Assignment - 3.1

Data Mining Tasks, Supervised vs Unsupervised Learning

Harsh Siddhapura
1230169813

IFT511 - Big Data Analysis
Prof. Asmaa Elbadrawy

18th January, 2024

PROBLEM - 1

1. **Classification**: Imagine sorting laundry. Classification is similar, but instead of clothes, you're sorting data points into predefined categories. You analyze existing data with known labels (e.g., emails labeled "spam" or "not spam") to build a model that can automatically categorize new data points into those categories. This is helpful for everything from predicting loan defaults to identifying fraudulent transactions.

2. **Regression**: Picture predicting the future. Regression goes beyond categories and focuses on continuous values. You train a model based on existing data (e.g., house prices and features) to estimate an outcome variable (e.g., future price of a new house). This is used for things like forecasting sales trends or predicting stock prices.

3. **Clustering**: Think of organizing your bookshelf. Clustering groups similar data points together without any predefined labels. It's like discovering natural "tribes" in your data based on shared features. This helps identify customer segments, product recommendations, and anomaly detection.

4. **Similarity Matching**: Ever wondered if two products are similar? Similarity matching finds data points that closely resemble each other based on specific features. Imagine searching for music similar to your favorite song or recommending movies based on your past preferences.

5. **Frequent Itemset Mining**: Picture analyzing grocery carts. This task discovers patterns of items frequently bought together. Imagine finding which products often end up in the same basket. This helps retailers stock shelves strategically and recommend complementary items to customers.

6. **Profiling**: Remember building a detailed character description in a story? Profiling does the same for data points. It analyzes individual data points to create detailed profiles based on their unique characteristics. This helps identify high-risk customers, target specific advertising campaigns, and personalize user experiences.

7. **Link Prediction**: Imagine connecting the dots. Link prediction tries to guess future connections between data points based on existing relationships. This is used for social network analysis, predicting product interactions, and even crime prevention.

8. **Data Reduction**: Think of decluttering your room. Data reduction helps condense massive datasets into a smaller, manageable form while preserving essential information. This is crucial for faster processing, improved model training, and efficient data storage.

9. **Causal Modeling**: Imagine untangling a web of causes and effects. Causal modeling aims to discover the true cause-and-effect relationships between variables, going beyond mere correlations. This helps in medical research, policy analysis, and understanding complex economic systems.

PROBLEM - 2

1. Regression:
    - Category: Supervised Learning
    - In regression, the algorithm is trained on a labeled dataset where the target variable is continuous. The goal is to learn a mapping from input features to a continuous output. The algorithm is supervised because it learns from labeled examples to predict a continuous value.

2. Frequent Itemset Mining:
    - Category: Unsupervised Learning
    - Frequent itemset mining is an unsupervised learning task. It involves discovering patterns or associations in a dataset without predefined labels. The algorithm identifies sets of items that frequently co-occur in a transactional database without the need for labeled data.

3. Profiling:
    - Category: Typically Unsupervised, but Context-Dependent
    - Profiling can be both supervised and unsupervised depending on the context. In an unsupervised setting, profiling may involve analyzing patterns or characteristics within a dataset without using predefined labels. In a supervised setting, profiling may involve creating user or entity profiles based on labeled data.

4. Link Prediction:
    - Category: Typically Supervised, but Context-Dependent
    - Link prediction can be both supervised and unsupervised. In a supervised setting, the algorithm may learn from labeled examples to predict the likelihood of a link between nodes in a network. In an unsupervised setting, the algorithm may infer potential links based on the structure of the network without using labeled data.

5. Data Reduction:
    - Category: Unsupervised Learning
    - Data reduction techniques, such as dimensionality reduction or feature selection, often fall under unsupervised learning. These methods aim to reduce the complexity of the

dataset without relying on labeled information. They focus on intrinsic data properties rather than specific labels.

6. Causal Modeling:
    - Category: Typically Unsupervised, but Context-Dependent
    - Causal modeling can be both supervised and unsupervised, depending on the context. In an unsupervised setting, the goal may be to identify causal relationships without using labeled data. In a supervised setting, the algorithm may learn causal relationships from labeled examples where cause-and-effect relationships are explicitly provided.