Knowledge Discovery and Data Mining

Unit # 13

Distance Computation

- Interval-Scaled Variables
- Binary Variables
- Categorical Variables
- Ordinal Variables

Interval-Scaled Variables

- *Interval-scaled variables are continuous measurements* of a roughly linear scale.
- Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.
- Both Euclidean distance and Manhattan distances are generally used for distance computation.

Binary Variables

- One approach involves computing a dissimilarity matrix from the given binary data.
- If all binary variables are thought of as having the same weight, we have the 2-by-2 contingency table, where
  - q is the number of variables that equal 1 for both objects i and j,
  - r is the number of variables that equal 1 for object i but that are 0 for object j,
  - s is the number of variables that equal 0 for object i but equal 1 for object j, and
  - t is the number of variables that equal 0 for both objects i and j.
- The total number of variables is p, where p = q+r+s+t.

Symmetric Binary Variables

- *A binary* variable is symmetric if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1.
- One such example could be the attribute *gender having the states male and female.*

$$d(i, j) = \frac{r+s}{q+r+s+t}.$$

Asymmetric Binary Variables

- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the *positive and negative outcomes of a disease test.*
- *By convention,* we shall code the most important outcome, which is usually the rarest one, by 1
- (e.g., *HIV positive) and the other by 0 (e.g., HIV negative).*

$$d(i, j) = \frac{r+s}{q+r+s}. \qquad sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j).$$

- The coefficient sim(i, j) is call the Jaccard coefficient, which is popularly referenced in the literature.

Categorical Variable

- A categorical variable is a generalization of the binary variable in that it can take on more than two states.
- For example, map color is a categorical variable that may have, say, five states: red, yellow, green, pink, and blue.
- The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i, j) = \frac{p - m}{p},$$

- where m is the number of matches (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables.

Ordinal Variables

- The treatment of ordinal variables is quite similar to that of interval-scaled variables when computing the dissimilarity between objects.
- The dissimilarity computation with respect to f involves the following steps:

1. The value of f for the ith object is x_{if}, and f has M_f ordered states, representing the ranking $1, \ldots, M_f$. Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$.
2. Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto [0.0,1.0] so that each variable has equal weight. This can be achieved by replacing the rank r_{if} of the ith object in the fth variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

3. Dissimilarity can then be computed using any of the distance measures described for interval-scaled variables

Mixed Type Variables

Suppose that the data set contains p variables of mixed type. The dissimilarity $d(i, j)$ between objects i and j is defined as

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}, \tag{7.15}$$

where the indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing (i.e., there is no measurement of variable f for object i or object j), or (2) $x_{if} = x_{jf} = 0$ and variable f is asymmetric binary; otherwise, $\delta_{ij}^{(f)} = 1$. The contribution of variable f to the dissimilarity between i and j, that is, $d_{ij}^{(f)}$, is computed dependent on its type:

- If f is interval-based: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for variable f.
- If f is binary or categorical: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as interval-scaled.

Vector Objects

- In some applications, such as information retrieval, text document clustering, and biological taxonomy, we need to compare and cluster complex objects (such as documents) containing a large number of symbolic entities (such as keywords and phrases).

There are several ways to define such a similarity function, $s(\boldsymbol{x}, \boldsymbol{y})$, to compare two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. One popular way is to define the similarity function as a **cosine measure** as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}|| ||\boldsymbol{y}||}, \tag{7.16}$$

where $\boldsymbol{x}^t$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$, and s is essentially the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

[1]The Euclidean normal of vector $x = (x_1, x_2, \ldots, x_p)$ is defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_p^2}$. Conceptually, it is the length of the vector.

Recap of K-Means

- The K-Means node provides a method of cluster analysis.
- It can be used to cluster the data set into distinct groups when you don't know what those groups are at the beginning.
- Instead of trying to predict an outcome, K-Means tries to uncover patterns in the set of input fields.
- Records are grouped so that records within a group or cluster tend to be similar to each other, but records in different groups are dissimilar.
- Note: The resulting model depends to a certain extent on the order of the training data. Reordering the data and rebuilding the model may lead to a different final cluster model.

Recap of K-Means (Cont'd)

- K-Means works by defining a set of starting cluster centers derived from data.
- It then assigns each record to the cluster to which it is most similar, based on the record's input field values.
- After all cases have been assigned, the cluster centers are updated to reflect the new set of records assigned to each cluster.
- The records are then checked again to see whether they should be reassigned to a different cluster, and the record assignment/cluster iteration process continues until either the maximum number of iterations is reached, or the change between one iteration and the next fails to exceed a specified threshold.

What is Fuzzy Logic

- Definition of Fuzzy
  - Fuzzy: "not clear, distinct, or precise; blurred"
- Definition of Fuzzy Logic
  - A form of knowledge representation suitable for notations that cannot be defined precisely but which depend upon their contexts.
- The term was coined by Lotfi Zadeh in 1965 with his mathematics of fuzzy set theory.

Examples of Linguistic Impression

- How was the weather like yesterday?
  - Oh! It was rainy with 98% humidity and hot with temperature of 35.5 deg C
  - Oh! It was very humid and really hot.

* Source: University Malaysian Pahang

Examples of Linguistic Impression (Cont'd)

- When you are at 10 meters from the junction start braking at 50% pedal level.
- When you are near the junction, start braking slowly.

