Introduction to Machine Learning Applications

Spring 2021

Lecture-8

Lydia Manikonda

manikl@rpi.edu

Rensselaer

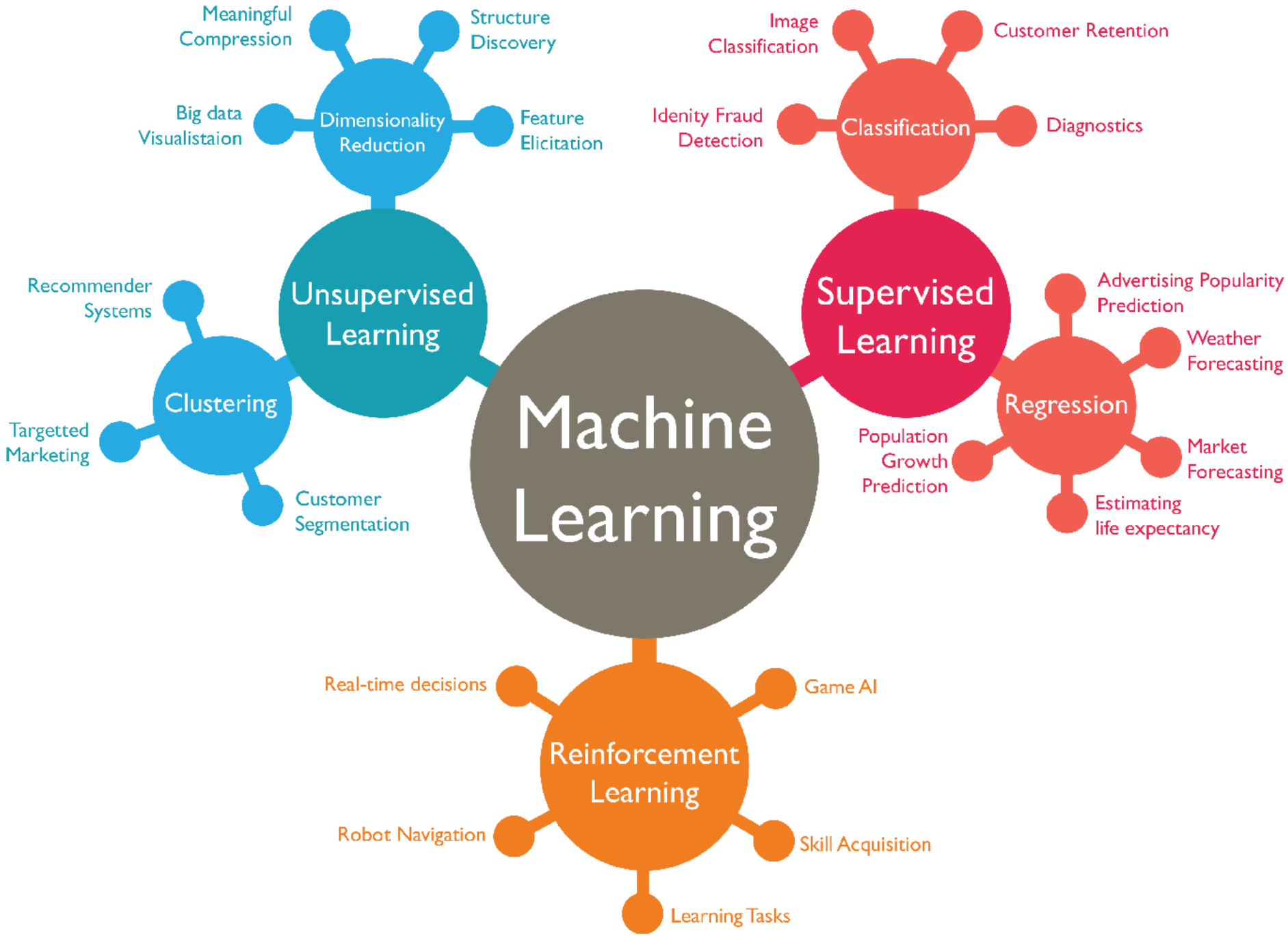
Today’s agenda

- Manipulating Strings and regular expressions
- Overview of Modeling

Announcements

- Homework-3 due on February 25th 11:59 pm ET via LMS

Python notebook on Regular Expressions


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Machine Learning

According to Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E

Well-defined learning task: <P, T, E>

Modeling

What is a model?

- Mathematical representation of a real-world process.
- In other words, description of a system using mathematical concepts.

