C(I)S 330: Applied Database Systems

A Break: A Mini-Introduction to Data Mining (Continued)

(Some slides courtesy of Rich Caruana)

Data Mining Techniques

- Supervised learning
  - Classification and regression
- Unsupervised learning
  - Clustering
- Dependency modeling
  - Associations, summarization, causality
- Outlier and deviation detection
- Trend analysis and change detection

Supervised Learning

- F(x): true function (usually not known)
- D: training sample drawn from F(x)

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	0

Supervised Learning

- F(x): true function (usually not known)
- D: training sample (x,F(x))

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1

- G(x): model learned from D

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0	?

- Goal: $E[(F(x)-G(x))^2]$ is small (near zero) for future samples

Supervised Learning

Well-defined goal:

Learn G(x) that is a good approximation to F(x) from training sample D

Well-defined error metrics:

Accuracy, RMSE, ROC, …

Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0	?

Un-Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0 ?

Un-Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

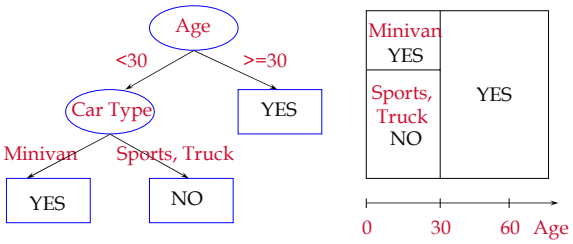
~~71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0~~ ?

Un-Supervised Learning

Data Set:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1

What are Decision Trees?



Decision Trees: Summary

- Many application of decision trees
- There are many algorithms available for:
  - Split selection
  - Pruning
  - Handling Missing Values
  - Data Access
- Decision tree construction still active research area (after 20+ years!)
- Challenges: Performance, scalability, evolving datasets, new applications

Un-Supervised Learning

Data Set:

```
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
18,M,165,0,110,80,41,30,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1
```

Supervised vs. Unsupervised Learning

Supervised

- y=F(x): true function
- D: labeled training set
- D: $\{x_i, F(x_i)\}$
- Learn:
  G(x): model trained to predict labels D
- Goal:
  $E[(F(x)-G(x))^2] \approx 0$
- Well defined criteria:
  Accuracy, RMSE, ...

Unsupervised

- Generator: true model
- D: unlabeled data sample
- D: $\{x_i\}$
- Learn
  ??????????
- Goal:
  ??????????
- Well defined criteria:
  ??????????

What to Learn/Discover?

- Statistical Summaries
- Generators
- Density Estimation
- Patterns/Rules
- Associations (see previous segment)
- Clusters/Groups (this segment)
- Exceptions/Outliers
- Changes in Patterns Over Time or Location

Clustering: Unsupervised Learning

- Given:
  - Data Set D (training set)
  - Similarity/distance metric/information
- Find:
  - Partitioning of data
  - Groups of similar/close items

Similarity?

- Groups of similar customers
  - Similar demographics
  - Similar buying behavior
  - Similar health
- Similar products
  - Similar cost
  - Similar function
  - Similar store
  - …
- Similarity usually is domain/problem specific

Distance Between Records

- *d* dimensional vector space representation and distance metric

r_1: 57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0

r_2: 78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0

…

r_N: 18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Distance (r_1,r_2) = ???

- Pairwise distances between points (no *d* dimensional space)
  - Similarity/dissimilarity matrix (upper or lower diagonal)
  - Distance: 0 = near, ∞ = far
  - Similarity: 0 = far, ∞ = near

--	1	2	3	4	5	6	7	8	9	10
1	-	d	d	d	d	d	d	d	d	d
2		-	d	d	d	d	d	d	d	d
3			-	d	d	d	d	d	d	d
4				-	d	d	d	d	d	d
5					-	d	d	d	d	d
6						-	d	d	d	d
7							-	d	d	d
8								-	d	d
9									-	d

Properties of Distances: Metric Spaces

- A metric space is a set *S* with a global distance function *d*. For every two points *x, y* in *S*, the distance *d(x,y)* is a nonnegative real number.
- A metric space must also satisfy
  - d(x,y) = 0 iff *x = y*
  - d(x,y) = d(y,x) (symmetry)
  - d(x,y) + d(y,z) >= d(x,z) (triangle inequality)

Minkowski Distance (L_p Norm)

- Consider two records $x=(x_1,\ldots,x_d)$, $y=(y_1,\ldots,y_d)$:

$$d(x,y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\ldots+|x_d-y_d|^p}$$

Special cases:

- p=1: Manhattan distance

$$d(x,y)=|x_1-y_1|+|x_2-y_2|+\ldots+|x_p-y_p|$$

- p=2: Euclidean distance

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_d-y_d)^2}$$

Only Binary Variables

2x2 Table:

	0	1	Sum
0	a	b	a+b
1	c	d	c+d
Sum	a+c	b+d	a+b+c+d

- Simple matching coefficient: (symmetric)

$$d(x,y)=\frac{b+c}{a+b+c+d}$$

- Jaccard coefficient: (asymmetric)

$$d(x,y)=\frac{b+c}{b+c+d}$$

Nominal and Ordinal Variables

- Nominal: Count number of matching variables
  - m: # of matches, d: total # of variables

$$d(x,y)=\frac{d-m}{d}$$

- Ordinal: Bucketize and transform to numerical:
  - Consider record x with value x_i for i^{th} attribute of record x; new value x_i':

$$x_i'=\frac{x_i-1}{dom(X_i)-1}$$

Mixtures of Variables

- Weigh each variable differently
- Can take "importance" of variable into account (although usually hard to quantify in practice)

Clustering: Informal Problem Definition

Input:

- A data set of *N* records each given as a *d* dimensional data feature vector.

