Classification Techniques

Đỗ Thanh Hà
Bài giảng của DSLab
Viện nghiên cứu cao cấp về Toán (VIASM)

Vietnam Institute for
Advanced Study in Mathematics

Content

- Introduction
- Base Classifiers
  - Decision Tree based Methods
  - Bayesian Classification
  - Neural Networks - Deep Learning for Computer Vision
- Ensemble Classifiers
  - Bagging
  - Random Forests
- Practical problems

Classification: definition

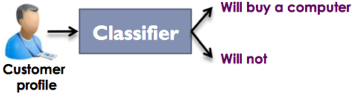
- Given a collection of records (training set)
  - Each record is by characterized by a tuple $(\mathbf{x}, y)$, where $\mathbf{x}$ is the attribute set and y is the class label
- Task: Learn a model that maps each attribute set $\mathbf{x}$ into one of the predefined class labels y
- Example:

Task	Attribute set, $\mathbf{x}$	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam

Vietnam Institute for Advanced Study in Mathematics

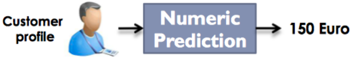
Classification vs. Prediction

- Classification
  - Predicts categorical class labels (discrete or nominal)
  - Use labels of the training data to classify new data
- Example



- Classifier is contsructed to predict **categorical labels** such as *safe* or *risky* for a loan application data

- Prediction
  - Models continuous-valued functions, i.e., predicts unknown or missing values
- Example



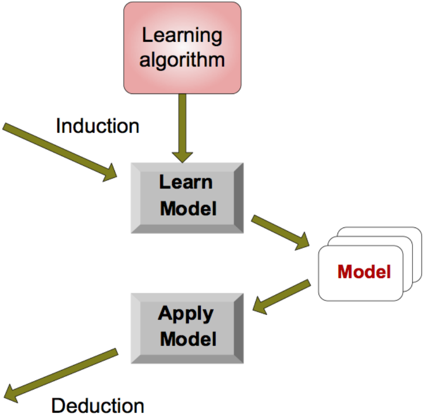
  - Predict how much a given costumer will spend during a sale
  - Unlike classification, it provides ordered values
  - **Regression** analysis is used for prediction

Vietnam Institute for Advanced Study in Mathematics

General Approach for Building 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

Decision Tree

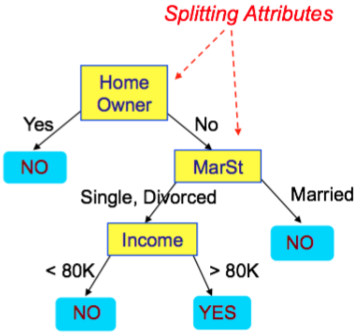
Based on the book *Introduction to Data Mining (2rd Edition)* of P.N Tan *et al*

Example: A Decision Tree

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



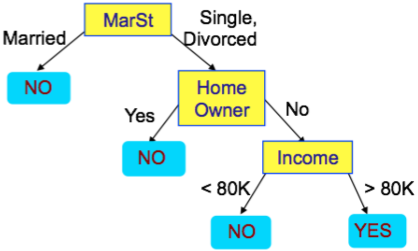
Model: Decision Tree

Vietnam Institute for Advanced Study in Mathematics

Example: A Decision Tree (con't)

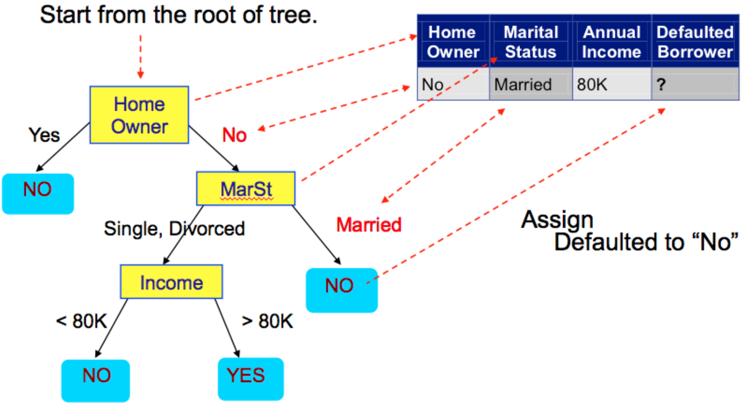
categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



- There could be more than one tree that fits the same data

Example: Apply Model to Test Data



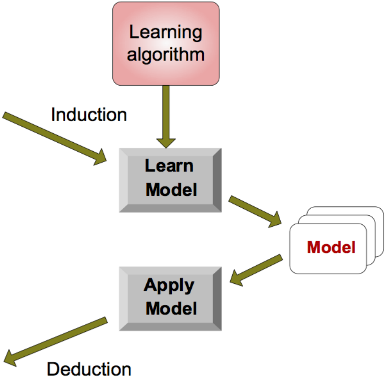
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Vietnam Institute for Advanced Study in Mathematics

Decision Tree Classification Task

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

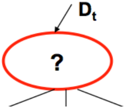
Learning Algorithm = Tree Induction Algorithm

For example: Hunt's Algorithm, CART, ID3, C4.5, SLIQ, SPRINT

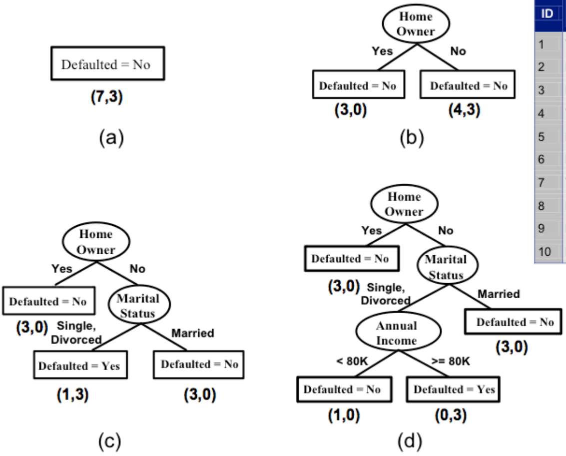
Decision Tree Induction: Hunt's Algorithm

