Ciro Donalek (Caltech)

Models and tools: overview

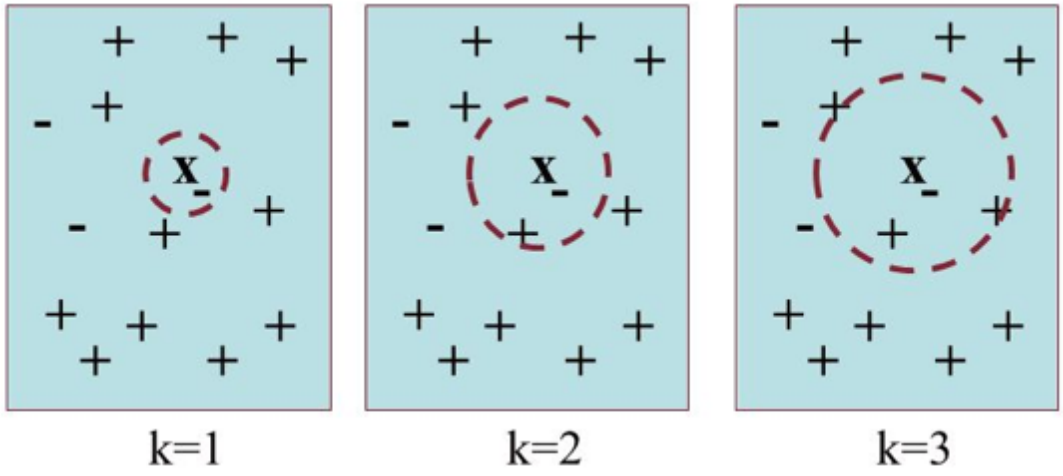
Jet Propulsion Laboratory
INSTITUTE FOR SPACE STUDIES
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

Summary

- Overview
  - SVM, kNN, DT, Bayesian Networks
- Combining classifiers
- Packages and Useful Resources

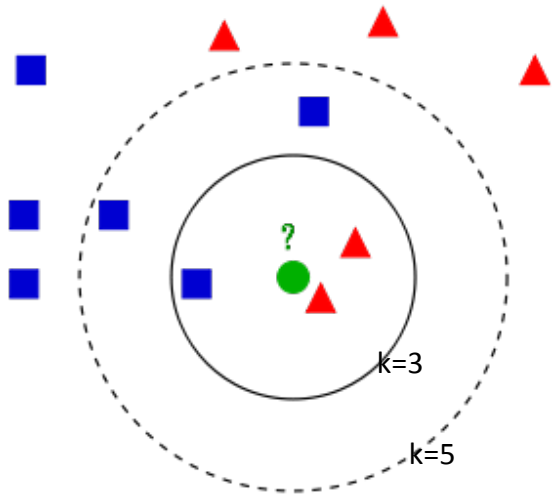
K-Nearest-Neighbors

- Simple, often used as benchmark model.
- Output: class membership.
- Each new object is assigned to the class most common among its k neighbors.
- Lazy learner.



K-Nearest-Neighbors

- Best choice of k depends upon the data set:
  - large k reduce the effect of the noise but make boundaries between classes less distinct
- Accuracy degrades in presence of irrelevant features.



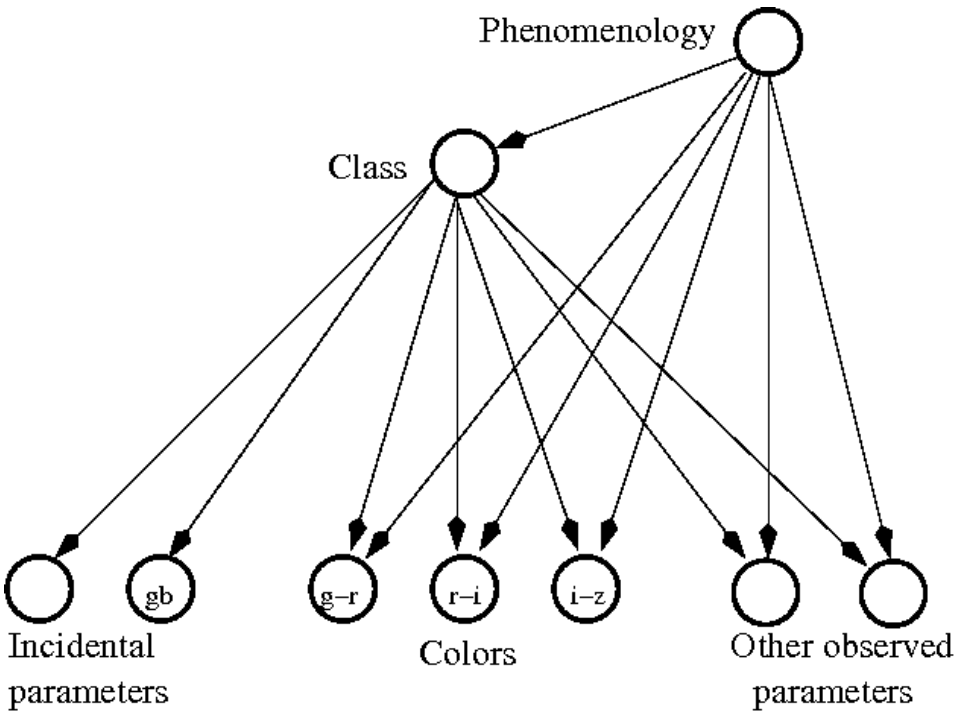
Green circle: test sample
Class 1: blue squares
Class 2: red triangles
If *k = 3* it is assigned to Class 2
If *k = 5* it is assigned to Class 1