Output:

- Determine a natural, useful "partitioning" of the data set into a number of (k) clusters and noise such that we have:
  - High similarity of records within each cluster (intra-cluster similarity)
  - Low similarity of records between clusters (inter-cluster similarity)

Types of Clustering

- Hard Clustering:
  - Each object is in one and only one cluster
- Soft Clustering:
  - Each object has a probability of being in each cluster

Clustering Algorithms

- Partitioning based clustering
  - K-means clustering
  - K-medoids clustering
  - EM (expectation maximization) clustering
- Hierarchical clustering
  - Divisive clustering (top down)
  - Agglomerative clustering (bottom up)
- Density Based Methods
  - Regions of dense points separated by sparser regions of relatively low density

K-Means Clustering Algorithm

Initialize k cluster centers

Do

Assignment step: Assign each data point to its closest cluster center

Re-estimation step: Re-compute cluster centers

While (there are still changes in the cluster centers)

Visualization at:

- http://www.delft-cluster.nl/textminer/theory/kmeans/kmeans.html

Issues

Why is K-Means working:

- How does it find the cluster centers?
- Does it find an optimal clustering
- What are good starting points for the algorithm?
- What is the right number of cluster centers?
- How do we know it will terminate?

K-Means: Distortion

- Communication between sender and receiver
- Sender encodes dataset: $x_i \rightarrow \{1,\ldots,k\}$
- Receiver decodes dataset: $j \rightarrow center_j$
- Distortion: $D = \sum_{1}^{N} \left(x_i - center_{encode(x_i)}\right)^2$
- A good clustering has minimal distortion.

Properties of the Minimal Distortion

- Recall: Distortion $D = \sum_{1}^{N} \left(x_i - center_{encode(x_i)}\right)^2$
- Property 1: Each data point x_i is encoded by its nearest cluster center $center_j$. (Why?)
- Property 2: When the algorithm stops, the partial derivative of the Distortion with respect to each center attribute is zero.

Property 2 Followed Through

- Calculating the partial derivative:

$$D = \sum_{1}^{N} \left(x_i - center_{encode(x_i)}\right)^2 = \sum_{j=1}^{k} \sum_{i \in Cluster(center_j)} (x_i - center_j)^2$$

$$\frac{\partial D}{\partial center_j} = \frac{\partial}{\partial center_j} \sum_{i \in Cluster(c_j)} (x_i - center_j)^2 = -2 \sum_{i \in Cluster(c_j)} (x_i - center_j) \overset{!}{=} 0$$

- Thus at the minimum:

$$center_j = \frac{1}{|\{i \in Cluster(center_j)\}|} \sum_{i \in Cluster(center_j)} x_i$$

K-Means Minimal Distortion Property

- Property 1: Each data point x_i is encoded by its nearest cluster center $center_j$
- Property 2: Each center is the centroid of its cluster.

- How do we improve a configuration:
  - Change encoding (encode a point by its nearest cluster center)
  - Change the cluster center (make each center the centroid of its cluster)

K-Means Minimal Distortion Property (Contd.)

- Termination? Count the number of distinct configurations …
- Optimality? We might get stuck in a local optimum.
  - Try different starting configurations.
  - Choose the starting centers smart.
- Choosing the number of centers?
  - Hard problem. Usually choose number of clusters that minimizes some criterion.

K-Means: Summary

- Advantages:
  - Good for exploratory data analysis
  - Works well for low dimensional data
  - Reasonably scalable
- Disadvantages
  - Hard to choose k
  - Often clusters are non spherical

K-Medoids

- Similar to K-Means, but for categorical data or data in a non-vector space.
- Since we cannot compute the cluster center (think text data), we take the "most representative" data point in the cluster.
- This data point is called the medoid (the object that "lies in the center").

Agglomerative Clustering

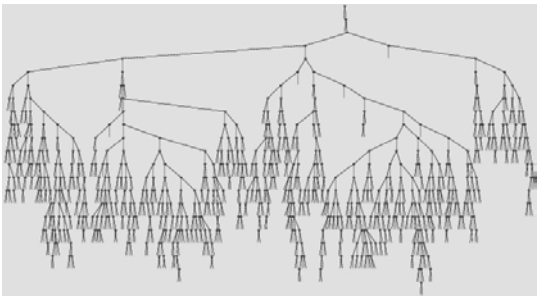
Algorithm:

- Put each item in its own cluster (all singletons)
- Find all pairwise distances between clusters
- Merge the two *closest* clusters
- Repeat until everything is in one cluster

Observations:

- Results in a hierarchical clustering
- Yields a clustering for each possible number of clusters
- Greedy clustering: Result is not "optimal" for any cluster size

Agglomerative Clustering Example



Density-Based Clustering

- A cluster is defined as a connected dense component.
- Density is defined in terms of number of neighbors of a point.
- We can find clusters of arbitrary shape

Clustering: Summary

- Unsupervised Learning
- Many different clustering algorithms
- We talked about:
  - K Means
  - Agglomerative clustering
  - Density based clustering