

# Comparison of Classification Techniques in Bioinformatics

Rashpal Ahluwalia, Ph.D, PE

Sundar Chidambaram, MS

Industrial and Management Systems Engineering

West Virginia University, Morgantown, WV

[rashpal.ahluwalia@mail.wvu.edu](mailto:rashpal.ahluwalia@mail.wvu.edu)

[schidamb@mix.wvu.edu](mailto:schidamb@mix.wvu.edu)

# Outline

---

- The Multivariate Approach
  - Discriminant Function Analysis (DFA)
  - Logistic Regression (LR)
- The Machine Learning Approach
  - Decision Trees (DT)
  - Artificial Neural Networks (ANN)
- Summary

# The Multivariate Approach

## DFA -1

- Goal:
  - Predict group membership from a set of predictors
  - Choice of predictors critical to achieving the goal
- Assumptions:
  - A linear relationship among dependent variable
  - Data are normally distributed
- Principle:
  - Interpretation of patterns of differences among the predictors as a whole – helps in understanding the dimension along which the groups differ

# The Multivariate Approach

## DFA -2

---

- Ideal Data Set:
  - Group sizes are equal
  - Independent Variables (IVs) are continuous and well distributed
- Types of Variables:
  - Predictors (Independent Variables)
  - Equal group sizes (Dependent Variables)

# The Multivariate Approach

## DFA -3

- Research Parameters:
  - Reliability of the prediction of group membership from a set of predictors
  - Number of dimensions along which the groups differ significantly
  - Interpretations of the dimensions along which the groups differ
  - Location of predictors along the discriminant function and correlation of predictors and the discriminant functions
  - Determination of a linear model for the classification of unknown cases

# The Multivariate Approach

## DFA - 4

- Research Parameters (cont/-)
  - The proportion of correct and incorrect classifications using the linear model
  - Degree of relationship between group membership and the set of predictors
  - The most important predictors in predicting group membership
  - Reliable prediction of group membership from a set of predictors after the removal of one or more covariates

# The Multivariate Approach

## DFA - 5

- Limitations:
  - Predicts membership only in naturally occurring groups rather random assignment
  - Assumption of normality
  - Sensitivity to outliers
  - The within cell error matrices must be homogenous to aid in pooling of the variance-covariance matrix
  - Assumes linear relationships among all pairs of dependent variables
  - For unreliable covariates, increased level of Type I and Type II errors

# The Multivariate Approach

## LR - 1

- Goal:
  - Establish a relationship between the outcome and the set of predictors.
  - If a relationship is found, the model is simplified by eliminating some predictors, while maintaining strong prediction
- Assumptions:
  - Dependent variables may be continuous or discrete, dichotomous, or a mix
- Principle:
  - The outcome variable is the probability of having the outcome based on a non-linear function of the best linear combination of predictors

# The Multivariate Approach

## LR - 2

---

- Ideal Datasets:
  - Compare models: Simple – Worst Fitting and Complex – Best Fitting
- Types of Variables:
  - Predictors (Independent Variables)
  - Groups (Dependent Variables)

# The Multivariate Approach

## LR - 3

- Research Parameters:
  - Prediction of outcome from a set of variables
  - Variables that predict and affect the outcome  
[Check if the variable increases, decreases or has no effect on the probability of the outcome]
  - Presence of interactions among the predictor variables
  - Coefficients of the predictors in the LR Model

# The Multivariate Approach

## LR - 4

- Research Parameters (cont/-):
  - Reliability of the model in classifying cases with unknown outcomes
  - Consideration of some predictors as covariates and others as independent variables
  - Strength of association between outcome and the set of predictors in the chosen model

# The Multivariate Approach

## LR - 5

- Limitations:
  - Outcome will always have to be discrete (a continuous variable can be converted into a discrete one)
  - Multivariate normality and linearity among the predictors are not required but help enhance power
  - When there are too few cases – it produces large parameter estimates and standard errors
  - It is sensitive to high correlation among predictor variables, signaled by high standard error for parameter estimates
  - One or more cases may be poorly predicted; a case in one category may show a high probability of being in another
  - Assumes responses for different cases are independent of each other

# The Machine Learning Approach

## DT - 1

- Goal:
  - Approximate discrete valued function that is robust to noisy data and capable of learning from disjunctive expressions
- Assumptions:
  - Target functions are discrete
  - Hypothesis can be learnt from a large set of examples
  - Hypothesis once learnt can approximate outcome for un-observed cases
- Principle:
  - Instance classified starting at the root node – testing the attribute specified by the node – moving down the branch corresponding to the value of the attribute

# The Machine Learning Approach

## DT - 2

- Ideal Datasets:
  - Instances represented by attribute value pairs
  - Target functions having discrete output values
  - Disjunctive descriptions required
  - Training data containing errors
  - Training data containing missing attribute values
- Types of Variables:
  - Instances, classified from the root to a leaf node
  - Attribute of the instance is represented by each node
  - Values of the attributes correspond to each branch descending from the node

