Notes on Classification

CS4881 Artificial Intelligence
Jay Urbain
Credits:
Nazli Goharian and David Grossman, IIT
Machine Learning, Tom Mitchell
AIMA, Russell and Norvig

Classification

Classification is a Supervised learning.

Learning by example:

- Use training set which has correct answers (class label attribute).
- Create a model by running the algorithm on the training data.
- Test the model. If accuracy is low, regenerate the model, after changing features, reconsidering samples,…
- Identify a class label for the incoming new data

Classification

F_1:color	F_2:shape	F_n	Class Label
Red	Round	5	Apple
Red	Round	3	Cherry
Green	Round	5	Apple
Yellow	Tall	?	banana

Attribute = Feature (F1, F2,…Fn)
Tuple = Sample
Class label is the result of the classification.
Each sample belongs to a pre-defined class label.
Training set is the set of samples used to build a model.
The model is represented as classification rules, decision trees, or mathematical formulae

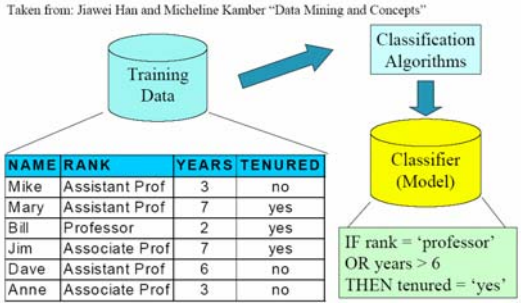
Classification

F_1:color	F_2:shape	F_n	Class Label
Red	Round	5	Apple
Red	Round	3	Cherry
Green	Round	5	Apple
Yellow	Tall	?	banana

Attribute = Feature (F1, F2,…Fn)
Tuple = Sample
Class label is the result of the classification.
Each sample belongs to a pre-defined class label.
Training set is the set of samples used to build a model.
The model is represented as classification rules, decision trees, or mathematical formulae

Classification

Taken from: Jiawei Han and Micheline Kamber "Data Mining and Concepts"



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Measuring Accuracy

Goal: We want to measure the effectiveness of a classification (supervised) algorithm.

- Take the training dataset
- Build model
- Test using the training dataset

This usually leads to a very optimistic result that has little ability to predict the real accuracy of the model.

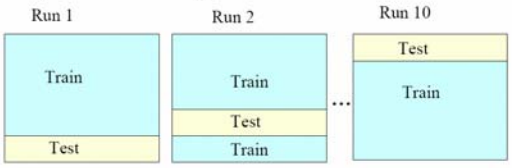
Measuring Accuracy (cont.)

Another approach is to take the training data set, cut it in half and use half on training and half for testing.

This leads to potential errors in estimating the real classification rate because the half we hold for testing may be very different that the half we used for training.

10-fold cross validation

- Take the data set and use the first 90 percent of the data for training and then test on the final ten percent. Then use the next 10 percent for testing, etc.



10-fold cross validation

Each run will result in a particular classification rate.

Ex: If we classified 50/100 of the test records correctly our classification rate for that run is 50%.

Choose the model that generated the highest classification rate. The final classification rate for the model is the average of the ten classification rates.

10-fold cross validation

Doing N-Fold cross validation creates N runs.

Should observe if the result of *N* runs are consistent.

If results not consistent in *N* runs, either model or data might have problems.

Potential Problems

Overfitting: This is when the generated model does not apply to the new incoming data.

- Either too small of training data, not covering many cases.
- Wrong assumptions

Outliers: This is when the generated model is based on outliers.

Summary

Training data is an important factor in building a model in supervised algorithms.

The classification results generated by each of the algorithms (Naïve Bayes, Decision Tree, Neural Networks,…) is not considerably different from each other.

Different classification algorithms can take different time to train and build models.