Bayesian Networks

- It is a probabilistic graphical model represented through directed acyclic graphs (DAG), whose nodes represent variables, and the missing arcs represent conditional independence assumptions.
- Can deal with missing data
- Probabilistic Classification

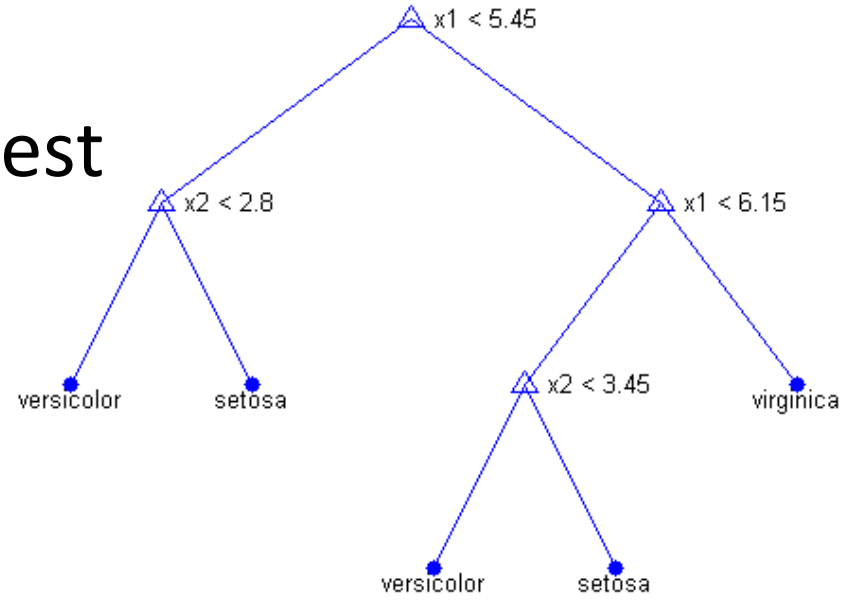
Algorithm

1. Obtain data
2. Choose topology
3. Generate priors and probabilities for each class
4. Run each new event through the network
5. Classify the new objects
6. Feed new data back in to refine the priors



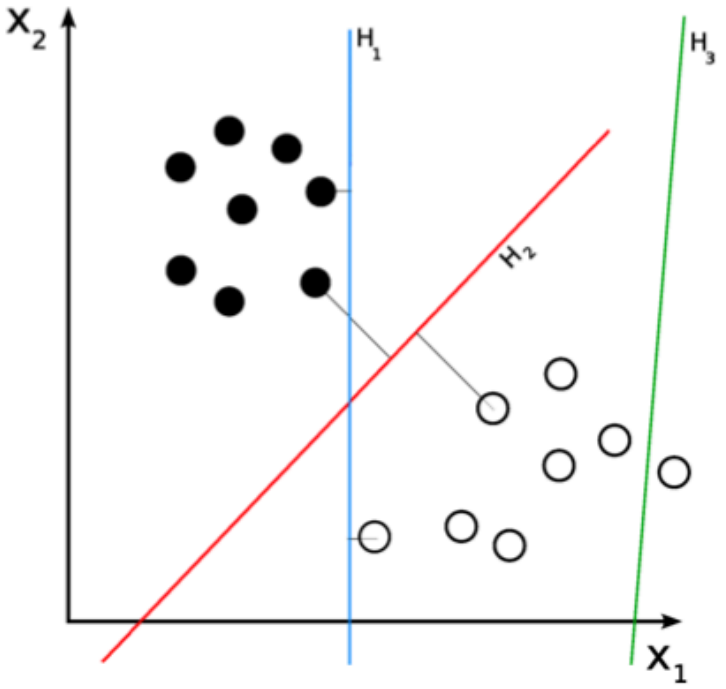
Decision Trees

- Supervised classification method.
- A decision tree is a set of simple rules, such as “if the feature 1 is less than x and feature 2 is greater than y, classify the specimen as AGN”.
- Non-parametric: they do not require any assumptions about the distribution of the variables in each class.
- Internal nodes denotes test on the attributes.
- Leaves represent the class labels.