- Three different types of models can be built:
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning

Definition of Classification

Given a collection of records (training set)

- Each record is by characterized by a tuple ($\boldsymbol{x}$,y), where $\boldsymbol{x}$ is the attribute set and y is the class label
  - # $\boldsymbol{x}$: attribute, predictor, independent variable, input
  - # y: class, response, dependent variable, output

Task:

- Learn a model that maps each attribute set $\boldsymbol{x}$ into one of the predefined class labels y

Example -- Classification tasks

Task	Attribute set, $\boldsymbol{x}$	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

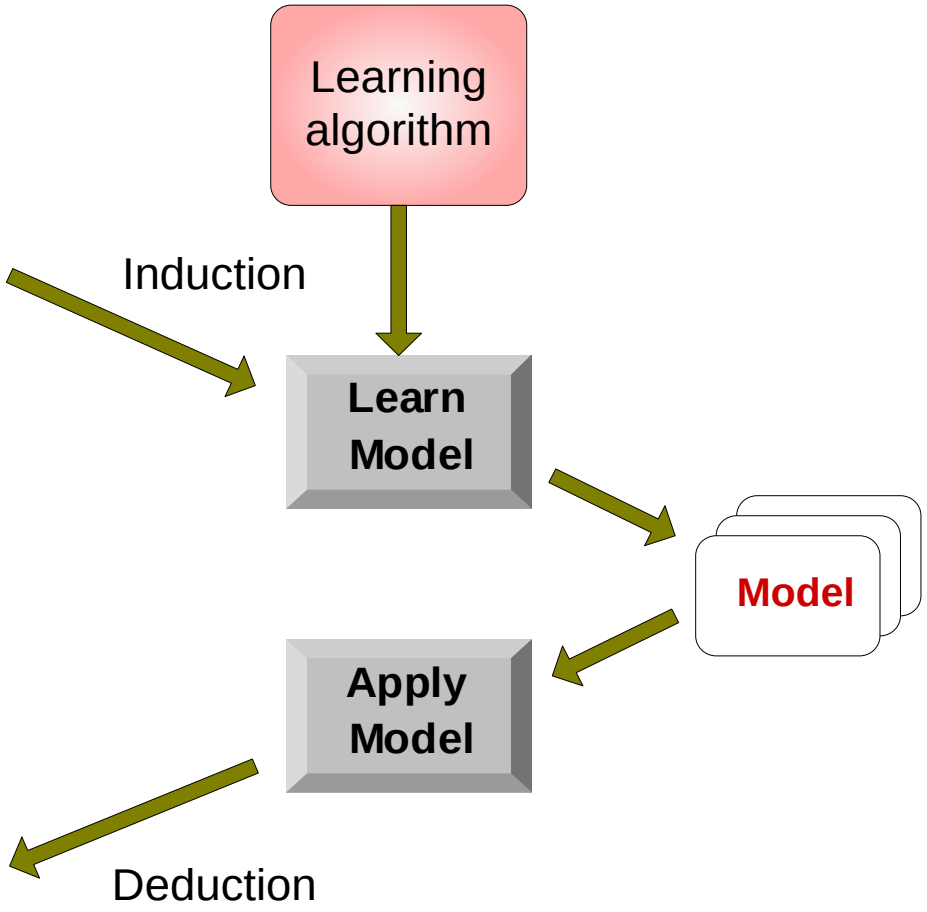
Classification model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



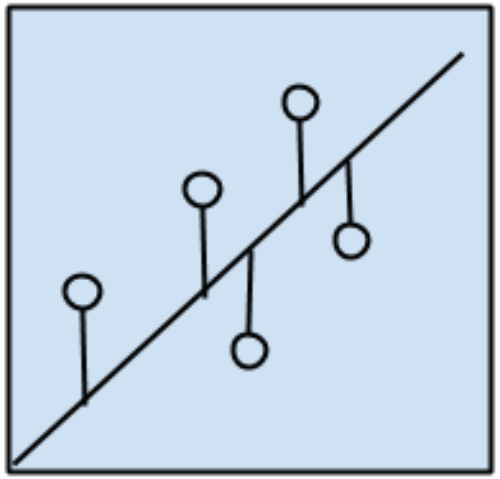
Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines

- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

Regression Algorithms

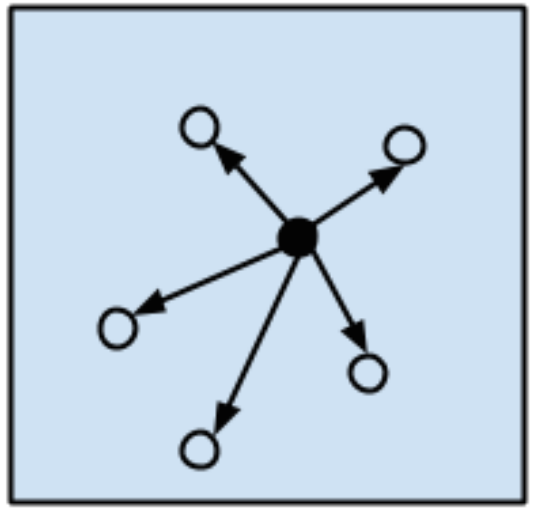
- Modeling the relationship between variables that are iteratively refined using a measure of error.
- Most popular regression algorithms are:
  - Ordinary least squares regression
  - Linear regression
  - Logistic regression
  - Multivariate adaptive regression splines
  - …



Regression Algorithms

Instance-based algorithms

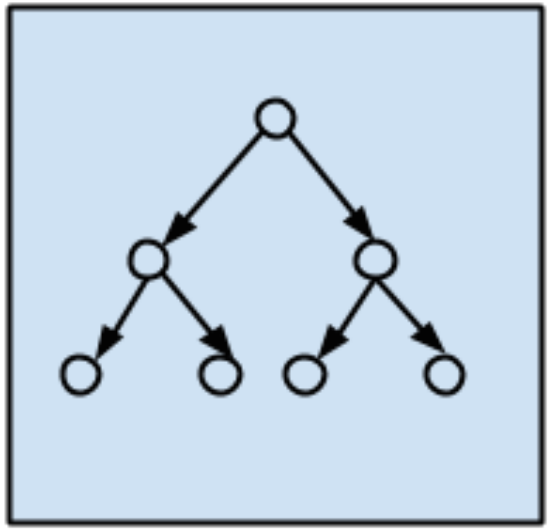
- This model is a decision problem with instances of training data that are deemed important or required to the model.
- Focus is put on the representation of the stored instances and similarity measures used between instances.
- Most popular instance-based algorithms are:
  - K-Nearest Neighbor (KNN)
  - Support Vector Machines (SVM)
  - Learning Vector Quantization
  - Self-Organizing Maps
  - …



Instance-based Algorithms

Decision Tree-based algorithms

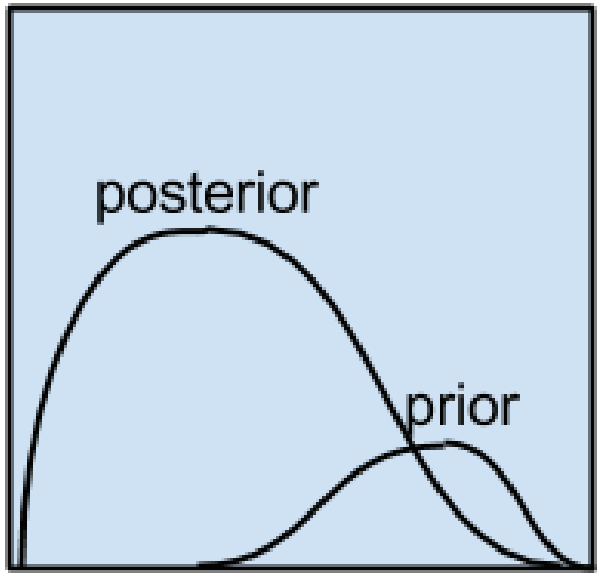
- These methods construct a model of decisions based on the actual values of attributes in the data.
- These decisions built are in the form of a tree.
- Most popular algorithms are:
  - Classification and Regression Tree
  - Conditional Decision Trees
  - ID3
  - C4.5 and C5.0
  - …



