Clustering
Topics

- Introduction
- K-means
- Self-organizing maps
How would you cluster these dogs?
What is a Cluster?

- A set of entities that are **alike**; entities from different clusters are not alike

- Compact clusters
  - within-cluster **distance** < between-cluster distance

- Connected clusters
  - within-cluster **connectivity** > between-cluster connectivity

- Ideal cluster: **compact and isolated**
Representation

Objects: pixels, images, time series, documents

Representation: features, similarity

Gene Expressions

Shamir et al. BMC Bioinformatics, 2005
Purpose of Grouping

Two different meaningful groupings of 16 animals based on 13 Boolean features (appearance & activity)

Mammals

Vs.

Birds

Large weight on appearance features

http://www.ofai.at/~elias.pampalk/kdd03/animals/

Predators

Vs.

Non-Predators

Large weight on activity features
Number of Clusters

The quality of a clustering solution is in the eye of the beholder
Cluster Validity

- Clustering algorithms find clusters, even if there are no natural clusters in the data!

100 2D uniform data points

K-Means with K=3

- Cluster stability (Lange et. al, 2004)
How Gene Expression Data Looks

Entries of the Raw Data matrix:

- Ratios
- Absolute values (RPKM)
- Distributions...

- Row = gene’s expression pattern / fingerprint vector
- Column = experiment/condition’s profile

Normalization is important!!
Gene expression: Applications

- Deduce function of unknown genes: similar expression pattern $\Rightarrow$ similar function “guilt by association”
- Decipher regulatory mechanisms co-expression $\Rightarrow$ co-regulation
- Classify biological conditions
- Identify disease profiles
- ...

Often, the very first step in analysis: **clustering** of genes/conditions.
Clustering: Objective

Group elements (genes) into clusters satisfying:

- **Homogeneity**: Elements inside a cluster are highly similar to each other.

- **Separation**: Elements from different clusters have low similarity to each other.

- Needs formal objective functions
- Most useful versions are NP-hard.
Clustering is NP-hard

- Input: set $X$, $d(i,j) \in \mathbb{Z}^+_0$ $i,j \in X$. $k$, $B \in \mathbb{Z}^+$
- Q: $\exists$ partition of $X$ into $X_1, \ldots, X_k$ s.t. $\forall i, \forall a,b \in X_i$, $d(a,b) \leq B$?
- Reduce from graph 3-coloring (Brucker, 78): for graph $G=(V,E)$ set $X=V$, $d(i,j)=1$ iff $ij \in E$ or $d(i,j)=0$. $k=3$, $B=0$.
- Hardness holds also when $d=\{0,1\}$, $k=3$, and with max or sum objective functions.
- Note: poly for $k=2$ (ex)
K-means Clustering
k-clustering problem

Input: a set of \( n \) elements \( V = \{v_1, \ldots, v_n\} \), integer \( k \)

Each partition \( P \) of \( V \) into \( k \) subsets has a cost \( E^P \)

Goal: find a \( k \)-partition of min cost
The key idea

• The number of clusters $k$ is given
• Repeatedly find the centroid of each cluster and then re-partition the input points according to which of these centroids is closest
K-means clustering
Lloyd 57, MacQueen 67

• Initialize an arbitrary partition $P$ of the elements into $k$ clusters.
• For cluster $j$, element $i \notin j$, $E^p(i,j) = \text{cost of soln. if } i \text{ is moved to cluster } j$. Pick $E^p(r,s)$ that is minimum; move $s$ to cluster $r$ if improving
• Repeat until no improvement possible
• Requires knowledge of $k$
Animation applet

http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html
Identification of Breast Cancer Subtypes using SOM

Dvir Netanely, Ayelet Avraham, Adit Ben-Baruch, Ella Evron

Breast Cancer Research 2016
Breast cancer

• The most common cancer among women
• Highly heterogeneous disease: distinct subtypes require distinct therapies
• Classical therapeutic biomarkers:
  • Estrogen receptor (ER)
  • Progesterone receptor (PR)
  • Epidermal growth factor receptor 2 (HER2/ERBB2)
• Luminal A subclass: largest, treated non-aggressively
• Goal: reanalyze TCGA large cohort of breast cancer omics profiles
534 Luminal-A RNA-seq samples

K-means Clustering reveals two distinct subgroups
LumA-R2 samples exhibit significantly reduced five-year recurrence rate
LumA-R2 over-expressed genes in the T Cell receptor signaling pathway
K-means Clustering of the Luminal-A Samples based on their methylation profiles

