise cluster 0} \]
The sign of the second eigenvector

The entries in the eigenvector corresponding to the second largest eigenvalue
Acknowledgment

These slides are based in part on slides from previous machine learning classes taught by Andrew Moore at CMU and Tommi Jaakkola at MIT. I thank Andrew and Tommi for letting me use their slides.