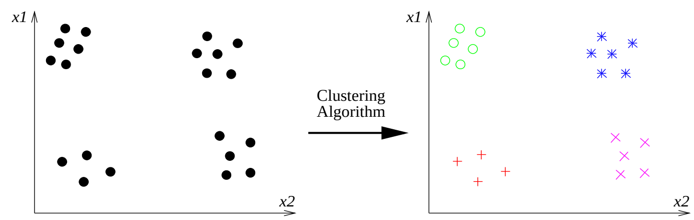
CSCE 970 Lecture 7:
Clustering: Basic Concepts

Stephen D. Scott

March 23, 2001

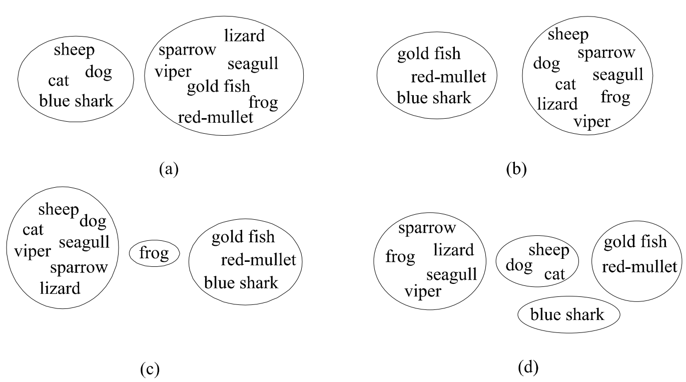
Introduction

- What if labels unavailable?
- E.g. feat. vectors are measurements of electromag. energy reflected from remote parts of Earth, can't afford to visit each area to determine labels
- Clustering (a.k.a. unsupervised PR) algs. group similar f.v.'s together based on a similarity measure
- If clustering is good, then can find label for one of each group & use it as label for entire group



Introduction (cont'd)

- Goal: Place patterns into "sensible" clusters (groups) that reveal similarities and differences, allowing for "useful" conclusions to be derived
- Definitions of "sensible" and "useful" depend on application and the humans involved:

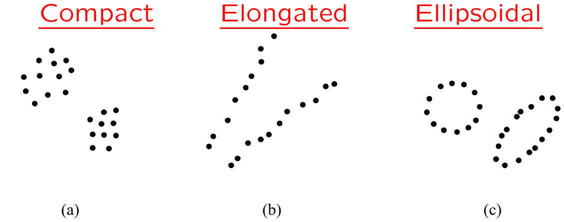


(a) How they bear young (b) Existence of lungs
(c) Environment (d) Both (a) & (b)
(e) [not shown] Vertebrates (all same cluster)

Clustering
Steps

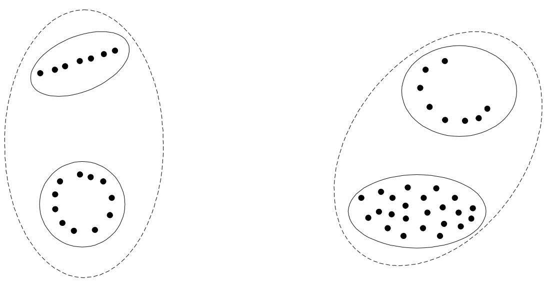
- Feature selection: Requirements and procedures same as in Chapter 5
- Proximity measure: Measures of "similarity" and "dissimilarity" between f.v.'s, between f.v. & a set, or between two sets
  - Preprocessing important to ensure all feats. treated equally
- Clustering criterion: Depends on defn of "sensible":

Compact Elongated Ellipsoidal



Clustering
Steps (cont'd)

- Verify clustering tendency (Sec. 16.6)
- Clustering algorithm: Chapters 12–15
- Cluster validation: Verify that choices of alg. params. & cluster shape match data's clustering structure (Chapt. 16)
- Interpretation: The expert interprets results with other information
- Warning: Each step is subjective and depends on expert's biases!