- Let D_t be the set of training records that reach a note t
- General Recursive Procedure:
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t is an empty set, then t is a leaf node labeled by the default class y_d
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset
- Stopping condition: All the records in the subset belong to the same class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Hunt's Algorithm (con't)



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Vietnam Institute for Advanced Study in Mathematics

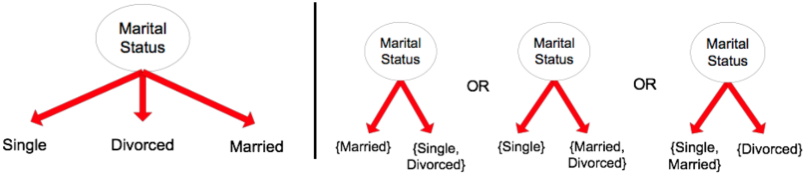
Design Issues of Decision Tree Induction

- How should training records be split?
  - Method for specifying test condition
    - Depending on attribute types: binary, nominal, ordinal, continuous
    - Depending on number of ways to split: 2-way split, multi-way split
  - Measure for evaluating the goodness of a test condition
- How should the splitting procedure stop?
  - Stop splitting if all the records belong to the same class or have identical attribute values
  - Early termination

Vietnam Institute for Advanced Study in Mathematics

Test Condition for Nominal Attributes

- Multi-way split
  - Use as many partitions as distinct values
- Binary split
  - Divides values into two subsets
  - Preserve order property among attribute values



Vietnam Institute for Advanced Study in Mathematics

Splitting Based on Continuous Attributes

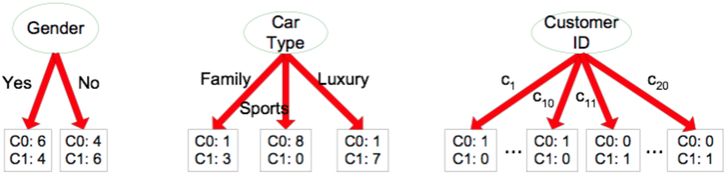
- Discretization to form an ordinal categorical attribute. Ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering
  - Static – discretize once at the beginning
  - Dynamic – repeat at each node
- Binary Decision: $(A < v)$ or $(A \geq v)$
  - consider all possible splits and finds the best cut
  - can be more compute intensive

Vietnam Institute for Advanced Study in Mathematics

How to determine the Best Split

- Before Splitting:
  - 10 records of class 0
  - 10 records of class 1
- Whis test condition is the best?

Customer Id	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1



VIASM
hematics

How to determine the Best Split

- Greedy approach:
  - Nodes with purer class distribution are preferred
- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
High degree of impurity	Low level of impurity

Vietnam Institute for Advanced Study in Mathematics

Measures of Node Impurity

- Gini Index

$$\text{GINI}(t) = 1 - \sum_j [p(j|t)]^2$$

- Entropy

$$\text{Entropy(t)} = -\sum p(j|t) \log p(j|t)$$

- Misclassification error

$$\text{Error}(t) = 1 - \max P(i|t)$$

Finding the Best Split

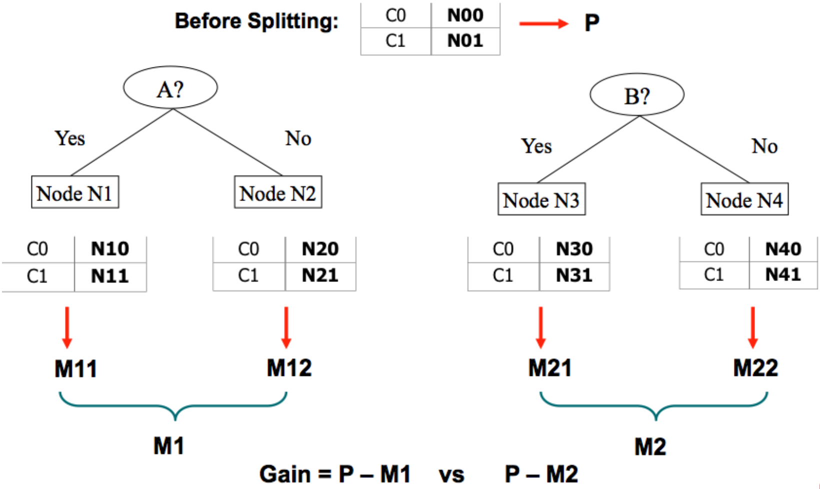
- Compute impurity measure (**P**) before splitting
- Compute impurity measure (**M**) after splitting
  - Compute impurity measure of each child node
  - **M** is the weighted impurity of children
- Choose the attribute test condition that produces the highest gain

$$\text{Gain} = \mathbf{P} - \mathbf{M}$$

  or equivalently, lowest impurity measure after splitting (**M**)