Support Vector Machines

- Support Vector Machines (SVM) are a group of supervised learning methods that can be applied to classification or regression.
- For any particular set of two-class objects, an SVM finds the unique hyperplane having the maximum margin.



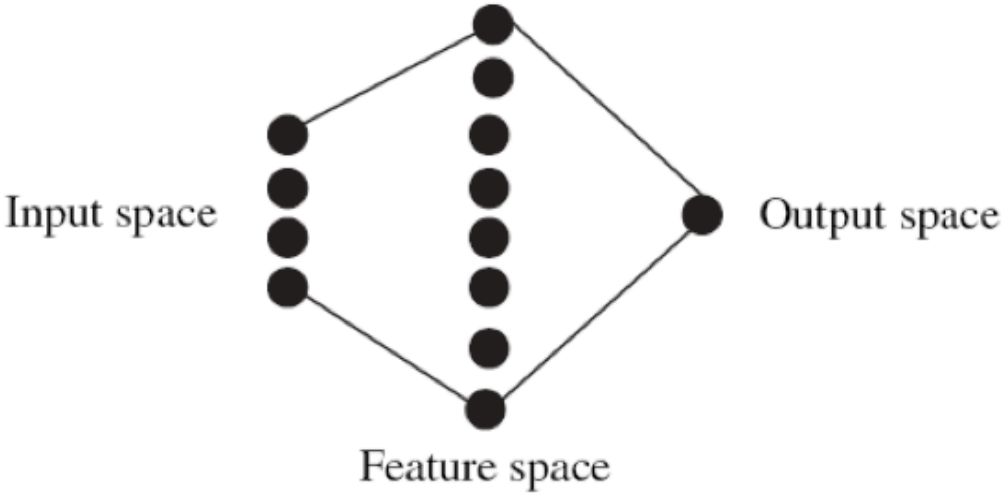
H3 (green) doesn't separate the 2 classes.

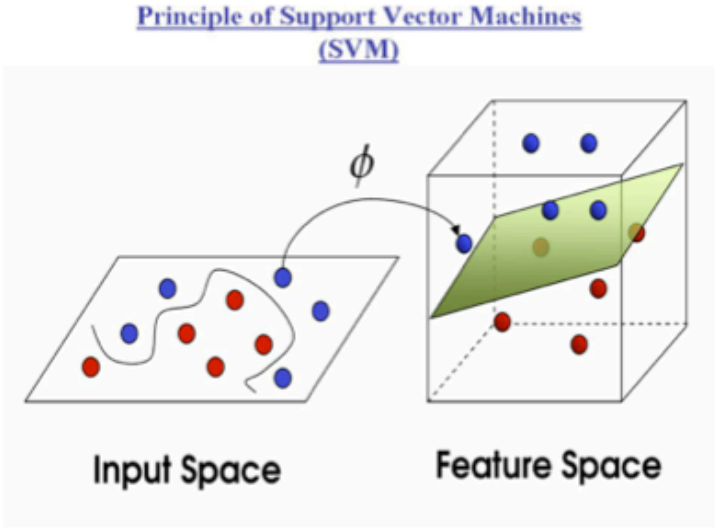
H1 (blue) does, with a small margin.

H2 (red) does with the maximum margin.

SVM: non linear classification

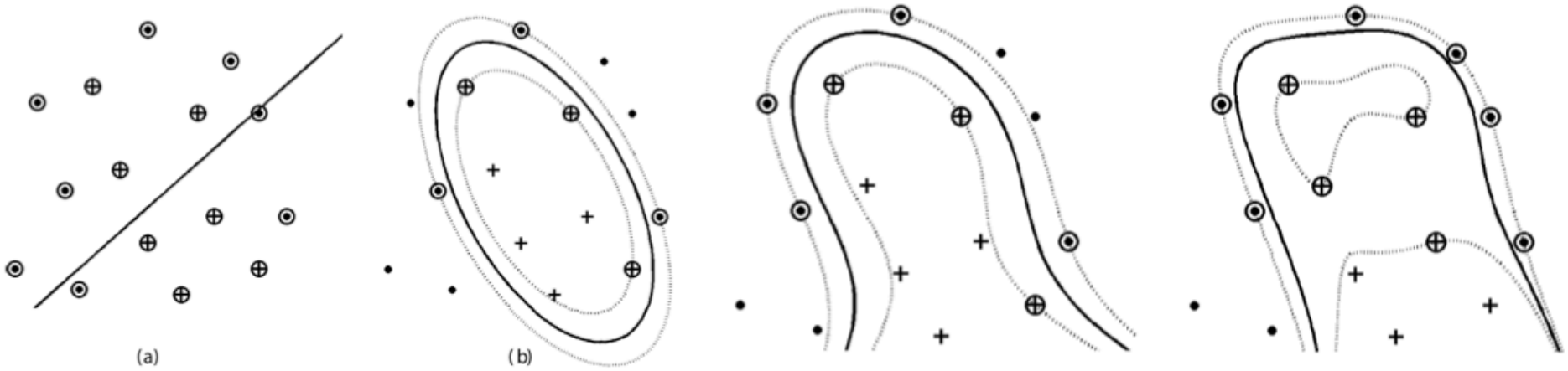
- SVM can be used to separate classes that cannot be separated with a linear classifier.
- Training vectors are mapped into an higher dimensional space using non linear function.
- The feature space is a high dimensional space in which to class can be linearly separated.





