Knowledge Discovery and Data Mining

Unit # 7

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

Mean and Variance based Feature Selection

- Suppose A and B are sets of feature values measured for two different classes, and n1 and n2 are the corresponding number of samples.
  - SE(A – B) = Sqrt (var(A)/n1 + var(B)/n2)
  - TEST: |mean(A) – mean(B)|/SE(A – B) > threshold value
- It is assumed that the given feature is independent of the others.

Mean-Variance Example

- $SE(X_A – X_B) = 0.17$
- $SE(Y_A – Y_B) = 0.31$
- $|mean(X_A) – mean(X_B)| / SE(X_A – X_B) = 0.196 < 0.5$
- $|mean(Y_A) – mean(Y_B)| / SE(Y_A – Y_B) = 0.750 > 0.5$

X	Y	C
0.3	0.7	A
0.2	0.9	B
0.6	0.6	A
0.5	0.5	A
0.7	0.7	B
0.4	0.9	B

Feature Ranking Exercise

- Given the data set X with three input features and one output feature representing the classification of samples

I1	I2	I3	O
2.5	1.6	5.9	0
7.2	4.3	2.1	1
3.4	5.8	1.6	1
5.6	3.6	6.8	0
4.8	7.2	3.1	1
8.1	4.9	8.3	0
6.3	4.8	2.4	1

- Rank the features using a comparison of means and variances

Entropy-based Measure for Feature Ranking

- Similarity Measure
  - $S_{ij} = e^{-\alpha Dij}$
  - Where $\alpha = -(\ln 0.5) / D$
  - D is the average distance among samples in the data set
  - Normalized Euclidean distance measure is used to calculate the distance Dij between two samples xi and xj (n is the number of dimensions):
  - $Dij = \left[\sum_{k=1}^{n}((x_{ik} - x_{jk})/max_k - min_k))^2\right]^{1/2}$

Entropy-based Measure for Feature Ranking (Cont'd)

- Since all features are not numeric, the similarity for nominal variables is measured directly using Hamming distance.
- $Sij = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}| \right) / n$
- Where $|x_{ik} - x_{jk}|$ is 1 if $x_{ik} = x_{jk}$, and 0 otherwise.

- For mixed data, we can discretize numeric values and transform numeric features into nominal features before we apply this similarity measure.

Entropy-based Measure for Feature Ranking (Cont'd)

- The distribution of all similarities for a given data set is a characteristic of the organization and order of data in an n-dimensional space. This may be measured by entropy.
- The proposed technique compares the entropy measure for a given data set before and after removal of a feature. If the two measures are close, then the reduced set of features will satisfactorily approximate the original set.

- $E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}))$

Sample	F1	F2	F3
R1	A	X	1
R2	B	Y	2
R3	C	Y	2
R4	B	X	1
R5	C	Z	3

	R1	R2	R3	R4	R5
R1		0/3	0/3	2/3	0/3
R2			2/3	1/3	0/3
R3				0/3	1/3
R4					0/3

Algorithm: Entropy based Ranking (Sequential Backward Ranking)

1. Start with the initial full set of features F.
2. For each feature $f \in F$, remove one feature F and obtain a subset F_f. Find the difference between entropy for F and entropy for all F_f.
3. Let f_k be a feature such that the difference between entropy for F and entropy for f_{fk} is minimum.
4. Update the set of features $F = F - \{f_k\}$.
5. Repeat steps 2-4 until there is only one feature.

Entropy-based Feature Ranking Exercise

- Given four-dimensional samples where the first two dimensions are numeric and last two are categorical

X1	X2	X3	X4
2.7	3.4	1	A
3.1	6.2	2	A
4.5	2.8	1	B
5.3	5.8	2	B
6.6	3.1	1	A
5.0	4.1	2	B

- Apply a method for unsupervised feature selection based on entropy measure to reduce one dimension from the given data set