Artificial Intelligence -- 15-381
Unsupervised Machine Learning Methods

Jaime Carbonell
4-April-2000

OUTLINE:

- What is unsupervised learning?
- Similarity computations
- Clustering algorithms
- Other kinds of unsupervised learning

Unsupervised Learning

Definition of Unsupervised Learning

Learning useful structure ***without*** labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data.

Examples:

- Find natural groupings of X's
  (X = human languages, stocks,
  gene sequences, animal species, ...)
  => Prelude to discovery of underlying properties
- Summarize the news for the past month
  => Cluster first, then report centroids.
- Sequence extrapolation:
  E.g. Predict cancer incidence next decade;
  predict rise in antibiotic-resistant bacteria

Methods:

- Clustering (n-link, k-means, GAC, ...)
- Taxonomy creation (hierarchical clustering)
- Novelty detection ("meaningful" outliers)
- Trend detection (extrapolation from multivariate partial derivatives)

Similarity Measures in Data Analysis

General Assumptions

- Each data item is a tuple (vector)
- Values of tuple are nominal, ordinal or numerical
- Similarity = $(\text{distance})^{-1}$

Pure Numerical Tuples

- $\text{sim}(\boldsymbol{d}_i,\boldsymbol{d}_j) = \Sigma d_{i,k} d_{j,k}$
- $\text{sim}(\boldsymbol{d}_i,\boldsymbol{d}_j) = \cos(\boldsymbol{d}_i,\boldsymbol{d}_j)$
- ...and many more (slide after next)

For Ordinal Values

- E.g. "small", "medium", "large", "X-large"
- Convert to numerical assuming constant Δ
  ...on a normalized [0,1] scale, where:
  max(v) = 1, min(v) = 0, others interpolate
- E.g. "small" = 0, "medium" = 0.33, etc.
- Then, use numerical similarity measures
- Or, use similarity matrix (next slide)

Similarity Measures [Cont.]

For Nominal Values

- E.g. "Boston", "LA", "Pittsburgh", or
  "male", "female", or* "diffuse", "globular", "spiral", "pinwheel"
- Binary rule:
  If $d_{i.k} = d_{j,k}$, then sim = 1, else 0
- Use underlying semantic property:
  E.g. Sim(Boston,LA) = αdist(Boston,LA)$^{-1}$, or
  Sim(Boston,LA) = α|size(Boston) - size(LA)|
- Use similarity matrix

Similarity Matrix

	tiny	little	small	medium	large	huge
tiny	1.0	0.8	0.7	0.5	0.2	0.0
little		1.0	0.9	0.7	0.3	0.1
small			1.0	0.7	0.3	0.2
medium				1.0	0.5	0.3
large					1.0	0.8
huge						1.0

- Diagonal must be 1.0
- Monotonicity property must hold
- Triangle inequality must hold
- Transitive property need *not* hold

Incremental Clustering Methods

Incremental Clustering

Given n data items:
D: D_1, D_2, ..., D_i, ... D_n
And given minimal similarity threshold: S_{min}

Cluster data incrementally as follows:

Procedure Singlelink(***D***)
  Let CLUSTERS = $\{D_1\}$
  For i = 2 to n
    Let $D_c = \underset{j < i}{Arg\max}[Sim(D_i,D_j)]$

    If $D_c > S_{min}$, add D_j to D_c's cluster
      Else Append(CLUSTERS,$\{D_j\}$;; new cluster

Procedure Averagelink(***D***)
  Let CLUSTERS = $\{D_1\}$
  For i = 2 to n
    Let $D_c = \underset{C \text{ in CLUSTERS}}{Arg\max}[Sim(D_i,centroid(C)]$

    If $D_c > S_{min}$, add D_j to cluster C
      Else Append(CLUSTERS,$\{D_j\}$) ;; new cluster

Observations

- Single pass over the data
  => easy to cluster new data incrementally
- Requires arbitrary S_{min} threshold
- $O(N^2)$ time, $O(N)$ space

Creating Taxonomies

Hierarchical clustering

- GAC trace creates binary hierarchy
- Incremental-link => Hierarchical version
  1. Cluster data with high S_{min}
     => 1st hierarchical level
  2. Decrease S_{min} (stop at S_{min}=0)
  3. Treat cluster centroids as data tuples
     and recluster, creating next level of hierarchy,
     then repeat steps 2 and 3.
- k-means => Hierarchical k-means
  1. Cluster data with large k
  2. Decrease k (stop at k=1)
  3. Treat cluster centroids as data tuples
     and recluster, creating next level of hierarchy,
     then repeat steps 2 and 3.

Postprocess Taxonomies

- Eliminate "no-op" levels
- Agglomerate "skinny" levels
- Label meaningful levels
  manually or with centroid summary