SVM: kernel functions

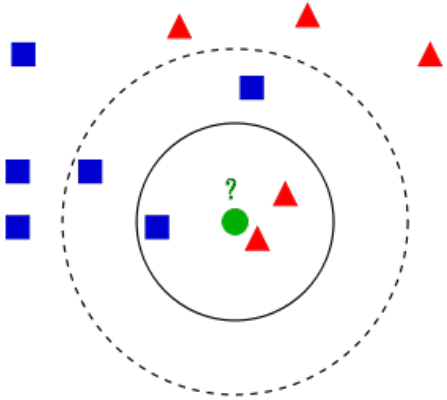
- The mapping into a higher dimensional space is done using the kernel functions.

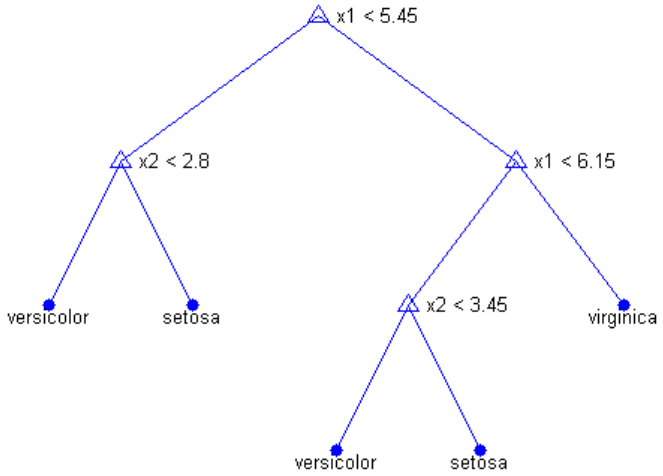


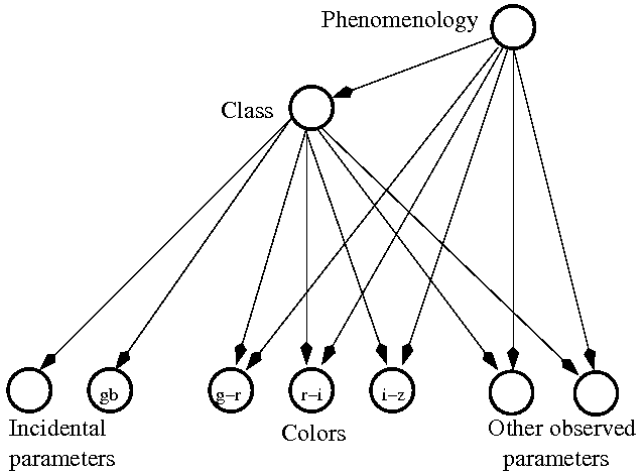
- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$.
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d$, $\gamma > 0$.
- radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)$.

Combining Models

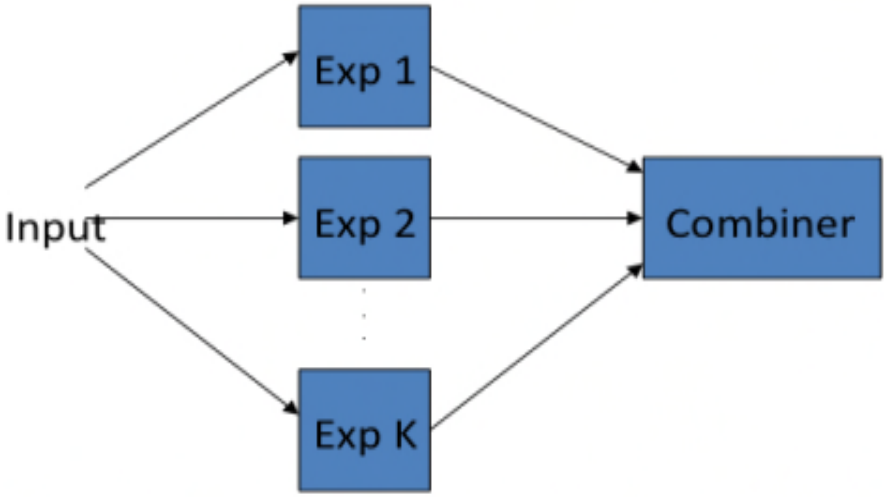
- It is often found that improved performance can be obtained combining model together.
  - individual classifiers may be optimized and trained differently;
  - some classifiers could work better than others in recognizing some classes when certain input attributes are present;
  - some can deal with missing data while some others not.



