L15 Unsupervised Learning

Unsupervised Learning

- Unsupervised procedures use unlabeled samples
- A collection of samples without being told their categories/function values
- With tons of data being generated every millisecond, it is no surprise that most of this data is unlabeled
- As there is no output variable to guide the learning process
  - data is explored by algorithms to find patterns.
  - since the data has no labels, the algorithm identifies similarities on the data points and groups them into clusters

Unsupervised Learning

- The purpose of supervised learning is to
  - answer a specific question ‘phrased’ as a target variable and
  - the various techniques to discover patterns in data to discern the value of the target.
- Historical data contains examples to find the best answer.
- Target variable is a way of putting domain expertise into the modeling process.
- A data miner looks for, ‘This is what is important’.
- With this information, supervised techniques have enough inputs to optimize their models.
- Unsupervised procedures have no such information.
- The data miner looks for ‘something interesting’ in the data.
- Requires more human understanding than supervised techniques

Unsupervised Learning

- Unsupervised learning vs Supervised learning: Differences and Similarities
- Similarities :
- Both work with data
- require exploration and understanding of the data with respect to the application domain
- Both are improved by incorporating intelligent variables into the data that identify the quality of different aspects of the process in domain of interest
- Because source data is usually not at the level required by the application in hand, building the data for the application requires many transformations in both the learning approaches

Unsupervised Learning

Differences:

Unsupervised learning differs in not having a target variable, and this poses challenges:

1. Unsupervised learning is more about creative endevours—exploration, understanding and refinements that do not lend themselves to a specific set of steps
   - So there is no specific methodology.
   - The unsupervised learning process cannot be automated.
   - Techniques are still not available to distinguish between the more useful and less useful.
   - Humans play a crucial role.

2. There is no right or wrong answer;
   - no simple statistical measure that summarizes the goodness of results.
   - Instead, descriptive statistics and visualization are key parts of the process.

3. The requirement of dataset partitioning— into training, validation and test sets—is of little importance

4. Supervised learning, with a few exceptions, is often based on statistical techniques that have been around for many decades or centuries. These have adapted to the increased volumes of data and computer power available in the modern world.

Clustering (a tool of unsupervised learning), on the other hand, is more recent and borne out of availability of lots of data and powerful computers

Unsupervised Learning

Techniques associated with unsupervised learning

- Clustering techniques
- Association rules techniques

Clustering

- In training datasets for clustering problems, outputs (class labels/numeric/ordinal) are not specified
- only the feature vectors representing different objects/instances/records/situations are known.
- Grouping a set of data objects to form several groups or clusters, in such a way that the components within a cluster are highly similar but differ from the components of other clusters, it is called clustering.
- The level of similarity and dissimilarity are evaluated on the basis of the characteristics of the variables that describe the objects or components.
- This assessment often involves distance measures

Clustering

- Application examples of clustering :
- Clustering can be used for data exploration, to understand the structure of the data.
- The data may contain complex structure that even the best data mining techniques are unable to coax out (arrange) meaningful patterns within the data.
- Cluster detection provides a way to learn about the structure of complex data.
- A natural way to make sense of complex data is to break the data into smaller clusters of data; then finding patterns within each cluster is often possible

Clustering

- Application examples of clustering :
- Image compression
- In an image, say we decide to color the same group with a single color
- If 24 bit pixels represent 16 million colors for an image
- But if there are shades of merely 64 main colors, then we will require 6 bits for each pixel rather than 24.
- Document clustering
- Aims to classify similar documents.
- Ex: news reports can be further divided pertaining to politics, entertainment, sports etc

Clustering

- Application examples of clustering :
- Bioinformatics
- in learning sequences of amino acids that occur repeatedly in protiens;
- Customer segmentation
- If a company has the data pertaining to past customers, along with demographic information and past transactions with the company
- the clustering approach ensures that customers with similar attributes are assigned the same group.
- As a result, the customers of the company end up being in natural groups.

Clustering

- Application examples of clustering :
- Customer relationship management
- Ex: if an organization's clients fit in one of the K groups known as segments,
- then there will be better understanding of the customer base letting the organization offer different strategies for different segments.
- Outlier detection
- Finding instances that do not lie in any of the main clusters and are exceptions.
- The outliers may be recording errors that should be detected and discarded in data cleansing process.
- However, an outlier may indicate abnormal behavior
- Ex: in a dataset of credit card transactions, it may indicate a fraud;
- in an image, it may indicate anomalies, for example, tumors; or it may be a novel, previously unseen but valid case;
- customers who do not fall in any large group may require special attention—churning customers (customers cease their relationship with a company by discontinuing their use of products or service)
- Outlier detection is given the name anomaly detection or novelty detection

ENGINEERING THE DATA

- Today's real-world data are typically of huge size and their origin is from multiple heterogeneous sources.
- The low-quality data will lead to low-quality results
- Careful preprocessing of the available raw data improves the efficiency and ease of data mining
- Data engineering - Engineering the data into a form suited to the selected learning scheme.
- Four important processes of data preparation for successful data mining:
  - data cleansing
  - derived attributes
  - discretizing numeric attributes and
  - attribution reduction
- All these processes result in data transformation to obtain high-quality mining performance

ENGINEERING THE DATA

Data Transformations:

- Because the source data is usually not at the level required by the application in hand, building the data for the application requires many transformations
- 1. Data Cleansing
- Discrepancies in data can be caused by various factors including poorly designed data collection, human-errors in data entry, deliberate errors (e.g., respondents not willing to divulge information), inconsistent data representations, errors in instrumentation that record data and system errors.
- Practical data mining process has to deal with errors in raw data such as
  - missing values,
  - inaccurate values,
  - outliers,
  - duplicate data,
  - sparse data,
  - skewed data,
  - noisy data,
  - anomalies because of some systematic errors, etc

ENGINEERING THE DATA

Data Transformations:

2. Derived Attributes

- Significant aspect of the data transformation procedure pertains to the definition of new variables.
- These convey the information intrinsic in the data in such a way that the information becomes more useful for data mining methods.
- Derived variables for data preparation may involve creation of new variables through creative transformation of existing variables.
- When measuring variables on different scales, their standardization also becomes essential.
- For data mining methods that work on numerical data alone, it is necessary to numerically represent categorical data in some way

ENGINEERING THE DATA

Data Transformations:

3. Discretizing Numeric Attributes

- Discretization of numeric attributes is essential if the task involves numeric attributes but the learning scheme chosen can only handle categorical ones.
- Even schemes that can handle numeric attributes often produce better results, or work faster, if the attributes are prediscretized.

ENGINEERING THE DATA

Data Transformations:

4. Attribute Reduction

- Datasets for analysis may contain hundreds of attributes, many of them may be irrelevant to the mining task.
- The role of domain expert is very important to pick out the relevant attributes.
- Keeping irrelevant attributes can result in performance of the learning scheme to deteriorate.
- In addition, the added volume of irrelevant or redundant attributes can slow down the mining process.
- Feature extraction using domain knowledge is thus an important initial step in data mining process.

ENGINEERING THE DATA

Data Transformations:

4. Attribute Reduction

- Data mining success will suffer more if relevant features are left out compared to the loss in keeping irrelevant redundant attributes.
- This is because data transformation techniques can help eliminate the redundant attributes
- But the relevant ones left out at the feature extraction stage cannot be recovered.
- Domain experts normally select a larger set of features, keeping all those for
- which they have doubts about redundancy.
- This larger set is then processed using transformation techniques for reduction of the attributes.