Decision Tree Algorithms

Bayesian Algorithms

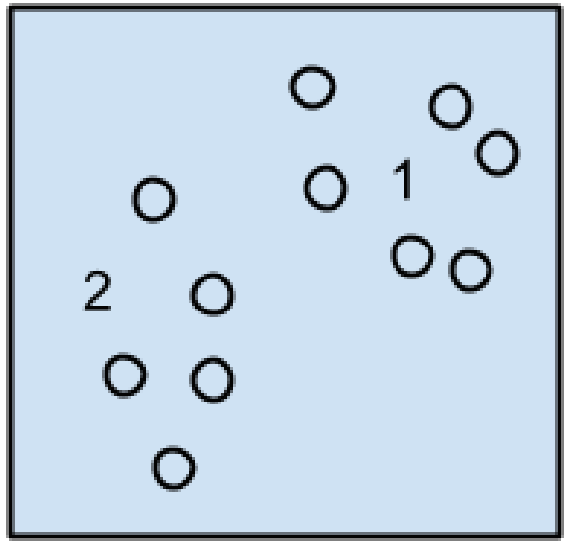
- Bayesian methods explicitly apply the Bayes Theorem for problems such as classification and regression.
- Bayes Theorem
- Most popular algorithms are:
  - Naïve Bayes
  - Gaussian Naïve bayes
  - Bayesian network
  - Bayesian belief network
  - …



Bayesian Algorithms

Clustering Algorithms

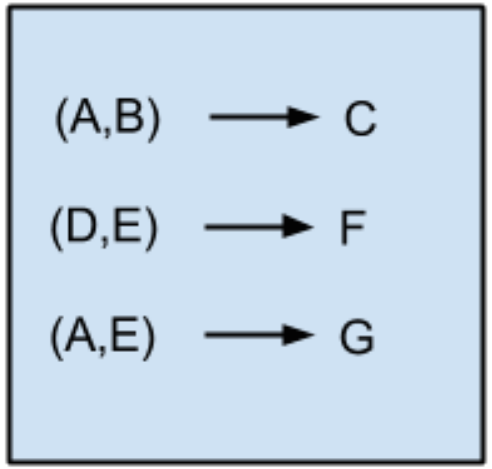
- These algorithms utilize the inherent structures in the data to organize them into various groups.
- Main goal is to find clusters that have high intra similarity and high inter similarity distances.
- Most popular clustering algorithms are:
  - K-Means
  - K-Medoids
  - Expectation Maximization
  - Hierarchical Clustering
  - …



Clustering Algorithms

Association Rule Learning Algorithms

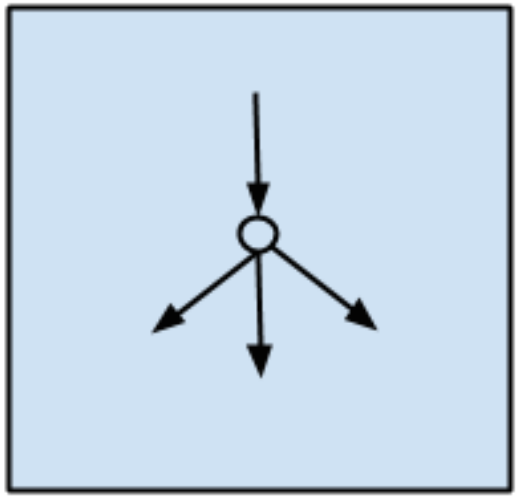
- These methods extract rules that best explain the observed relationships between variables in the data
- Most popular algorithms are:
  - Apriori
  - Eclat
  - FP-growth
  - …



Association Rule Learning Algorithms

Artificial Neural Network Algorithms

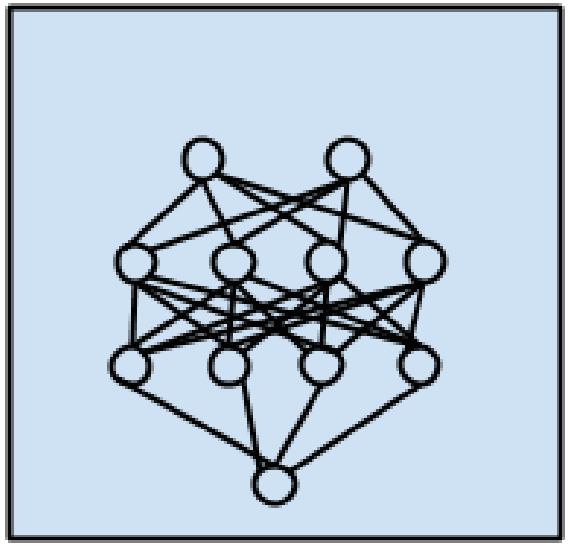
- Models that are inspired by the structure and function of biological neural networks.
- Most popular algorithms are:
  - Perceptron
  - Multilayer perceptron
  - Backpropagation
  - …