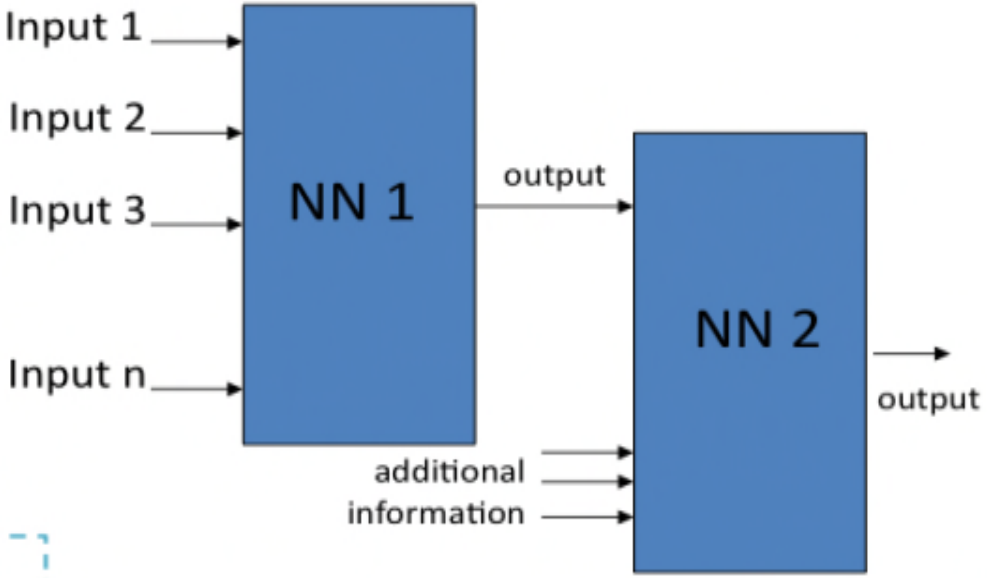




Committee of Machines



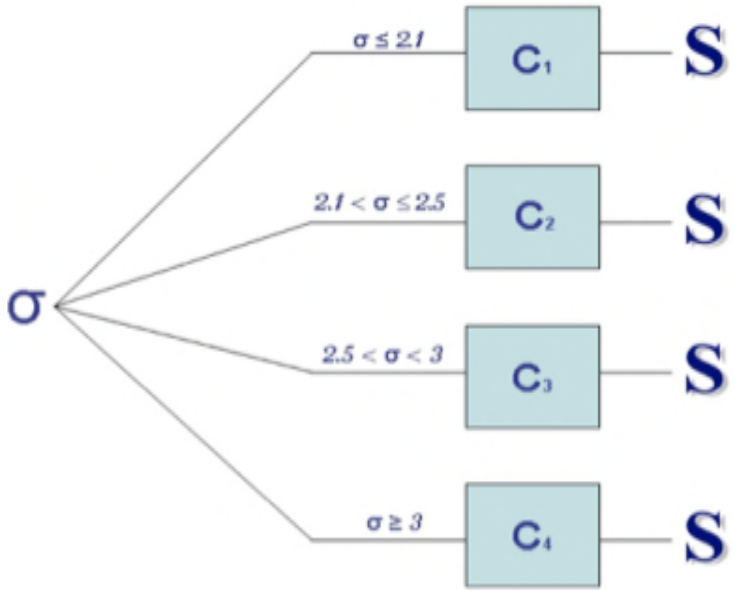
Committee Machines: combination of experts that "vote" together on a given example.

