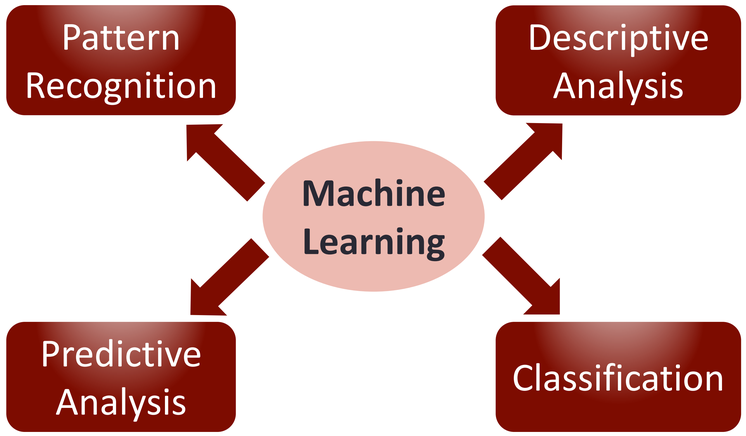
Machine Learning and its Applications

Yuexi Wang[1] and Hossein Pourreza[1]
[1]Research Computing Center

BACKGROUND

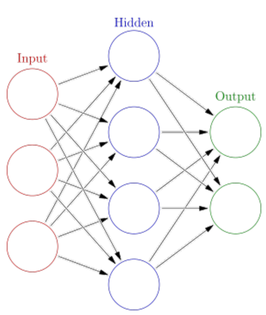
Machine learning gives computers the ability to learn and make decisions without being explicitly programmed. Machine learning is used in different disciplines from medical imaging to data analysis in social science.

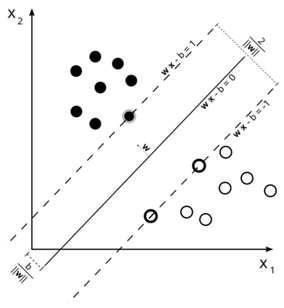


Machine learning algorithms can be supervised (trained on labeled data) or unsupervised (no training data).

Supervised learning algorithms:

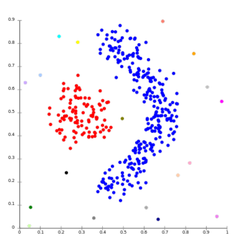
- Decision Tree
- Bagging
- Linear Regression
- Naïve Bayes
- SVM
- Neural Network





Unsupervised algorithms:

- Hierarchical Clusters
- K-means



STEPS OF MODELING

Data Collection
- Data is collected and preprocessed
  - Missing values
  - Typos and mixed languages

Feature Selection
- Indicative features are selected and described
  - Correlation among variables
  - Ordinal/categorical/dummy variables or responses
  - Overview of distribution of the data

Model Selection
- A machine learning algorithm is selected
  - Supervised vs unsupervised

Train and Test
- Data is divided into training and testing sets
  - Ratio between two sets
  - Cross validation

Compare and Evaluate
- Measure error rates
  - Accuracy, precision, and recall
  - Type I and II errors

CLOUD-BASED SOLUTIONS

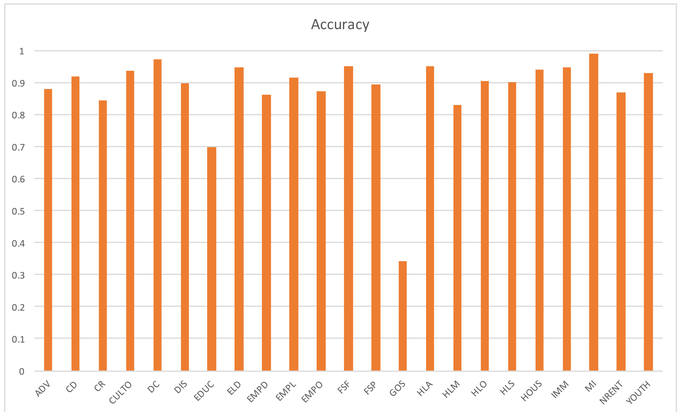
Due to the popularity of using machine learning techniques in different areas, cloud providers such as IBM and Microsoft offer machine learning solutions. Mostly working as a black box, such solutions receive training data and then build and train a model. Once it is ready, it can be applied to the test data.

The cost associated with using these services as well as working as a black box could be limiting factors to use these services.

RCC-ENABLED PROJECTS

RCC uses Natural Language Processing methods to process tweet data and classify them as gang-related. The following word cloud was generated based on the frequency of extracted words from collected tweets.

Machine learning techniques are used to automatically assign public contracts to pre-determined categories based on contract descriptions. The following chart shows the accuracy of the classification in different categories.



RCC uses GPU-enabled deep learning technique to predict cancer-prone regions within the Prostate