Vietnam Institute for Advanced Study in Mathematics

Finding the Best Split



Vietnam Institute for Advanced Study in Mathematics

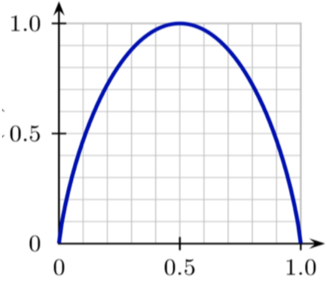
Measure of Impurity: Entropy

- Entropy at a given node t:

  $$\text{Entropy(t)} = -\sum p(j|t) \log p(j|t)$$

  NOTE: $p(j|t)$ is the relative frequency of class j at note t

- The higher the entropy, the less confident we are in the outcome



Vietnam Institute for Advanced Study in Mathematics

Computing Entropy of a Single Node

$$\text{Entropy(t)} = -\sum p(j|t) \log p(j|t)$$

C1	**0**
C2	**6**

$\mathbf{P}(\mathrm{C1}) = \frac{0}{6} = 0$; $\mathbf{P}(\mathrm{C2}) = \frac{6}{6} = 1$
Entropy = -0 log0 -1 log1 = -0 -0 = 0

C1	**1**
C2	**5**

$\mathbf{P}(\mathrm{C1}) = \frac{1}{6}$; $\mathbf{P}(\mathrm{C2}) = \frac{5}{6}$
Entropy $= -\frac{1}{6} \log_2 \left(\frac{1}{6}\right) - \frac{5}{6} \log_2\left(\frac{5}{6}\right) = 0.65$

C1	**2**
C2	**4**

$\mathbf{P}(\mathrm{C1}) = \frac{2}{6}$; $\mathbf{P}(\mathrm{C2}) = \frac{4}{6}$
Entropy $= -\frac{2}{6} \log_2 \left(\frac{2}{6}\right) - \frac{4}{6} \log_2\left(\frac{4}{6}\right) = 0.92$

VIASM
Vietnam Institute for Advanced Study in Mathematics

Computing Information Gain After Splitting

- Information Gain:

$$\text{GAIN}_{split} = \text{Entropy}(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i)\right)$$

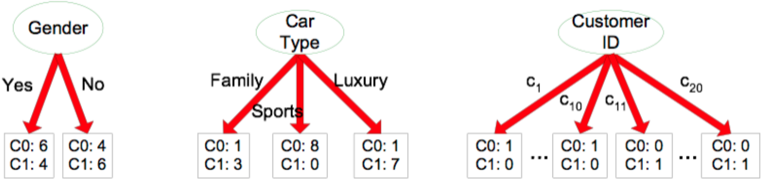
Parent Node, p is split into k partitions; n_i is number of records in partition i

- Choose the split that achieves most reduction (maximizes GAIN)

Vietnam Institute for Advanced Study in Mathematics

Problem with large number of partitions

- Node impurity measures tend to prefer splits that result in large number of partitions, each being small but pure



- Customer ID has highest information gain because entropy for all the children is zero

Vietnam Institute for Advanced Study in Mathematics

Gain Ratio

- Gain Ratio:
  $$GainRATIO_{split} = \frac{GAIN_{split}}{splitINFO} \quad \Bigg\| \quad splitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$
  Parent Node, p is split into k partitions; n_i is number of records in partition i
- Adjusts Information Gain by the entropy of the partitioning (SplitINFO)
  - Higher entropy partitioning (large number of small partitions) is penalized
  - SplitINFO $= 1.52$ (*Left*), 0.72 (*Middle*), and 0.97 (*Right*)

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3

	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10

Decision Tree Based Classification

- Advantages:
  - Inexpensive to construct
  - Extremely fast at classifying unknown records
  - Easy to interpret for small-sized trees
  - Robust to noise (especially when methods to avoid overfitting are employed)
  - Can easily handle redundant or irrelevant attributes (unless the attributes are interacting)
- Disadvantages:
  - Space of possible decision trees is exponentially large. Greedy approaches are often unable to find the best tree.
  - Does not take into account interactions between attributes
  - Each decision boundary involves only a single attribute

Bayesian Classifiers

Based on the book *Introduction to Data Mining (2rd Edition)* of P.N Tan *et al*

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

- Bayes theorem: "X" is feature, "Y" is class

$$\underbrace{\text{P(Class} \mid \text{Feature)}}_{\textbf{Posterior}} = \frac{\overbrace{\text{P(Feature} \mid \text{Class)}}^{\textbf{Likelihood}}\,\overbrace{\text{P(Class)}}^{\textbf{Prior}}}{\underbrace{\text{P(Feature)}}_{\textbf{Evidence}}}$$

Vietnam Institute for Advanced Study in Mathematics

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Vietnam Institute for Advanced Study in Mathematics

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables
- Given a record with attributes $(X_1, X_2, ..., X_d)$
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes $P(Y|X_1, X_2, ..., X_d)$
- Can we estimate $P(Y|X_1, X_2, ..., X_d)$ directly from data?
- For example:
  - Given $X = (Refund = No, Divorced, Income = 120K)$
  - Estimate $P(Evade = Yes|X)$ and $P(Evade = No|X)$?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Vietnam Institute for Advanced Study in Mathematics

Using Bayes Theorem for Classification

- Approach:
  - compute posterior probability $P(Y|X_1X_2....X_d)$ using the Bayes theorem

$$\boxed{P(Y|X_1X_2....X_d) = \frac{P(X_1X_2....X_d|Y)P(Y)}{P(X_1X_2....X_d)}}$$

  - *Maximum a-posteriori*: Choose Y that maximizes

$$P(Y|X_1X_2....X_d)$$

  - Equivalent to choosing value of Y that maximizes

$$P(X_1X_2....X_d|Y)P(Y)$$

- How to estimate $P(X_1X_2...X_d|Y)$?