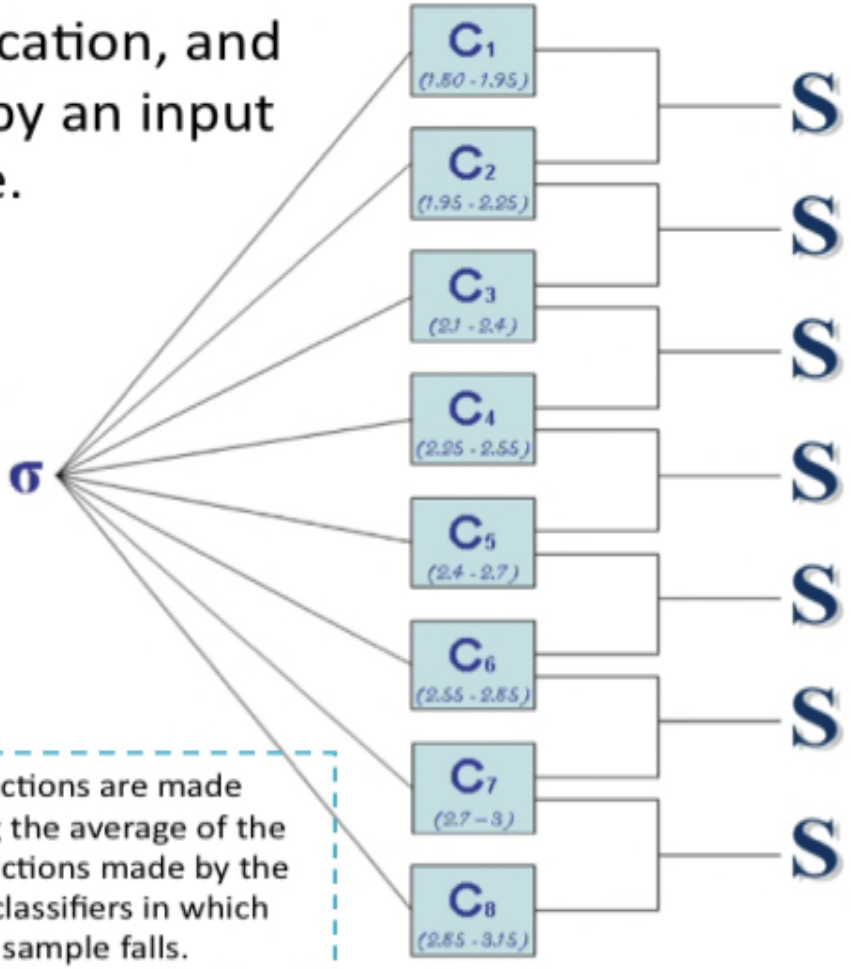


Two-level Network.

A priori Knowledge

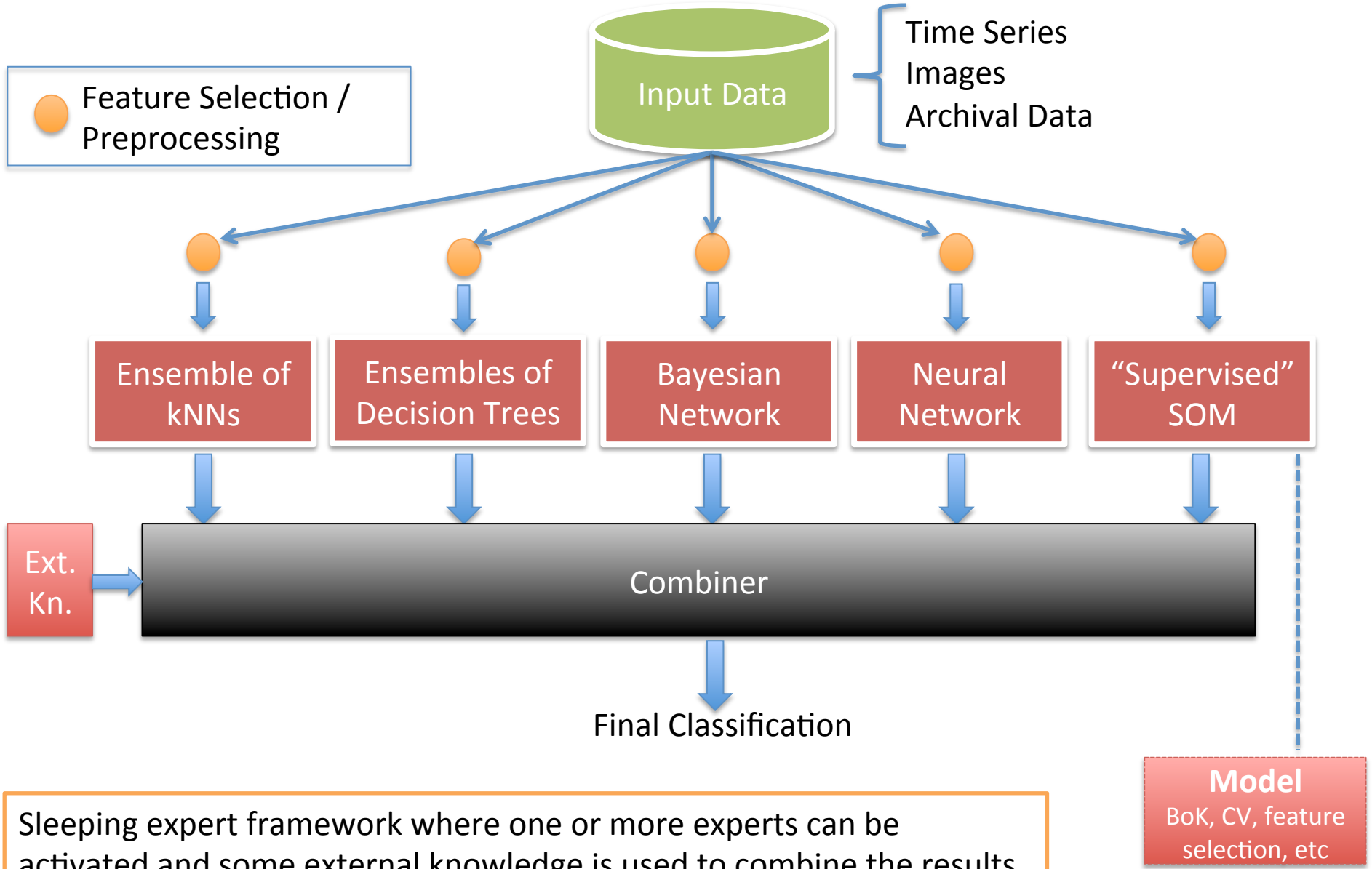
An alternative of model combinations is to select one of the models to make the classification, and let the choice of the model be driven by an input parameter or by an a priori knowledge.





