Clustering (Gene Expression Data)

Plan
- Gene Expression Data/DNA Microarrays
- Feature selection and Clustering

DNA MicroArrays
- To measure levels of messages in a cell
  - Construct an array with DNA sequences for multiple genes
  - Hybridize each RNA in your sample to a sequence in your array (All sequences from the same gene hybridize to the same spot)
  - Measure the number of hybridizations for each spot

Result
- 6000 genes in one shot
- Entire transcriptome observable in one experiment
- Can perform multiple experiments under varying conditions
  - Time
  - Temperature
  - Sugar level
  - Other chemicals
  - Gene knock-outs
  - ...

Noise
- Sources of Noise
  - Cross-hybridization
  - Non-uniform hybridization kinetics
  - Non-linearity of array response to concentration
  - Non-linear amplification
  - Improper probe sequence
  - Difference in materials/procedures
**Expression Value Normalization**

- Noise model: \( y_{ij} = n_i \alpha_{ij} (c_j t_{ij}) + \epsilon_{ij} \)
  - \( y_{ij} \): observed level for gene \( j \) on chip \( i \)
  - \( t_{ij} \): true level
  - \( c_j \): gene constant
  - \( n_i \): multiplicative chip normalization
  - \( \alpha_{ij}, \epsilon_{ij} \): multiplicative and additive noise terms

**Gene expression data**

- For each gene \( j \) we have a vector \( t_j = (t_{ij}, t_{i2}, \ldots, t_{id}) \)
- Now what?
- I.e., what can we do with this data?

**The goal**

- Group genes into co-regulated sets
  - Observe cells under different environmental changes
  - Find genes whose expression profiles are affected in a similar way
  - These genes are potentially co-regulated, i.e., regulated by the same transcription factor
- Clustering!

**Clustering expression levels**

- Clustering process:
  1. How to tell if two expression profiles are similar?
     - Define the (dis)-similarity measure between two profiles
  2. How to group multiple profiles into meaningful subsets?
     - Describe the clustering procedure
  3. Are the results meaningful?
     - Evaluate statistical significance of a clustering
- And don’t forget about:
  - De-noising
  - Choice of experiments/features

1. **(Dis)-similarity measures**

   - Distance metrics (between vectors \( x \) and \( y \))
     - "Manhattan" distance: \( MD(x,y) = \sum |x_i - y_i| \)
     - Euclidean distance: \( ED(x,y) = [\sum (x_i - y_i)^2]^{1/2} \)
     - SSE: \( SSE(x,y) = \sum (x_i - y_i)^2 \)
   - Correlation:
     \( C(x,y) = \sum x_i y_i \)
     (possibly take absolute value)

   - Data pre-processing: Instead of clustering on direct observation of expression values...
     - … can cluster based on differential expression from the mean, \( \bar{x}, \bar{y} \):
       \( \sum |x_i - \bar{x} - (y_i - \bar{y})| \)
     - … or differential expression normalized by standard deviation, \( \bar{x}, \bar{y} \):
       \( \sum |(x_i - \bar{x})/\text{stddev}(x) - (y_i - \bar{y})/\text{stddev}(y)| \)

2. **Clustering Algorithms**

   - Hierarchical: Merge data successively to construct tree
   - Non-Hierarchical: place k-means to best explain data
Hierarchical clustering

- **Bottom-up algorithm:**
  - Initialization: each point in a separate cluster
  - At each step:
    - Choose the pair of **closest** clusters
    - Merge
  - The exact behavior of the algorithm depends on how we define the distance $CD(X,Y)$ between clusters $X$ and $Y$
  - How would you define $CD(X,Y)$?

Distance between clusters

- $CD(X,Y)=\min_{x\in X, y\in Y}D(x,y)$
  - Single-link method
- $CD(X,Y)=\max_{x\in X, y\in Y}D(x,y)$
  - Complete-link method
- $CD(X,Y)=\text{avg}_{x\in X, y\in Y}D(x,y)$
  - Average-link method
- $CD(X,Y)=D(\text{avg}(X),\text{avg}(Y))$
  - Centroid method

Example I

Example II

K-means algorithm

- Each cluster $X_i$ has a center $c_i$
- Define the clustering cost criterion
  $$\text{COST}(X_1, \ldots, X_k) = \sum_{i=1}^{k} \sum_{x \in X_i} \text{SSE}(x,c_i)$$
- Algorithm tries to find clusters $X_1, \ldots, X_k$ and centers $c_1, \ldots, c_k$ that minimize COST
- K-means algorithm:
  - Initialize centers "somehow"
  - Repeat:
    - Compute best clusters for given centers
      - Attach each point to the closest center
    - Compute best centers for given clusters
      - Choose the centroid of points in a cluster
    - Until the COST is "small"
  - How?

Choosing optimal center

- Consider a cluster $X$ and a center $c$
  (not necessarily a centroid)
- Want to minimize
  $$\sum_{x \in X} \text{SSE}(x,c) = \sum_{x \in X} (x_i - c)^2$$
- Can optimize each $c_i$ separately:
  $$\sum_{x \in X} (x_i - c_i)^2 = \sum_{x \in X} x_i^2 - 2c_i \sum_{x \in X} x_i + |X|c_i^2$$
- Optimum:
  $$c_i = \sum_{x \in X} x_i / |X|$$
3. Evaluating clustering output

Idea I: Computing statistical significance of clusters

\[
P(\text{pos} \geq r) = \sum_{m \geq r} \binom{p}{m} \binom{n-p}{k-m} \binom{N}{k}
\]

- \(N\) experiments, \(p\) labeled +, \((N-p)\) –
- Cluster: \(k\) elements, \(m\) positive
- \(P\)-value of single cluster containing \(k\) elements out of which \(r\) are same

Idea II: Visualizing clustering output

What we have learned

- Gene expression data
  - Microarray technology
  - De-noising
- Two methods for clustering
  - Hierarchical clustering
    - non-parametric, top-down
  - K-means clustering
    - ‘model’-based
- Evaluating clustering results
- Visualizing clustering output
- Next: running time analysis

Links

- http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html
- http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletH.html