Vietnam Institute for Advanced Study in Mathematics

Example Data

- Given $X = (Refund = No, Divorced, Income = 120K)$
- Using Bayes Theorem:

$$P(Yes|X) = \frac{P(X|Yes)P(Yes)}{P(X)}$$

$$P(No|Y) = \frac{P(X|No)P(No)}{P(X)}$$

- How to estimate $P(X|Yes)$ and $P(X|No)$?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

VIASM
Vietnam Institute for Advanced Study in Mathematics

Naïve Bayes Classifier

Assume independence among attributes X_i when class is given:

- $P(X_1 X_2 ... X_d | Y_j) = P(X_1|Y_j)P(X_2|Y_j)...P(X_d|Y_j)$
- Now we can estimate $P(X_i|Y_j)$ for all X_i and Y_j combinations from the training data
- New point is classified to Y_j if $P(Y_j)\prod P(X_i|Y_j)$ is maximal

Conditional Independence

- **X** and **Y** are conditionally independent given **Z** if $P(\mathbf{X}|\mathbf{YZ}) = P(\mathbf{X}|\mathbf{Z})$
- Example: Arm length and reading skills
  - Young child has shorter arm length and limited reading skills, compared to adults
  - If age is fixed, no apparent relationship between arm length and reading skills
  - Arm length and reading skills are conditionally independent given age

Naïve Bayes on Example Data

Given $X = (Refund = No, Divorced, Income = 120K)$

$$P(X|Yes) = P(Refund = No|Yes) \times P(Divorced|Yes) \times P(Income = 120K|Yes)$$

$$P(X|No) = P(Refund = No|No) \times P(Divorced|No) \times P(Income = 120K|No)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

VIASM
Vietnam Institute for Advanced Study in Mathematics

Naïve Bayes on Example Data

- Class: $P(Y) = N_c/N$
  - e.g., $P(No) = \frac{7}{10}$; $P(Yes) = \frac{3}{10}$
- For categorical attributes: $P(X_i|Y_k) = |X_{ik}|/N_{ck}$
  - where $|X_{ik}|$ is number of instances having attribute value X_i and belonging to class Y_k
  - eg.:

$$P(Status = Married|No) = \frac{4}{7}$$

$$P(Refund = Yes|Yes) = 0$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Vietnam Institute for Advanced Study in Mathematics

Estimate Probabilities from Data

For continuous attributes:

- Discretization: Partition the range into bins:
    - Replace continuous value with bin value
        - Attribute changed from continuous to ordinal
- Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

Estimate Probabilities from Data

- Normal distribution

$$P(X_i|Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (X_i, Y_i) pair
- For (Income, Class = No):
  - If Class = No
    - sample mean = 110
    - sample variance = 2975

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Vietnam Institute for Advanced Study in Mathematics

Example of Naïve Bayes Classifier

$$\text{Given } X = (Refund = No, Divorced, Income = 120K)$$

$$\begin{aligned}
P(Refund = Yes|No) &= 3/7\\
P(Refund = No|No) &= 4/7\\
P(Refund = Yes|Yes) &= 0\\
P(Refund = No|Yes) &= 1\\
P(MaritalStatus = Single|No) &= 2/7\\
P(MaritalStatus = Divorced|No) &= 1/7\\
P(MaritalStatus = Married|No) &= 4/7\\
P(MaritalStatus = Single|Yes) &= 2/3\\
P(MaritalStatus = Divorced|Yes) &= 1/3\\
P(MaritalStatus = Married|Yes) &= 0
\end{aligned}$$

For Taxable Income:

$$\begin{aligned}
Ifclass &= No : samplemean = 110; samplevariance = 2975\\
&= Yes : samplemean = 90; samplevariance = 25
\end{aligned}$$

- $$\begin{aligned}
P(X|No) = P(Refund &= No|No)\\
&\times P(Divorced|No)\\
&\times P(Income = 120K|No)\\
&= \frac{4}{7} \times \frac{1}{7} \times 0.0072 = 0.0006
\end{aligned}$$

- $$\begin{aligned}
P(X|Yes) = P(Refund &= No|Yes)\\
&\times P(Divorced|Yes)\\
&\times P(Income = 120K|Yes)\\
&= 1 \times \frac{1}{3} \times \frac{1}{2} \times 10^{-9} = 4 \times 10^{-10}
\end{aligned}$$

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$
Therefore $P(No|X) > P(Yes|X)$

$\longrightarrow$ Class = No

Vietnam Institute for Advanced Study in Mathematics

Example of Naïve Bayes Classifier

Given $X = (Refund = No, Divorced, Income = 120K)$

$$
\begin{aligned}
P(Refund = Yes|No) &= 3/7 \\
P(Refund = No|No) &= 4/7 \\
P(Refund = Yes|Yes) &= 0 \\
P(Refund = No|Yes) &= 1 \\
P(MaritalStatus = Single|No) &= 2/7 \\
P(MaritalStatus = Divorced|No) &= 1/7 \\
P(MaritalStatus = Married|No) &= 4/7 \\
P(MaritalStatus = Single|Yes) &= 2/3 \\
P(MaritalStatus = Divorced|Yes) &= 1/3 \\
P(MaritalStatus = Married|Yes) &= 0
\end{aligned}
$$

For Taxable Income:

$$
\begin{aligned}
If class &= No : sample mean = 110; sample variance = 2975 \\
&= Yes : sample mean = 90; sample variance = 25
\end{aligned}
$$

- $P(Yes) = \frac{3}{10}; P(No) = \frac{7}{10}$
- $P(Yes|Divorced) = \frac{1}{3} \times \frac{3}{10} / P(Divorced)$
- $P(No|Divorced) = \frac{1}{7} \times \frac{7}{10} / P(Divorced)$
- $P(Yes|Refund = No, Divorced) =$ $1 \times \frac{1}{3} \times \frac{3}{10} / P(Divorced, Refund = No)$
- $P(No|Refund = No, Divorced) =$ $\frac{4}{7} \times \frac{1}{7} \times \frac{7}{10} / P(Divorced, Refund = No)$