Predictions are made using the average of the predictions made by the two classifiers in which each sample falls.

Sleeping Experts



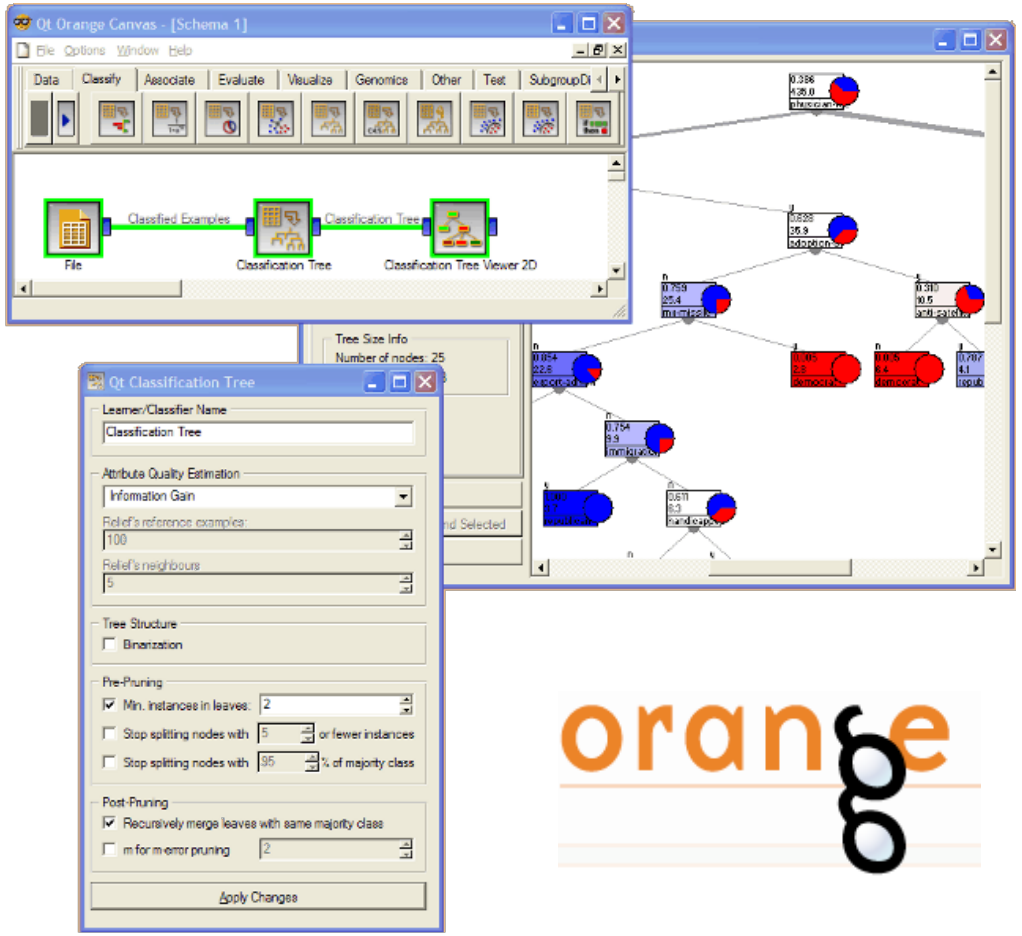
Sleeping expert framework where one or more experts can be activated and some external knowledge is used to combine the results.

DM Tools

- Tools
  - standalone codes
  - DAME
  - Weka
  - Orange
  - Rapid Miner
  - …
- Tasks
  - Regression
  - Classification
  - Clustering

Orange

- Open source data analysis, machine learning and data visualization tool.
- Many classification and clustering models implemented
  - DT, SVM, kNN, Random Forest, Naive Bayes, KMeans, SOM...
- Support for data stored in database (ver. 3.0+).
- Python scripting.