Clustering
Applications

- Data reduction (compression): Represent each cluster with single item
- Suggest hypotheses about nature of data
- Test hypotheses about data, e.g. that certain feats. are correlated while others are independent
- Prediction based on groups: e.g. Slide 7.2

Clustering
Types of Features

- Nominal: Name only, no quantitative comparisons possible, e.g. {male, female}
- Ordinal: Can be meaningfully ordered, but no quantitative meaning on the differences, e.g. {4, 3, 2, 1} to represent {excellent, very good, good, poor}
- Interval-scaled: Difference is meaningful, ratio is not, e.g. temperature measures on Celsius scale
- Ratio-scaled: Difference and ratio both meaningful, e.g. weight
- Each type possesses the properties of the preceding types

Clustering
Cluster Types

- Start with $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and place into m clusters $C_1, \ldots, C_m$
- Type 1: Hard (crisp)

$$C_i \neq \emptyset,\ i = 1, \ldots, m \qquad \bigcup_{i=1}^{m} C_i = X$$
$$C_i \cap C_j = \emptyset,\ i \neq j,\ i, j \in \{1, \ldots, m\}$$

  - F.v.'s in C_i "more similar" to others in C_i than those in C_j, $j \neq i$

- Type 2: Fuzzy: C_j has membership function $\mu_j : X \to [0, 1]$ s.t.

$$\sum_{j=1}^{m} \mu_j(\mathbf{x}_i) = 1,\ i \in \{1, \ldots, N\}$$
$$0 < \sum_{i=1}^{N} \mu_j(\mathbf{x}_i) < N,\ j \in \{1, \ldots, m\}$$

Proximity Measures
Definitions

- Dissimilarity measure is func. $d : X \times X \to \Re$ s.t.

$$\exists d_0 \in \Re : -\infty < d_0 \le d(\mathbf{x},\mathbf{y}) < +\infty, \ \ \forall \mathbf{x},\mathbf{y} \in X$$
$$d(\mathbf{x},\mathbf{x}) = d_0 \ \ \forall \mathbf{x} \in X$$
$$d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}) \ \ \forall \mathbf{x},\mathbf{y} \in X$$

- d is a metric DM if $d(\mathbf{x},\mathbf{y}) = d_0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ and $d(\mathbf{x},\mathbf{z}) \le d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) \ \ \forall \mathbf{x},\mathbf{y},\mathbf{z} \in X$
  - E.g. $d_2(\cdot,\cdot)$ = Euclidean distance, $d_0 = 0$

- Similarity measure is func. $s : X \times X \to \Re$ s.t.

$$\exists s_0 \in \Re : -\infty < s(\mathbf{x},\mathbf{y}) \le s_0 < +\infty, \ \ \forall \mathbf{x},\mathbf{y} \in X$$
$$s(\mathbf{x},\mathbf{x}) = s_0 \ \ \forall \mathbf{x} \in X$$
$$s(\mathbf{x},\mathbf{y}) = s(\mathbf{y},\mathbf{x}) \ \ \forall \mathbf{x},\mathbf{y} \in X$$

- s is a metric SM if $s(\mathbf{x},\mathbf{y}) = s_0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ and $s(\mathbf{x},\mathbf{y})\, s(\mathbf{y},\mathbf{z}) \le [s(\mathbf{x},\mathbf{y}) + s(\mathbf{y},\mathbf{z})]\, s(\mathbf{x},\mathbf{z})$ $\forall \mathbf{x},\mathbf{y},\mathbf{z} \in X$

Proximity Measures
Definitions (cont'd)

- Can also define proximity measures between sets of f.v.'s

- Let $U = \{D_1, \ldots, D_k\}$, $D_i \subset X$, PM $\alpha : U \times U \to \Re$