Vietnam Institute for Advanced Study in Mathematics

Issues with Naïve Bayes Classifier

$$\text{Given } X = (Refund = No, Divorced, Income = 120K)$$

$$
\begin{aligned}
P(Refund = Yes|No) &= 3/7 \\
P(Refund = No|No) &= 4/7 \\
P(Refund = Yes|Yes) &= 0 \\
P(Refund = No|Yes) &= 1 \\
P(MaritalStatus = Single|No) &= 2/7 \\
P(MaritalStatus = Divorced|No) &= 1/7 \\
P(MaritalStatus = Married|No) &= 4/7 \\
P(MaritalStatus = Single|Yes) &= 2/3 \\
P(MaritalStatus = Divorced|Yes) &= 1/3 \\
P(MaritalStatus = Married|Yes) &= 0
\end{aligned}
$$

For Taxable Income:

$$
\begin{aligned}
If class &= No : sample mean = 110; sample variance = 2975 \\
&= Yes : sample mean = 90; sample variance = 25
\end{aligned}
$$

- $P(Yes) = \frac{3}{10}; P(No) = \frac{7}{10}$
- $P(Yes|Married) = 0x\frac{3}{10}/P(Married)$
- $P(No|Married) = \frac{4}{7}x\frac{7}{10}/P(Married)$

Vietnam Institute for Advanced Study in Mathematics

Issues with Naïve Bayes Classifier

$$P(Refund = Yes|No) = 2/6$$
$$P(Refund = No|No) = 4/6$$
$$P(Refund = Yes|Yes) = 0$$
$$P(Refund = No|Yes) = 1$$
$$P(MaritalStatus = Single|No) = 2/6$$
$$P(MaritalStatus = Divorced|No) = 0$$
$$P(MaritalStatus = Married|No) = 4/6$$
$$P(MaritalStatus = Single|Yes) = 2/3$$
$$P(MaritalStatus = Divorced|Yes) = 1/3$$
$$P(MaritalStatus = Married|Yes) = 0/3$$

For Taxable Income:

$$Ifclass = No : samplemean = 91; samplevariance = 685$$
$$= Yes : samplemean = 90; samplevariance = 25$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Given $X = (Refund = Yes, Divorced, 120K)$

$P(X|No) = \frac{2}{6} \times 0 \times 0.0083 = 0$

$P(X|Yes) = 0 \times \frac{1}{3} \times 1.2 \times 10^{-9} = 0$

Cannot be able to classify X as Yes or No!

Issues with Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero
- Need to use other estimates of conditional probabilities than simple fractions
- Probability estimation:
  - Original: $P(A_i|C) = \frac{N_{ic}}{N_c}$
  - Laplace: $P(A_i|C) = \frac{N_{ic}+1}{N_c+c}$
  - m - estimate: $P(A_i|C) = \frac{N_{ic}+mp}{N_c+m}$

c: number of classes; p: prior probability of the class, m: parameter; N_c: number of instances in the class;

N_{ic}: number of instances having attribute value A_i in class c

Vietnam Institute for Advanced Study in Mathematics

Naïve Bayes Classifier (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN) that *provides graphical representation of probabilistic relationships among a set of random variables*

Vietnam Institute for Advanced Study in Mathematics

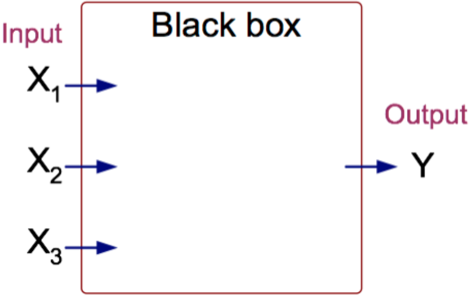
Neural Netwok

Deep Learning for Computer Vision

Based on the book *Introduction to Data Mining (2rd Edition)* of P.N Tan *et al*

Artificial Neural Network (ANN)

X_1	X_2	X_3	Y
1	0	0	-1
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	-1
0	1	0	-1
0	1	1	1
0	0	0	-1

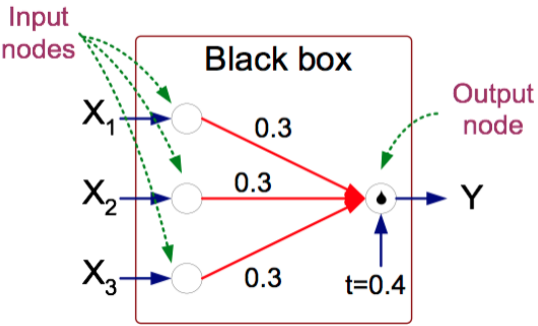


Output Y is 1 if at least two of the three inputs are equal to 1

Vietnam Institute for Advanced Study in Mathematics

Artificial Neural Network (ANN)

X_1	X_2	X_3	Y
1	0	0	-1
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	-1
0	1	0	-1
0	1	1	1
0	0	0	-1

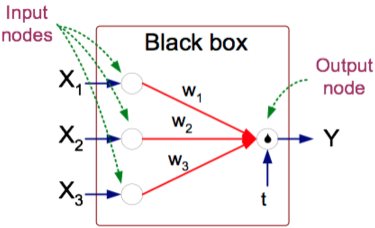


$$Y = sign(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4)$$

$$\text{where } sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Artificial Neural Networks (ANN)

- Model is an assembly of inter-connected nodes and weighted links
- Output node sums up each of its input value according to the weights of its links
- Compare output node against some threshold t

