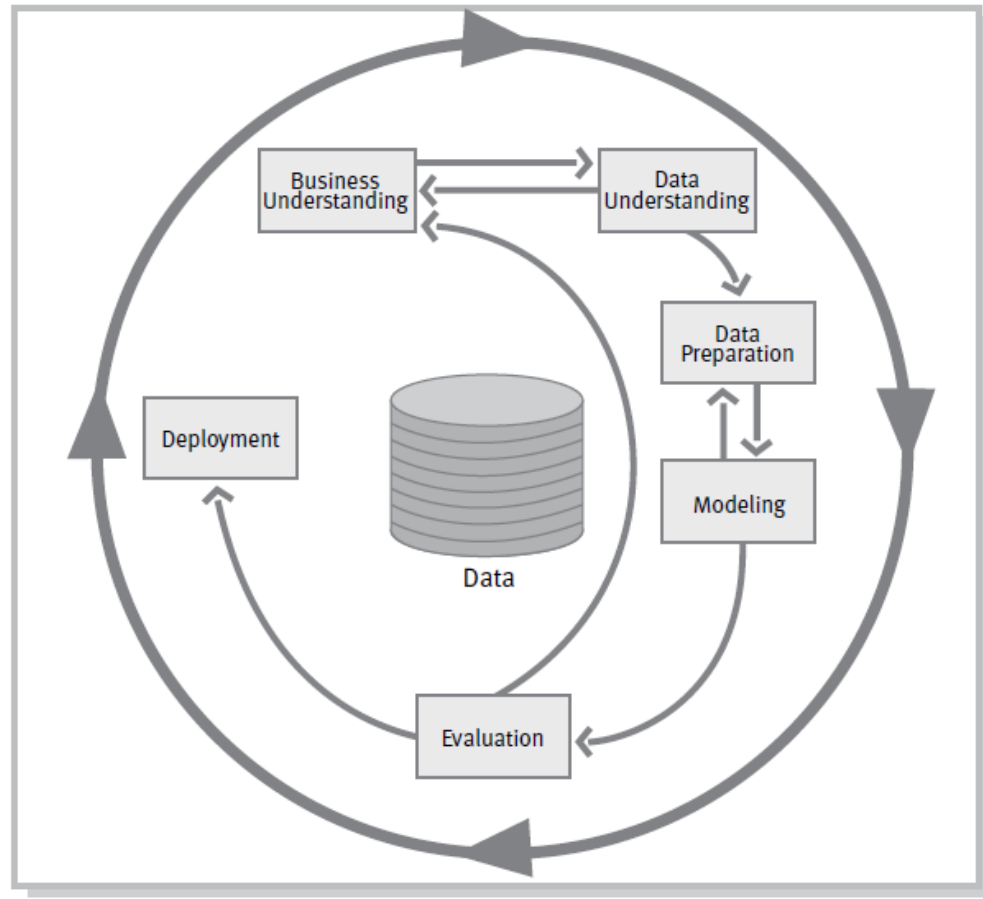
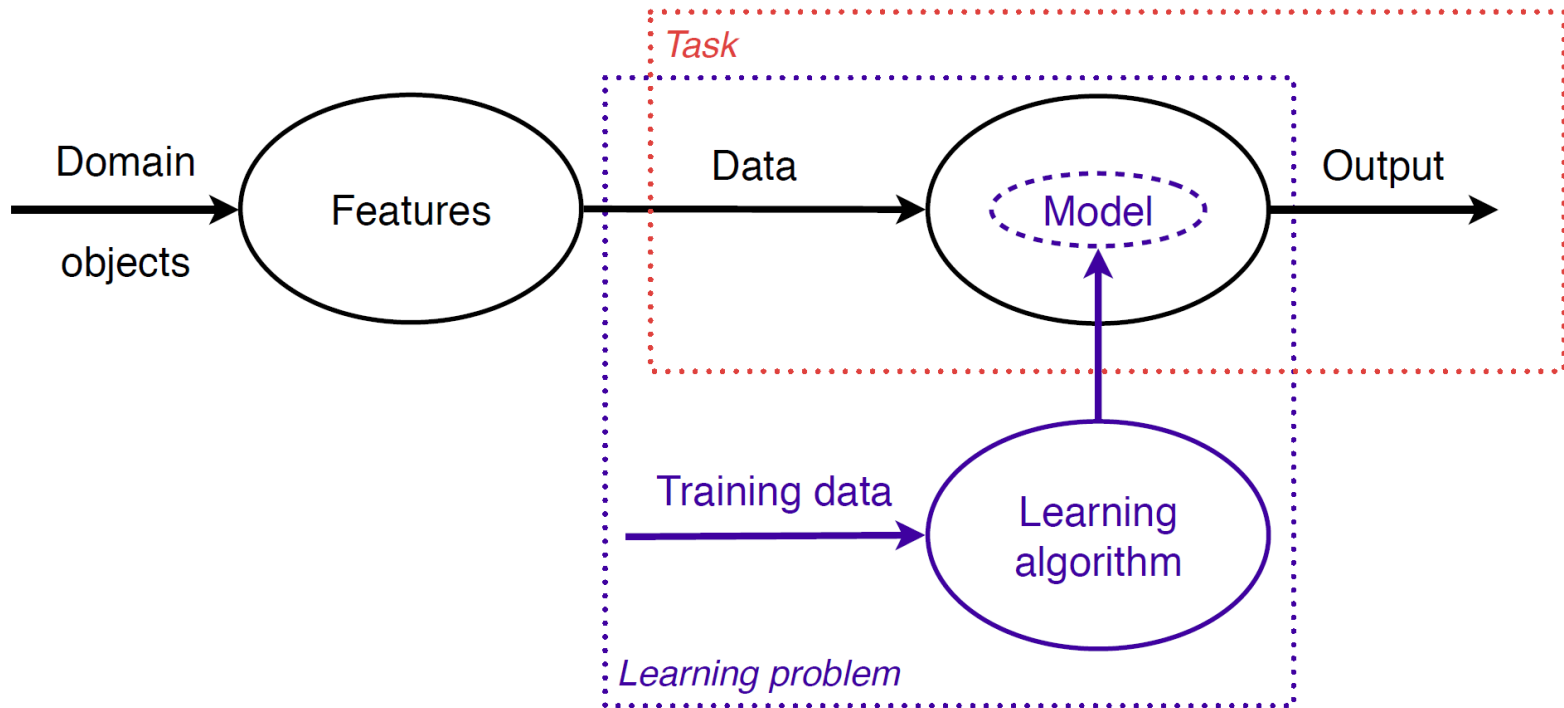