- E.g. $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, $U = \{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_4\}, \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}, \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}\}$,

$$d^{ss}_{min}\left(D_i, D_j\right) = \min_{\mathbf{x} \in D_i, \mathbf{y} \in D_j} d_2(\mathbf{x},\mathbf{y})$$

- Min. value is $d^{ss}_{min,0} = 0$, $d^{ss}_{min}(D_i, D_i) = d^{ss}_{min,0}$, and $d^{ss}_{min}\left(D_i, D_j\right) = d^{ss}_{min}\left(D_j, D_i\right)$, so $d^{ss}_{min}(\cdot,\cdot)$ is a DM

- However, $d^{ss}_{min}(\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_4\}) = d^{ss}_{min,0}$ and $\{\mathbf{x}_1, \mathbf{x}_2\} \ne \{\mathbf{x}_1, \mathbf{x}_4\}$, so not a metric DM

Proximity Measures Between Points
Real-Valued Vectors
Example Dissimilarity Measures (pp. 361–362)

- Common, general-purpose metric DM is weighted L_p norm:

$$d_p(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{\ell} w_i \left|x_i - y_i\right|^p\right)^{1/p}$$

- Special cases include weighted Euclidian distance ($p = 2$), weighted Manhattan distance

$$d_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{\ell} w_i \left|x_i - y_i\right|,$$

  and weighted L_∞ norm

$$d_\infty(\mathbf{x},\mathbf{y}) = \max_{1 \le i \le \ell} \{w_i \left|x_i - y_i\right|\}$$

- Generalization of weighted L_2 norm is

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T B\, (\mathbf{x}-\mathbf{y})}\ ,$$

  e.g. Mahalanobis distance

Proximity Measures Between Points
Real-Valued Vectors
Example Similarity Measures (pp. 362–363)

- Inner product:

$$s_{inner}(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{\ell} x_i\, y_i$$

- If $\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 \le a$, then $-a^2 \le s_{inner}(\mathbf{x},\mathbf{y}) \le a^2$

- Tanimoto distance:

$$s_T(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \mathbf{x}^T\mathbf{y}} = \frac{1}{1 + \frac{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}{\mathbf{x}^T\mathbf{y}}},$$

  which is inversely prop. to (squared Euclid. dist.)/(correlation measure)

Proximity Measures Between Points

Discrete-Valued Vectors

- If the coordinates of f.v.'s come from $\{0, \ldots, k-1\}$, can use SMs and DMs defined for real-valued f.v.'s, (e.g. weighted L_p norm) plus:
  - Hamming distance: DM measuring number of places where $\mathbf{x}$ and $\mathbf{y}$ differ
  - Tanimoto measure: SM measuring number of places where $\mathbf{x}$ and $\mathbf{y}$ are same, divided by total number of places
    - Ignore places i where $x_i = y_i = 0$
      - Useful for ordinal features where x_i is degree to which $\mathbf{x}$ possesses ith feature

Proximity Measures Between Points

Fuzzy Measures

- Let $x_i \in [0, 1]$ be measure of how much $\mathbf{x}$ possesses ith feature
- If $x_i, y_i \in \{0, 1\}$, then

$$(x_i \equiv y_i) = ((\neg x_i \wedge \neg y_i) \vee (x_i \wedge y_i))$$

- Generalize to fuzzy values:

$$s(x_i, y_i) = \max\left\{\min\left\{1 - x_i, 1 - y_i\right\}, \min\left\{x_i, y_i\right\}\right\}$$

- To measure similarity between vectors:

$$s_F^p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{\ell} s(x_i, y_i)^p\right)^{1/p}$$

$$\left(\ell^{1/p}\right)/2 \leq s_F^q(\cdot, \cdot) \leq \ell^{1/p}$$

- So $s_F^{\infty} = \max_{1 \leq i \leq \ell} s(x_i, y_i)$ and $s_F^1 = \sum_{i=1}^{\ell} s(x_i, y_i)$ = generalization of Hamming distance