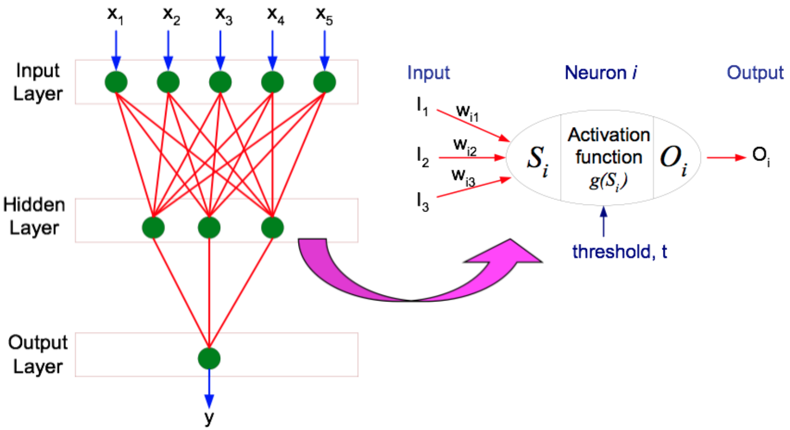


Perceptron Model

$$Y = sign(\sum_{i=1}^{d} w_i X_i - t)$$
$$= sign(\sum_{i=0}^{d} w_i X_i)$$

Vietnam Institute for Advanced Study in Mathematics

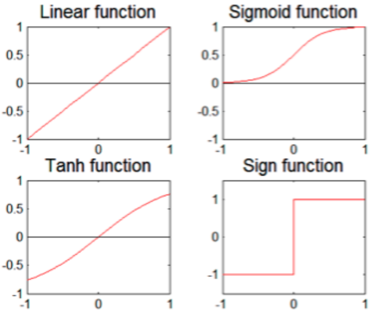
General Structure of ANN



Training ANN means learning the weights of the neurons

Artificial Neural Networks (ANN)

- Various types of neural network topology
  - single - layered network (perceptron) versus multi - layered network
  - Feed - forward versus recurrent network
- Various types of activation functions (f):



Perceptron

- Single layer network
  - Contains only input and output nodes
- Activation function: $f = sign(w \cdot x)$
- Applying model is straightforward

$$Y = sign(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4)$$

$$\text{where } sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

  - $X_1 = 1, X_2 = 0, X_3 = 1 \rightarrow y = sign(0.2) = 1$

Vietnam Institute for Advanced Study in Mathematics

Perceptron Learning Rule

- Initialize the weights $(w_0, w_1, ..., w_d)$
- Repeat: for each training example (x_i, y_i)
  - Compute $f(w, x_i)$
  - Update the weights based on error, in which λ is learning rate

$$w^{(k+1)} = w^{(k)} + \lambda[y_i - f(w^k, x_i)]x_i$$

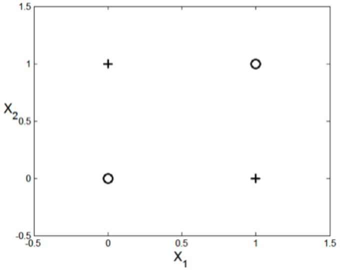
- Until stopping condition is met

Vietnam Institute for Advanced Study in Mathematics

Perceptron Learning Rule

- Since $f(w, x)$ is a linear combination of input variables, decision boundary is linear
- For nonlinearly separable problems, perceptron learning algorithm will fail because no linear hyperplane can separate the data perfectly
- Example of Nonlinearly Separable Data

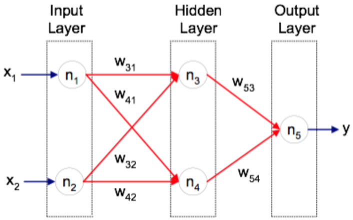
x_1	x_2	y
0	0	-1
1	0	1
0	1	1
1	1	-1

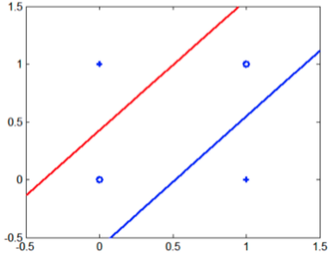


Vietnam Institute for Advanced Study in Mathematics

Multilayer Neural Network

- Hidden layers: intermediary layers between input & output layers
- More general activation functions (sigmoid, linear, etc)
- Multi-layer neural network can solve any type of classification task involving nonlinear decision surfaces





Vietnam Institute for Advanced Study in Mathematics

Learning Multilayer Neural Network

- Can we apply perceptron learning rule to each node, including hidden nodes?
  - Perceptron learning rule computes error term $e = y - f(w, x)$ and updates weights accordingly
    - Problem: how to determine the true value of y for hidden nodes?
  - Approximate error in hidden nodes by error in the output nodes. However, there are problems:
    - Not clear how adjustment in the hidden nodes affect overall error
    - No guarantee of convergence to optimal solution

Gradient Descent for Multilayer NN

- Weight update: $w_j^{(k+1)} = w_j^{(k)} - \lambda \frac{\partial E}{\partial w_j}$
- Error function:

$$E = \frac{1}{2}\sum_{i=1}^{N}(t_i - f(\sum_j w_j x_{ij}))$$

- Activation function f must be differentiable
- For sigmoid function:

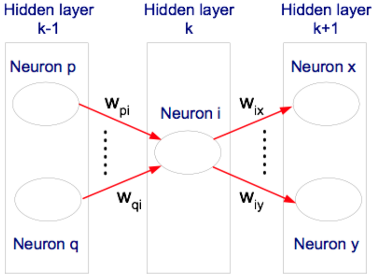
$$w_j^{(k+1)} = w_j^{(k)} - \lambda \sum_i (t_i - o_i) o_i (1 - o_i) x_{ij}$$