- Three groups, one with significantly poorer survival and hyper-methylation of developmental genes

Conclusion: Analysis reveals a subgroup within the Lum-A patients with poor prognosis: may benefit from more aggressive treatment
Multivariate Cox analysis of Luminal-A subgroups for five-year survival and five-year recurrence

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survival</th>
<th></th>
<th>Recurrence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>p-value</td>
<td>HR</td>
<td>p-value</td>
</tr>
<tr>
<td>LumA-R (1 vs 2)</td>
<td>0.56</td>
<td>0.36991</td>
<td>0.06</td>
<td>0.00693</td>
</tr>
<tr>
<td>LumA-M (2,3 vs 1)</td>
<td>6.68</td>
<td>0.00484</td>
<td>3.04</td>
<td>0.07028</td>
</tr>
<tr>
<td>Age (&lt;60 vs.&gt;=60 years)</td>
<td>11.20</td>
<td>0.0037</td>
<td>1.03</td>
<td>0.96530</td>
</tr>
<tr>
<td>Pathologic stage (I,II vs. III,IV)</td>
<td>2.12</td>
<td>0.25519</td>
<td>1.93</td>
<td>0.26992</td>
</tr>
<tr>
<td>ER Status</td>
<td>0.47</td>
<td>0.50039</td>
<td>0.29</td>
<td>0.29092</td>
</tr>
<tr>
<td>PR Status</td>
<td>1.48</td>
<td>0.72659</td>
<td>0.64</td>
<td>0.68789</td>
</tr>
<tr>
<td>Her2 Status</td>
<td>0.56</td>
<td>0.36991</td>
<td>0.06</td>
<td>0.00693</td>
</tr>
</tbody>
</table>

Cox multivariate analysis showed the independent prognostic contribution of each pattern to outcome
Geometric k-clustering

Input: a set of $n$ points $V = \{v_1, \ldots, v_n\}$ $v_i \in \mathbb{R}^m$, integer $k$

distance $d(x, y)$ = distance between points $x$ and $y$ (e.g. Euclidean)

Given a set $X = \{x_1, \ldots, x_k\}$ of $k$ points ("centers"), define

distance $d(v, X) = \min_i d(v, x_i)$

distance $d(V, X) = (\sum_i d(v_i, X)^2) / n$ mean squared error

Goal: find $X$ that minimizes $d(V, X)$

$X$ implies a partition of $V$ into $k$ subsets

For a cluster $C$, its centroid, or center of gravity is

c = (\sum_{i \in C} v_i) / |C|
K-clustering variations

- **Input**: vector $v_i$ for each element $i$
- $c_p$: a centroid for cluster $p$
- **Objective**:
  - $\sum_{\text{clusters } p} \sum_{i \text{ in cluster } p} d(v_i, c_p)^2$ \textbf{k-means problem}
    - NP-hard even for $k=2$ (Drineas et al. ML 04)
  - $\sum_{\text{clusters } p} \sum_{i \text{ in cluster } p} d(v_i, c_p)$ \textbf{k-median problem}
    - NP-hard on graphs (Kariv, Hakimi 76)
  - $\max_{\text{clusters } p} \max_{i \text{ in cluster } p} d(v_i, c_p)$ \textbf{k-center problem}
    - NP-hard on graphs (Kariv, Hakimi 76)

- **K-medioids alg**: use data points as centers, measure distances using the Manhattan distance.
comments

- **Parallel version**: move each elt. to the cluster with the closest centroid simultaneously
- **Sequential version**: one elt. each time
- “moving centers” approach
- Objective = homogeneity only (k fixed)
- Variations for changing k
Stuart P. Lloyd

- **Stuart P. Lloyd**, SB’43, a physicist, died October 20 2007 in Rahway, NJ. He was 84. *A member of the Manhattan Project, Lloyd was a fellow at the Institute for Advanced Studies before joining Bell Telephone Laboratories’ math department. His research, in particular work now known as Lloyd’s Theorem or Lloyd’s Algorithm, helped improve communication with space probes, increase credit-card security, and advance computer graphics.*

  http://magazine.uchicago.edu/0812/peer_review/deaths.shtml

- Lloyd’s work was not published outside Bell Labs till 1982.