Prox. Measures Between a Point and a Set

- Might want to measure proximity of point $\mathbf{x}$ to existing cluster C
- Can measure proximity α by using all points of C or by using a representative of C
- If all points of C used, common choices:

$$\alpha_{max}^{ps}(\mathbf{x}, C) = \max_{\mathbf{y} \in C}\{\alpha(\mathbf{x}, \mathbf{y})\}$$

$$\alpha_{min}^{ps}(\mathbf{x}, C) = \min_{\mathbf{y} \in C}\{\alpha(\mathbf{x}, \mathbf{y})\}$$

$$\alpha_{avg}^{ps}(\mathbf{x}, C) = \frac{1}{|C|}\sum_{\mathbf{y} \in C} \alpha(\mathbf{x}, \mathbf{y}) ,$$

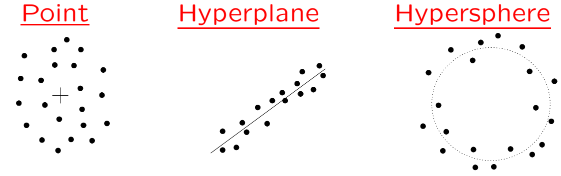
where $\alpha(\mathbf{x}, \mathbf{y})$ is any measure between $\mathbf{x}$ and $\mathbf{y}$

Prox. Measures Between a Point and a Set

Representatives

- Alternative: Measure distance between point $\mathbf{x}$ and a representative of the set C
- Appropriate choice of representative depends on type of cluster

Compact	Elongated	Hyperspherical
Point	Hyperplane	Hypersphere



(a) (b) (c)

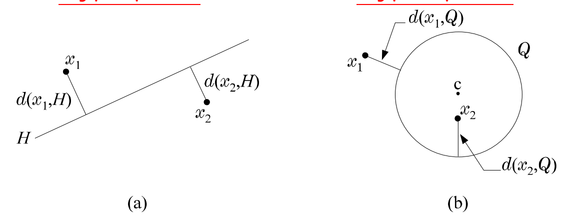
Prox. Measures Between a Point and a Set
Examples of Point Representatives

- Mean vector: $\mathbf{m}_p = \frac{1}{|C|} \sum_{\mathbf{y} \in C} \mathbf{y}$
- Works well in $\Re^\ell$, but might not exist in discrete space
- Mean center $\mathbf{m}_c \in C$:
$$\sum_{\mathbf{y} \in C} d(\mathbf{m}_c, \mathbf{y}) \le \sum_{\mathbf{y} \in C} d(\mathbf{z}, \mathbf{y}) \quad \forall \mathbf{z} \in C,$$
where $d(\cdot, \cdot)$ is DM (if SM used, reverse ineq.)
- Median center: For each point $\mathbf{y} \in C$, find median dissimilarity from $\mathbf{y}$ to all other points of C, then take min; so $\mathbf{m}_{med} \in C$ is defined as
$\text{med}_{\mathbf{y} \in C} \{d(\mathbf{m}_{med}, \mathbf{y})\} \le \text{med}_{\mathbf{y} \in C} \{d(\mathbf{z}, \mathbf{y})\} \quad \forall \mathbf{z} \in C$
- Examples p. 375
- Now can measure proximity between C's rep and $\mathbf{x}$ with standard measures

Prox. Measures Between a Point and a Set
Hyperplane & Hyperspherical Representatives

- Definition of hyperplane H and dist. function:
$$\mathbf{a}^T \mathbf{x} + a_0 = 0 \qquad d(\mathbf{x}, H) = \min_{\mathbf{z} \in H} d(\mathbf{x}, \mathbf{z})$$
- Definition of hypersphere Q and dist. function:
$$(\mathbf{x} - \mathbf{c})^T (\mathbf{x} - \mathbf{c}) = r^2 \qquad d(\mathbf{x}, Q) = \min_{\mathbf{z} \in Q} d(\mathbf{x}, \mathbf{z})$$