- Stochastic gradient descent (update the weight immediately)

Vietnam Institute for Advanced Study in Mathematics

Gradient Descent for Multilayer NN

- For output neurons, weight update formula is the same as before (gradient descent for perceptron)
- For hidden neurons:



$$w_{pi}^{(k+1)} = w_{pi}^{(k)} + \lambda o_i(1 - o_i) \sum_{j \in \Phi_i} \sigma_j w_{ij} x_{pi}$$

  - Output neurons:

$$\sigma_j = o_j(1 - o_j)(t_j - o_j)$$

  - Hidden neurons:

$$\sigma_j = o_j(1 - o_j) \sum_{k \in \Phi_i} \sigma_k w_{jk}$$

Vietnam Institute for Advanced Study in Mathematics

Design Issues in ANN

- Number of nodes in input layer
  - One input node per binary/continuous attribute
  - k or $\log_2 k$ nodes for each categorical attribute with k values
- Number of nodes in output layer
  - One output for binary class problem
  - k or $\log_2 k$ nodes for k-class problem
- Number of nodes in hidden layer
- Initial weights and biases

Vietnam Institute for Advanced Study in Mathematics

Characteristics of ANN

- Multilayer ANN are universal approximators but could suffer from overfitting if the network is too large
- Gradient descent may converge to local minimum
- Model building can be very time consuming, but testing can be very fast
- Can handle redundant attributes because weights are automatically learnt
- Sensitive to noise in training data
- Difficult to handle missing attributes

Vietnam Institute for Advanced Study in Mathematics

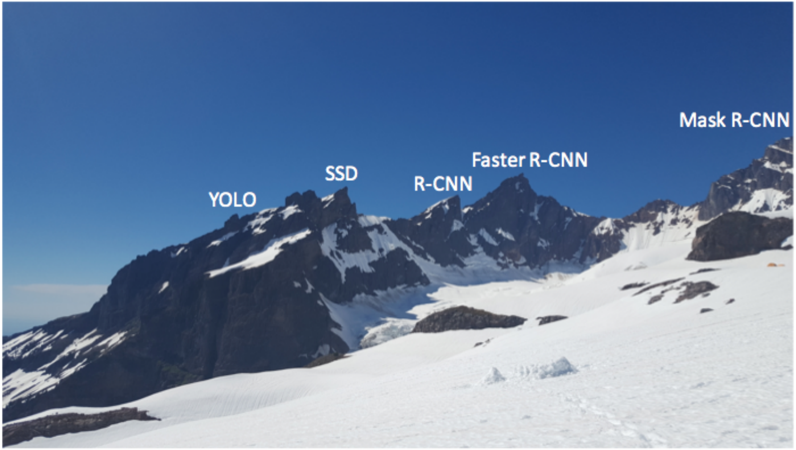
Developments of ANN in Computer Vision

- In computer vision: deep learning learns good representation of the input

R-CNN OverFeat DetectorNet DeepMultibox SPP-net Fast R-CNN MR-CNN SSD YOLO YOLOv2 G-CNN AttractioNet Mask R-CNN R-FCN RPN FPN Faster R-CNN ...

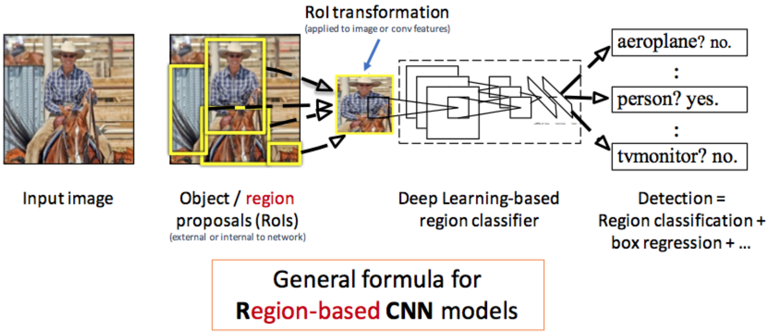
Vietnam Institute for Advanced Study in Mathematics

Landscape of deep learning methods



A random landscape scene on Mt. Baker, just because I like mountains. Photo credit: Ross Girshick

General formula for Region-based Convolutional Neural Netwrok models



Example of Using Deep Learning for Object Understanding

VIASM
Vietnam Institute for
Advanced Study in Mathematics

Ensemble methods

Bagging & Random Forests

Based on the Lecture of Data Science, Harvard University, 2016

Ensemble methods

- A single decision tree does not perform well
- But, it is super fast
- What if we learn multiple trees?

> Make sure that they do not all just learn the same

Bagging (Bootstrap) aggregating

- If we split data in random different ways, decision trees give different results, **high variance**
- Bagging: is a method that result in low variance
- If we had multiple realizations of the data (or multiple samples) we could calculate the predictions multiple times and take the average of the fact that averaging multiple onerous estimations produce less uncertain results

