Tools and Algorithms in Bioinformatics

GCBA815, Fall 2013

Week6: Introduction to Machine Learning and WEKA package

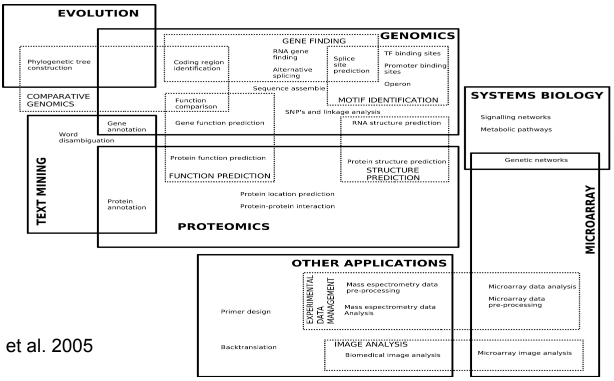
Guest lecture: Akram Mohammed

Department of Genetics, Cell Biology and Anatomy

University of Nebraska Medical Center

Machine Learning (ML)

- Develop an automated system that learns from the input data and build models for predicting the unknown instances
- Making predictions or decisions from data
- ML is ideally suited for areas where there is a lot of data but little theory



Larranaga et al. 2005

Components of Machine Learning

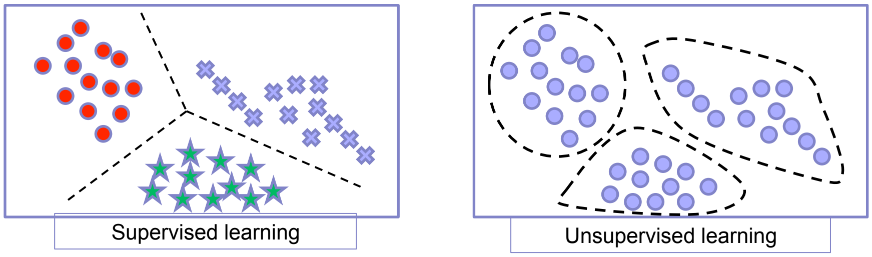
- the data (data-driven approach)
- Attributes extracted from the data
- the classifier model

Data

- labeled data (input data with label)
- unlabeled data (input data without label)

Types of Machine Learning

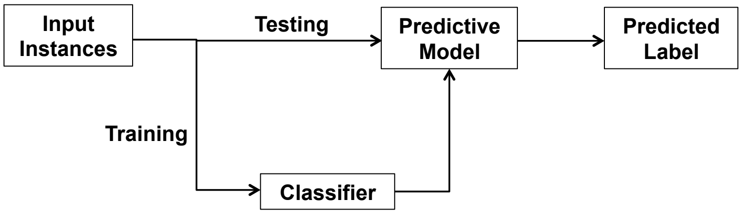
- **Supervised:** Classification: predicting an item class; given input data and class label
- **Unsupervised:** Clustering: finding clusters in data; given input data



Supervised Learning

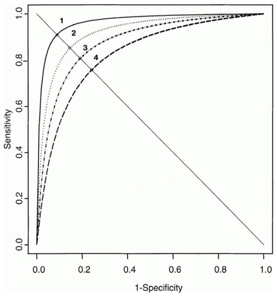
- Involves determining what attributes make the input instance unique to the class, and adjusting the classifier model to best exploit those attributes.
- The training data is analyzed to establish patterns between the input data and the classes the data is assigned to.

Supervised Learning Workflow



Classifiers

- Support Vector Machine (SMO)
- Naïve Bayes
- Bayesian Network
- Random Forest
- K-nearest Neighbor
- Decision Tree (J48)
- Neural network



Metrics for validation

- N-fold cross validation
- Sensitivity
- Specificity
- ROC curves
- A probability distribution for all classes is generated for each sequence, where the class with highest probability is assigned as the predicted class

WEKA (<u>W</u>aikato <u>E</u>nvironment for <u>K</u>nowledge <u>A</u>nalysis)

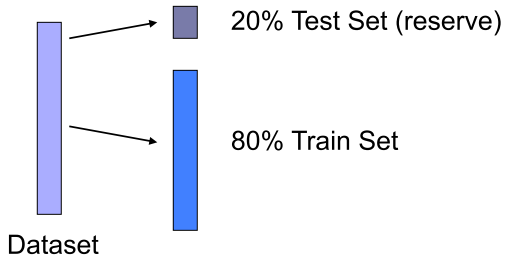
- Machine learning and data mining framework
- Collection of machine learning algorithms
- Written in java
- Simple (GUI) and Advanced (Command-Line)
- Bird found only on the islands of New Zealand

Tools (or Functions)

- Data Preprocessing (filters for Add/remove attribute etc.)
- Attribute Selection (subsets of attributes which are the most predictive ones)
- Classification
- Clustering
- Association Rules
- Data Visualization

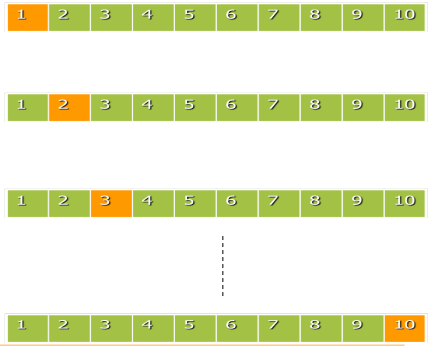
Data import : ARFF and CSV
Row(Instance)
Column(attribute)

Data Partitioning
Random partition into train and test set
Stratification: Unbalanced data; Class distributions is retained



Evaluation: n-fold cross validation

- In each iteration, one fold is used for testing and (n-1) folds are used for training the classifier
- Class distributions is retained in each fold
- Test results are averaged over all folds
- This allows for nearly unbiased estimates of model performance
- Assess the performance of the fully-trained classifiers using the test set (testing)



Classification Demo:

iris.arff
breast-cancer.arff
diabetes.arff