Hyperplane Hypersphere



- Given set of points, can find representative via regression techniques, minimizing sum of distances between points and representative

Prox. Measures Between Two Sets

- Given sets of f.v.'s D_i and D_j and prox. meas. $\alpha(\cdot, \cdot)$
- Max: $\alpha^{ss}_{max}(D_i, D_j) = \max_{\mathbf{x} \in D_i, \mathbf{y} \in D_j} \{\alpha(\mathbf{x}, \mathbf{y})\}$ is a measure (but not necessarily a metric) iff α is a SM
  - E.g. α is Euclid. dist. (a DM), $\ell = 1$, $D_1 = \{(1), (10)\}$, $D_2 = \{(4), (7)\}$:
  $\alpha^{ss}_{max}(D_1, D_1) = 9 \ne 3 = \alpha^{ss}_{max}(D_2, D_2)$
  - α is SM $\Rightarrow \alpha(\mathbf{x}, \mathbf{y}) \le s_0 \ \forall \mathbf{x}, \mathbf{y}$ and $\alpha(\mathbf{x}, \mathbf{x}) = s_0 \ \forall \mathbf{x}$, so

$$\alpha^{ss}_{max}(D_i, D_j) \le s_0 \ \forall D_i, D_j, \quad \text{and} \ \forall D$$
$$\alpha^{ss}_{max}(D, D) = \max_{\mathbf{x} \in D, \mathbf{y} \in D} \{\alpha(\mathbf{x}, \mathbf{y})\} = \max_{\mathbf{x} \in D} \{\alpha(\mathbf{x}, \mathbf{x})\} = s_0$$
$$\alpha^{ss}_{max}(D_i, D_j) = \alpha^{ss}_{max}(D_j, D_i)$$

Prox. Measures Between Two Sets
(cont'd)

- Min: $\alpha^{ss}_{min}(D_i, D_j) = \min_{\mathbf{x} \in D_i, \mathbf{y} \in D_j} \{\alpha(\mathbf{x}, \mathbf{y})\}$
is a measure (but not a metric) iff α is a DM
- Average: $\alpha^{ss}_{avg}(D_i, D_j) = \frac{1}{|D_i| \, |D_j|} \sum_{\mathbf{x} \in D_i} \sum_{\mathbf{y} \in D_j} \alpha(\mathbf{x}, \mathbf{y})$
is not necessarily a measure even if α is
- Representative (mean):
$\alpha^{ss}_{rep}(D_i, D_j) = \alpha(\mathbf{m}_{D_i}, \mathbf{m}_{D_j})$,
($\mathbf{m}_{D_i}$ is point rep. of D_i) is a measure whenever α is

Overview of Clustering Algorithms
Exhaustive Search

- Want to find set of clusters that maximizes SM or minimizes DM
- Option 1: Try all possible clusters of size m for various values of m
- Number of ways to partition N items into m nonempty subsets is exactly given by the Stirling numbers of the second kind, which are:

$$\gg \binom{N}{m} \geq \left(\frac{N}{m}\right)^m$$

- Thus brute-force approach infeasible

Overview of Clustering Algorithms
Categories of Algorithms

- Sequential algorithms (Chapt. 12) produce a single clustering, use straightforward greedy approaches, and output depends on the order the f.v.'s are presented to the algorithm
- Hierarchical algorithms (Chapt. 13) produce a sequence (hierarchy) of clusterings, and are of two types:
  - Agglomerative: Repeatedly merge two clusters into one
  - Divisive: Repeatedly divide one cluster into two
- Algorithms based on cost function optimization (Chapt. 14) evaluate the goodness of a clustering with a cost function, typically m is fixed
  - Crisp (hard): Each f.v. belongs to only one cluster, e.g. Isodata algorithm
  - Fuzzy: Each f.v. can belong to a cluster up to a certain degree, as indicated by a membership function
  - Many more
- Other various methods (Chapt. 15)