- James MacQueen of UCLA coined the term k-means in 1967.
Self Organizing Maps

Kohonen 97
The ingredients

• Fixed k
• Moving centers approach
• More structure: some topology on the centers (in a different space)
• Center movement depends not only on the points in their clusters - also on others (with weaker effect to points belonging to farther centers, according to the topology.)
Nudge centers towards points

- Point \( p \), center \( f \)
- Move \( f \) a fraction \( a \) of the distance towards \( p \)
- Move also close enough centers
- Smaller steps later

\[
\text{new } f = f + a(p-f)
\]
Self-Organizing Maps
Kohonen 97

• Data: n-dim vector for each element (data point) \( p \)
• Fix a 2-D grid of \( k=lxm \) nodes; \( d(u,v) = \) dist in the grid
• Start with \( k \) arbitrary n-dim “centers” \( f_0(v) \), one corresponding to each node \( v \)
• Iteration \( i \):
  - Pick random data pt. \( p \),
  - Find center \( f_i(v) \) closest to \( p \)
  - Update all centers \( r \):
    * \( f_{i+1}(r) \leftarrow f_i(r) + H(v,r,i)[p-f_i(v)] \)
    * \( H \) : learning function. decreases with iteration no., and with \( d(v,r) \)
Genes – data points

Clusters – map nodes
SOM - Scheme

- Randomly choose a data point (gene).
- Find its closest map node.
- Move this map node towards the data point.
- Move the neighbor map nodes towards this point, but to lesser extent.
- Iterate over data points.
• The extent of node displacements is reduced with the iteration number.

• After thousands of iterations:
  • Assign each gene to the closest map node (cluster)
Teuvo Kohonen
GENECLUSTER

SOM software version for GE, Tamayo et al 99

• Pick random data pt. \( p \),
• Find node \( n \) with center \( f_r(n) \) closest to \( p \)
• Update all centers:
  - \( f_{i+1}(n) \leftarrow f_i(n) + H(n,r,i)[p-f_i(n)] \)
  - \( (H \) decreases with iter, and with \( d(n,r) \) in the grid)\)
• \( H(n,r,i)=0.02T/(T+100i) \) if \( d(n,r)\leq\rho(i), H=0 \) o/w
• \( T = \text{max no. of iterations} \)
• \( \rho(i) = \text{"radius of influence"; linearly decreasing with } i, \rho(0)=3, \rho(T)=0 \)
Yeast cell cycle data: 828 genes, 17 conditions over 2 cell cycles; 6X5 SOM
Eric Lander (& Craig Venter)
SOM clustering of the ENCODE data

• “Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps” Mortazavi et al. Genome Res 2013
Encode project goal: delineate all functional elements encoded in the human genome

https://www.encodeproject.org/

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)
ENCODE (Encyclopedia of DNA elements)

• Project goal: delineate all functional elements encoded in the human genome

• Elements mapped:
  - RNA transcribed regions
  - Protein coding regions
  - Transcription factor binding sites
  - Chromatin structure
  - DNA methylation sites

• 147 cell types, 26 assays, 1640 data sets

• 6 years, 32 groups, >440 scientists, 30 papers, $185M
Grand plan

• **Goal**: Cluster genome segments based on their assay characteristics

• **Method**: Cluster the data using SOM
Forming the input matrix

Generate signal density vector for each segment across the genome

<table>
<thead>
<tr>
<th>segment</th>
<th>Data1</th>
<th>....</th>
<th>DataN</th>
</tr>
</thead>
<tbody>
<tr>
<td>chrA:a-b</td>
<td>rpk_m12</td>
<td>....</td>
<td>rpk_m1n</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

1.5 million segments times 72 datasets
Method

- Break the genome into segments based on all ENCODE data (using ChromHMM)
- Result: 1.5M segments
- Data: 72-long vector of features for each segments, 12 features (histone modifications and DNAse-seq) x 6 cell lines (RPKM)
- Apply **30x45 SOM** (1350 units)
Details

- Toroidal map (why?)
- Hexagonal cells – 6 direct neighbors
  - $T=5M$ iterations
  - Start with update bubble radius $\rho=15$, learning rate $\alpha=0.2$
  - Exponential decrease for $\rho, \alpha$
- Ran SOM 10 times, chose the best based on mean center-point Euclidean distance
1350 states on a torus

Is this clustering meaningful? Does it give new biological insights?
Plot distribution of other genomic features on each cluster

Cell type specific patterns
GO enrichment in clusters

Genes within 20 kb of a genomic segment in a SOM unit are assigned to that unit.

Two neighbor clusters with distinct patterns

Many more results in the refs!