# The Machine Learning Approach

## DT - 3

- Typical DT Algorithms (ID3 and C4.5):
  - Basic algorithm constructs a decision tree (top-down), selecting an attribute that needs to be tested at the root of the tree
  - Each instance of the attribute is evaluated using a statistical test (to determine how well it classifies the training examples)
  - The best attribute is selected and used as a test at the root node of the tree
  - A descendant of the root node is then created for each possible value of the attribute
  - Overfitting the training data – important issue in decision tree learning

# The Machine Learning Approach

## ANN - 1

- Goal:
  - A method for learning and interpretation of complex real world data
- Assumptions:
  - Depends on the type of algorithm used
- Principle (for Back Propagation Algorithm):
  - Algorithm learns from error patterns (gradient descent)
  - Uses error propagation to identify patterns

# The Machine Learning Approach

## ANN - 2

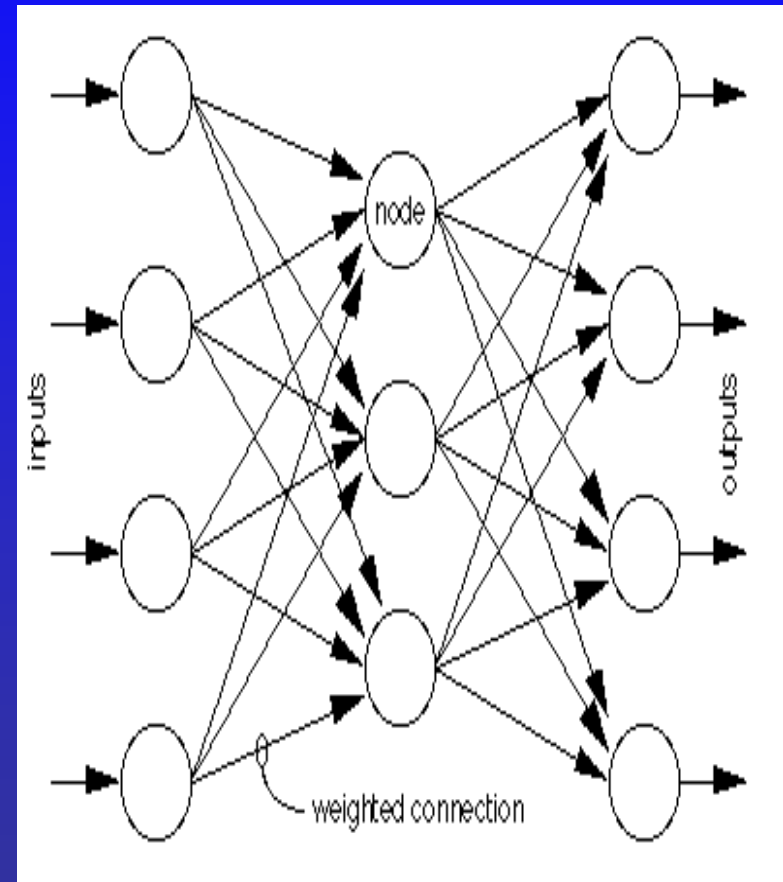
---

- Ideal Datasets:
  - Training samples may have errors
  - Instances represented by many attributes
- Types of Variables:
  - Real valued, discrete valued and vector valued target functions
- Typical ANN Algorithms:
  - Supervised Learning (Back Propagation Algorithm)
  - Unsupervised Learning (Self-Organizing Maps)

# The Machine Learning Approach

## ANN – 3 (BPA)

- BPA Steps
  - Feedforward: Input training pattern
  - Backpropagation of associated error
  - Use gradient descent to reduce the error function



# Summary

	DFA	LR	DT	ANN
Goals	Predict - group membership - dimension along which the groups differ	Predict discrete outcome from a set of variables that [continuous, discrete or dichotomous]	Method of approximating discrete valued functions – robust to noisy data	A method for learning and interpretation of complex real world data
Assumptions	Normally distributed, Linear relationship between DVs	No assumptions on the distribution of predictor variables	Target function – discrete valued An inductive learning method	Assumptions depend on the type of the algorithm used
Types of variables	Predictors – IVs Groups – DVs	Predictors – IVs Groups – DVs	Instances Attributes	Real, discrete and vector valued fn.
Principle	Interpretation of pattern differences among predictors along which the dimensions differ	Output is the probability of having the output based on a non-linear function of the best linear combination of the predictors	Instance classified starting at the root node – testing the attribute specified by the node – moving down the branch corresponding to the value of the attribute	<u>Prin. of BP algorithm</u> - Uses error propagation to identify patterns - Algorithm learns from error patterns (gradient descent)
Ideal Dataset	Group sizes are equal and IVs are continuous and well distributed	Compare Models Simple-Worst fitting Complex-Best fitting	Instances rep. by attribute values Target functions - discrete output	- Instances rep. by many attributes - Training samples may have errors



# Thank You!