Artificial Neural Network Algorithms

Deep Learning Algorithms

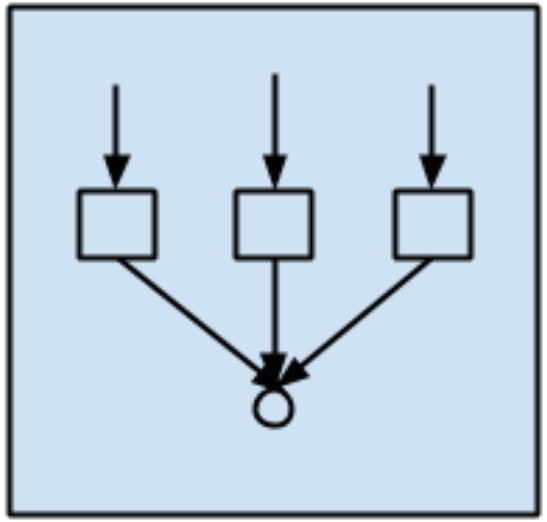
- Update to Artificial Neural Networks
- Main goal is to build a much larger and more complex neural networks.
- Most popular algorithms are:
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Networks (RNNs)
  - Long Short-Term Memory Networks (LSTMs)
  - Deep Belief Networks (DBN)



Deep Learning Algorithms

Ensemble Algorithms

- These are the models composed of multiple weaker models that are independently trained and the predictions are combined to make the overall prediction.
- Some of the popular algorithms are:
  - Boosting
  - Boostrapped Aggregation
  - AdaBoost
  - Gradient Boosting Machines
  - Random Forest
  - …



Ensemble Algorithms

How can we measure the quality of this model?

k-fold Cross-validation

- Resampling procedure to evaluate machine learning models on a given data sample.
- The parameter *k* refers to the number of groups that a given data sample is to be split into.
- If k=10, it is 10-fold cross-validation where the sample data is divided into 10 groups.

k-fold Cross-validation

> Shuffle the dataset (better)

> Split the dataset into k disjoint groups

> For each unique group:

  > Take the group as a hold out or test (validation) data set

  > Take the remaining groups as a training data set

  > Fit a model on the training set and evaluate it on the test set

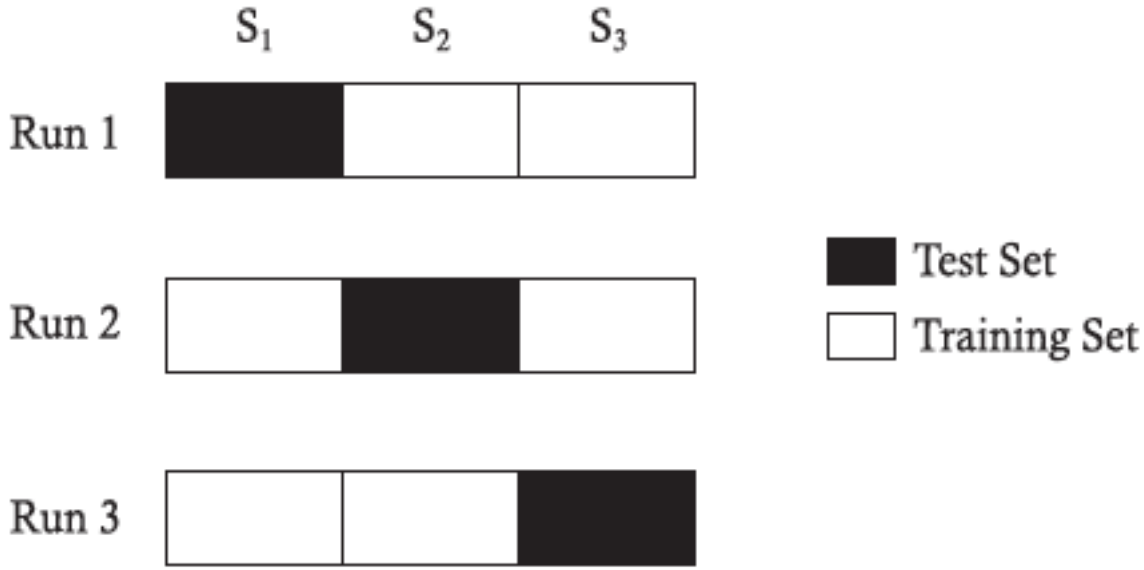
  > Record the evaluation score

> Find the mean of all the sample of model evaluation scores

k-fold Cross-validation

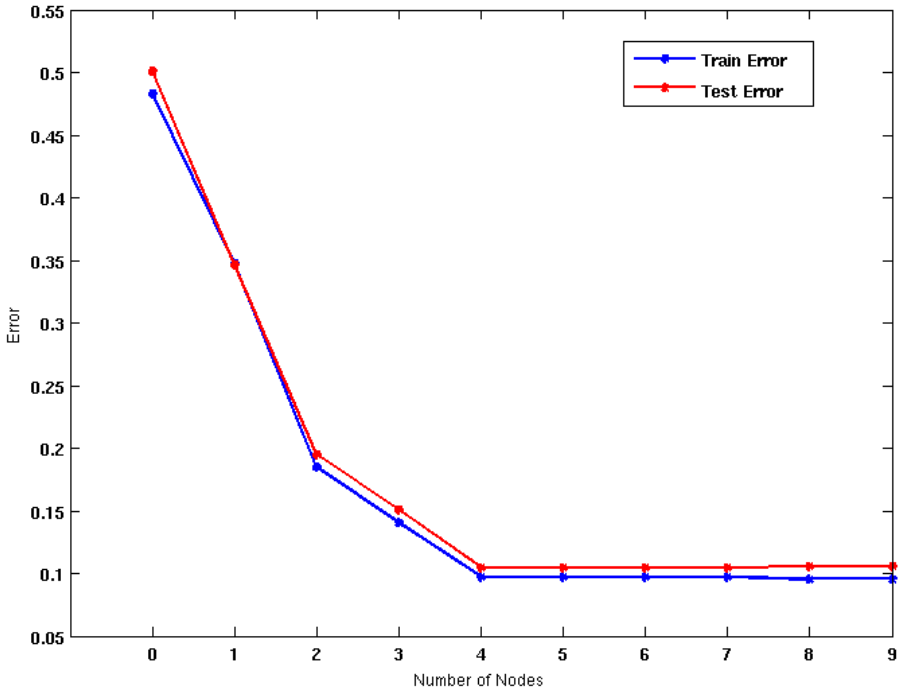
[1, 2, 3, 4, 5, 6]

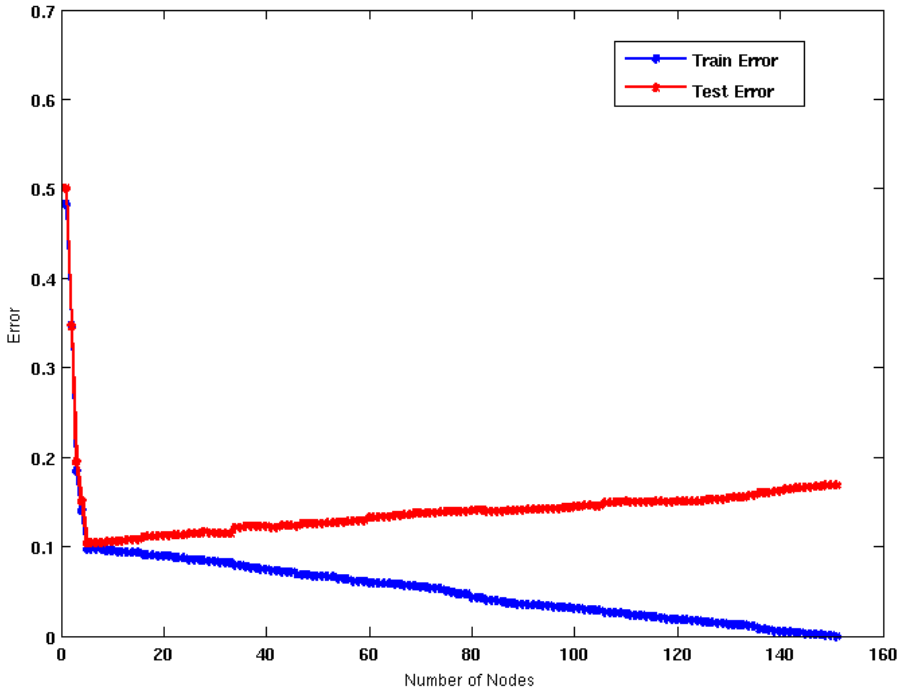
Fold1: [5, 3]
Fold2: [1, 6]
Fold3: [2,4]



Model1: Trained on Fold2 + Fold3, Tested on Fold1
Model2: Trained on Fold1 + Fold3, Tested on Fold2
Model3: Trained on Fold1 + Fold2, Tested on Fold3

Model Overfitting & Underfitting





Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large