VIASM
Vietnam Institute for
Advanced Study in Mathematics

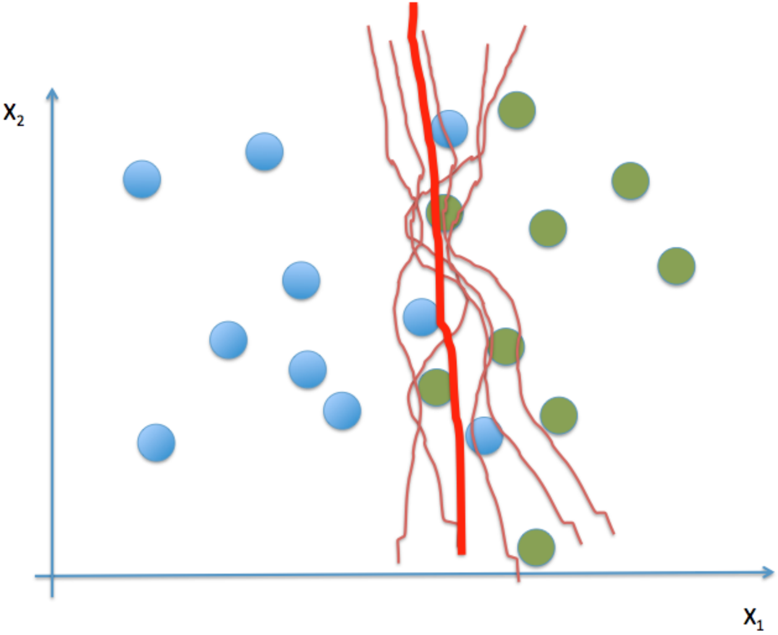
Bagging

- For each sample b, calculate $f^b(x)$:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

- **How?**:
  - Construct B (hundreds) of trees (no pruning)
  - Learn a classifier for each bootstrap sample
  - Average classifiers

Vietnam Institute for Advanced Study in Mathematics


X2
X1

Out-of-Bag Error Estimation

- No cross validation?
- In bootstrapping, **not all observations are used for each bootstrap sample**. On average $1/3$ of them are not used, and they are called as out-of-bag samples (OOB)
- The response of the $i-th$ observation can be predicted using each of the trees in which that observation was OOB. Do this for n observations
- Calculate overall OOB MSE or classification error

Vietnam Institute for Advanced Study in Mathematics

Bagging

- Reduces overfitting (variance)
- Normally uses one type of classifier
- Decision trees are popular
- Easy to parallelize

Vietnam Institute for
Advanced Study in Mathematics

Bagging - issues

- Each tree is identically distributed (i.d.)
  - The expectation of the average of B such trees is the same as the expectation of any one of them
  - the bias of bagged trees is the same as that of the individual trees
- An average of B i.i.d. random variables, each with variance σ^2, has variance: σ^2/B
  - Tree is i.d. and pair correlation ρ is present, thus the variance is $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$
  - As B increases, the second term $\frac{1-\rho}{B}\sigma^2$ disappears but the first term remains
- Suppose: the dataset has one very strong predictor and number of other moderately strong predictors
  - $\longrightarrow$ All bagged trees will select the strong predictor at the top of the tree and therefore all trees will look similar

Vietnam Institute for Advanced Study in Mathematics

Bagging - issues

We want B i.i.d. random variables such as the bias to be the same and variance to be less

- Solution:
  - Consider each only a subset of the predictors at each split?
    - Still get correlated trees unless..
    - Randomly select the subset!


Random Forests

Random Forests

- Building a number of decision trees on bootstrapped training samples each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors.
- if $m = p$, then it is bagging

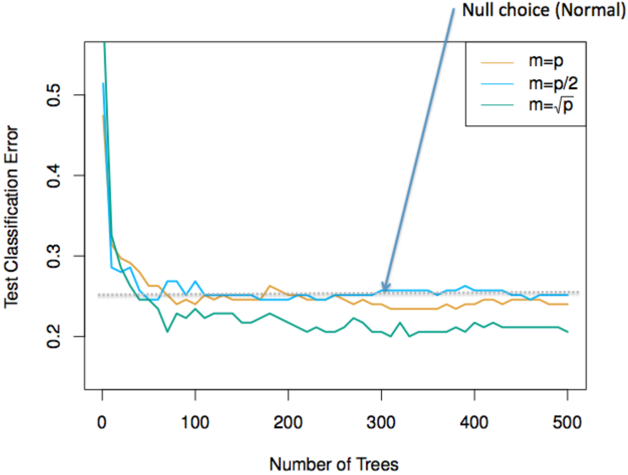
Vietnam Institute for Advanced Study in Mathematics

Random Forests Algorithm

- For $b = 1$ to B
  (a) Draw a bootstrap sample $Z*$ of size N from the training data
  (b) Grow a random-forest tree to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached
    - Select **m** variables at random from the p variables
    - Pick the best variable/split-point among the m
    - Split the node into two daughter nodes
- Output the ensemble of trees
- Make a prediction at a new point x: majority vote

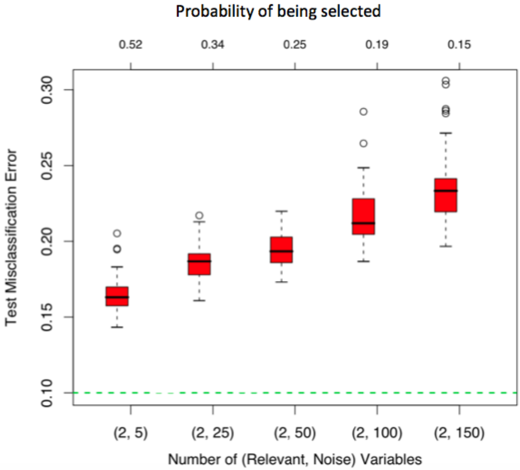
Random Forests: Parameters for Classification

- In theory, the default value for m is $\sqrt{p}$ and the minimum node size is one
- In practice, the parameters depend on the problem



Random Forests Issues

The number of variables is large, but the fraction of relevant variables is small $\longrightarrow$ random forests perform poorly when m small



Vietnam Institute for Advanced Study in Mathematics

Practical Section

Some available tools

- Scikit - learn Data Classification and Regression (Python)
- Apache Mahout Machine Learning Library (Classification)
- ENTOOL for Ensemble Learning and Classification

Thank for your attention!