Lesson B-1

Introduction to Data Mining Modeling and Algorithms

The life cycle of a data mining project





Learning Styles (1)

- Supervised learning
 - The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations and new data is classified based on the training set
 - Classification, regression, neural networks
- Unsupervised learning
 - The class labels of training data is unknown and the process of grouping a set of objects into classes of similar objects
 - Clustering, association rules

Learning Styles (2)

- Semi-supervised learning
 - Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data.
 - RNN, DBN, RMS
- Reinforced learning
 - The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return or penalties in the form of negative rewards
 - DeepMind's AlphaGo

Types of Algorithms ⁽¹⁾

- Classification
 - algorithms can be used to build a model that predicts the outcome class for a given dataset
 - Decision tree, kNN, Bayesian, ensemble, etc.
- Regression
 - a statistical method for examining the relationship between two or more variables to predict numeric values
 - Regression (simple linear, nonlinear, polynomial), regularization, support vector machine, etc.

Types of Algorithms (2)

- Neural networks
 - an artificial model based on the human brain and it learns task without being told any specific rules
 - Perceptron, back-propagation, MLP etc.
- Deep learning
 - Special type of neural networks and deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks
 - RNN, CNN, DBN, LVQ, RBM, etc.

Types of Algorithms (3)

- Ensemble
 - meta-algorithms that combine several machine learning methods into a single predictive model to increase the overall performance
 - Random forest, bagging, AdaBoost, stacking etc.
- Clustering
 - Clustering is the practice of assigning labels to unlabeled data using the patterns that exist in it
 - Partitioning (k-means, k-medians, k-modes), hierarchical (agglomerative, divisive), density-based (DBSCAN, OPTICS)

Types of Algorithms (4)

- Association rule
 - uncover how items are associated with each other
 - Frequent pattern mining, Apriori algorithm
- Anomaly detection
 - anomaly detection is used to find rare occurrences or suspicious events in your data
 - Isolation forest, PCA-based anomaly detection, IQR-based, standard deviation