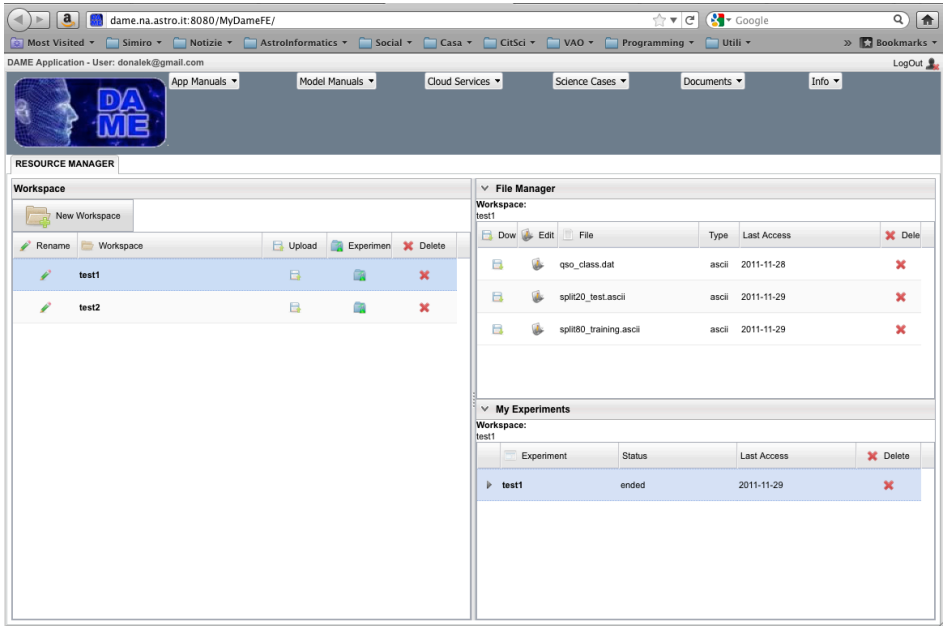


orange

University of Lubjiana - http://orange.biolab.si

DAME

- WebApp
- Method implemented: MLP (many versions), KMeans, SOM, SVM. Many more models are beeing added.

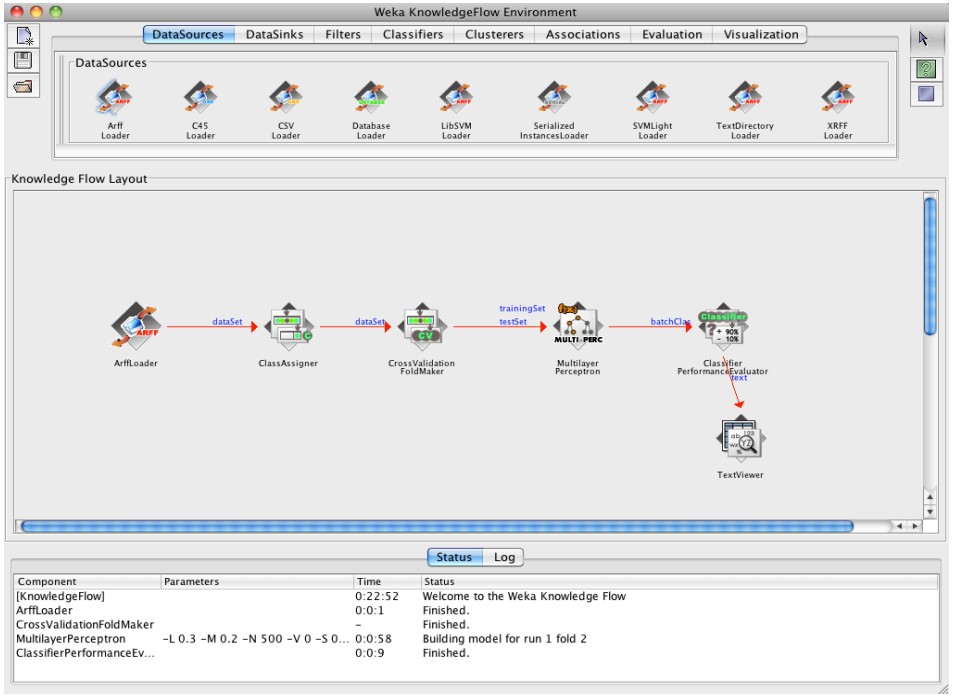


- Scalability: scalable, can deal with large datasets.
- Used in many Astronomical Publication.

DAME is a joint effort among University of Napoli Federico II, INAF and Caltech
http://dame.caltech.edu

Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- The algorithms can either be applied directly to a dataset or called from your own Java code.
- Tools for data pre-processing, classification, regression, clustering, association rules, and visualization.



University of Wakaito - http://www.cs.waikato.ac.nz/~ml/weka/

Other packages

- DM toolboxes and libraries available for any programming language: Matlab, R, C++, Java, Python...

- Useful resources: http://bigdata.astro.caltech.edu/Resources.html

Summary

- Overview of other classification models
- How to combine classifiers
- Useful resources