* Source: University Malaysian Pahang

Fuzzy c-Means

- The fuzzy *c*-means algorithm is very similar to the *k*-means algorithm:
  - Choose a number of clusters.
  - Assign randomly to each point coefficients for being in the clusters.
  - Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than , the given sensitivity threshold) :
    - Compute the centroid for each cluster, using the formula on the next slide.
    - For each point, compute its coefficients of being in the clusters, using the formula on the next slide.
  - The algorithm minimizes intra-cluster variance as well, but has the same problems as *k*-means, the minimum is a local minimum, and the results depend on the initial choice of weights.

Fuzzy c-Means (Cont'd)

$$\forall x \left(\sum_{k=1}^{\text{num. clusters}} u_k(x) = 1 \right).$$

With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$\text{center}_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}.$$

The degree of belonging is related to the inverse of the distance to the cluster center:

$$u_k(x) = \frac{1}{d(\text{center}_k, x)},$$

- then the coefficients are normalized

Example

- Data: {8, 12, 3, 7, 15, 4, 10, 20, 6, 19}
- Perform K-Means (where K = 2)
- Perform the same exercise using Fuzzy c-Means (where c=2)

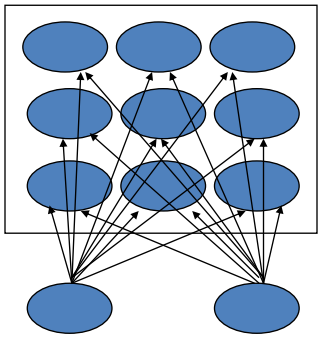
KNIME Demo (Clustering)

Kohonen Map

- Kohonen networks are a type of neural network that perform clustering, also known as a knet or a self-organizing map.
- This type of network can be used to cluster the data set into distinct groups when you don't know what those groups are at the beginning.
- Records are grouped so that records within a group or cluster tend to be similar to each other, and records in different groups are dissimilar.
- The basic units are neurons, and they are organized into two layers: the input layer and the output layer (also called the output map).

Kohonen Map (Cont'd)

- Formalized by Teuvo Kohonen in 1982 for unsupervised clustering.
- All of the input neurons are connected to all of the output neurons, and these connections have strengths, or weights, associated with them.
- During training, each unit competes with all of the others to "win" each record.
- Input data is presented to the input layer, and the values are propagated to the output layer. The output neuron with the strongest response is said to be the winner and is the answer for that input.



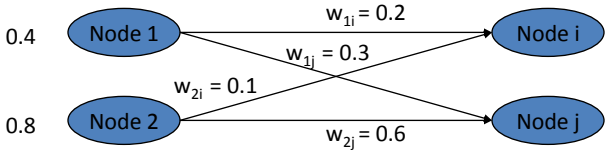
Kohonen Map (Cont'd)

- Initially, all weights are random. When a unit wins a record, its weights (along with those of other nearby units, collectively referred to as a neighborhood) are adjusted to better match the pattern of predictor values for that record.
- All of the input records are shown, and weights are updated accordingly. This process is repeated many times until the changes become very small.
- As training proceeds, the weights on the grid units are adjusted so that they form a two-dimensional "map" of the clusters (hence the term self-organizing map).

Kohonen Map (Cont'd)

- When the network is fully trained, records that are similar should appear close together on the output map, whereas records that are vastly different will appear far apart.
- Usually, a Kohonen net will end up with a few units that summarize many observations (strong units), and several units that don't really correspond to any of the observations (weak units). The strong units (and sometimes other units adjacent to them in the grid) represent probable cluster centers.

Working of Kohonen Maps



- The score for classifying a new instance with output node j is given by

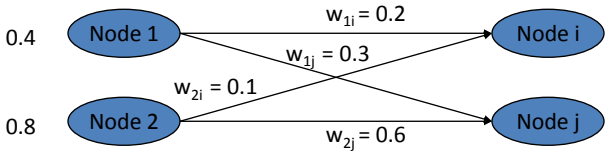
$$\text{sqrt}\left(\sum (n_i - w_{ij})^2\right)$$

  - n_i is the attribute value for the current instance at input i.
  - w_{ij} is the weight associated with the ith input node and output node j.
- Weights are updated according the following formula:

$$w_{ij}\,(\text{new}) = w_{ij}\,(\text{current}) + \Delta w_{ij}$$

  - where $\Delta w_{ij} = r(n_i - w_{ij})$, r is the learning parameter and 0<r<1.

Working of Kohonen Maps (Cont'd)



- Score of Node i: sqrt(0.4-0.2)^2 + (0.8-0.1)^2 = 0.53
- Score of Node j: sqrt(0.4-0.3)^2 + (0.8-0.6)^2 = 0.05

- Thus, the record belongs to Cluster j.

- Next we update the weights of incoming links to node j. Let r = 0.8
- Δw_{1j} = 0.8 x (0.4 – 0.3) = 0.08
- Δw_{2j} = 0.8 x (0.8 – 0.6) = 0.16
- w_{1j} = 0.3 + 0.08 = 0.38
- w_{2j} = 0.6 + 0.16 = 0.78

Kohonen Map (Cont'd)

- The simplicity of this algorithm makes it a great choice for clustering.
- One primary disadvantage of the algorithm is that the number of output classes must be defined upfront.
- This is significant because it assumes that we have some general knowledge of the data and how it should be classified.