UNIT-4

* **Supervised Learning**
- → training data is accompanied by labels.
- → classification
- → data is classified based on training set.

Unsupervised learning
- → class labels are unknown.
- → clustering
- → with the help of measurements; classes or clusters are formed.

→ **Classification**
- * predicts class labels (discrete/nominal)

→ **Numeric Prediction:**
- * predicts unknown / missing values.

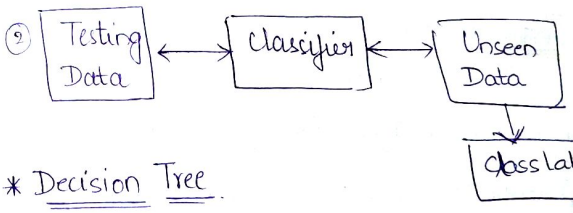
→ **Applications:**
- → Credit/loan approval.
- → Diagnosis
- → fraud detection
- → Webpage categorization.

* Classification is a 2-step process.

① Model construction.
- ↳ a training set formed ~~using~~ by assigning each tuple with a label.

② Model usage:
- ↳ checking for accuracy of the model using test set.

→ with the help of accuracy; we decide whether to use model / not.



* **Decision Tree**

→ Attributes are selected based on heuristic or statistical measure.
- ↳ Information gain.

Entropy = measure of uncertainity

$$= -\sum_{i=1}^{m} P_i \log(P_i)$$

* Attribute Selection measures:

→ Information gain: multi-valued attributes
→ Gain Ratio: unbalanced splits
→ Gini index: multi-valued attributes; large classes.

* Overfitting:

↳ Too many branches, some may reflect anomalies due to noise.

* Approaches to avoid overfitting

① Prepruning: Stop tree construction; if splitting a node; ~~&~~ measures falls below threshold.

② Postpruning: Remove branches from a fully grown tree.

* Enhancements to Decision Tree Induction:

→ continuous valued attributes
→ handle missing attribute values
→ attribute construction.

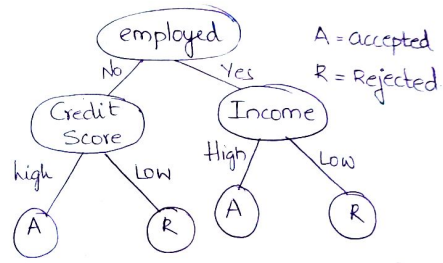
→ AVC list: (Attribute, value, class_label)

* BOAT: Bootstrapped Optimistic Algorithm for Tree Construction.

→ Statistical technique called bootstrapping to create several ~~the~~ smaller samples; each fits in memory.

→ Each subset is used to create a tree.

Ex:



A = accepted
R = Rejected

Information Gain:

$$\frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

Ex:

Age	Completion	Type	Profit
old	Yes	S/W	Down
old	No	S/W	Down
old	No	H/W	Down
mild	Yes	S/W	Down
mild	Yes	H/W	Down
mild	No	H/W	Up
mild	No	S/W	Up

new	Yes	s/w	up
new	No	H/W	up
new	No	s/w	up

→ Information Gain: for Target attribute = profit

$$IG = \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

P = count(down) = 5
N = count(Up) = 5

$$= \frac{-5}{10}\log\left(\frac{5}{10}\right) - \frac{5}{10}\log\left(\frac{5}{10}\right)$$

$$= \frac{-1}{2}(-1) - \frac{1}{2}(-1) = 1$$

→ Entropy for remaining attributes:

$$E(A) = \sum \frac{P_i+N_i}{P+N} I(P_i N_i)$$

① Age:

	down	up
old	3	0
mid	2	2
new	0	3

P = down count
=
N = up count
=

$IG(old) = \frac{-3}{3}\log\left(\frac{3}{3}\right) - \frac{0}{3}\log\left(\frac{3}{3}\right) = 0$

Probability = $\frac{3}{10}$

E(old) = 0

$$IG(mid) = \frac{-2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right)$$

$$= \frac{-1}{2}\log_2(2^{-1}) - \frac{1}{2}\log_2(2^{-1})$$

$$= 1$$

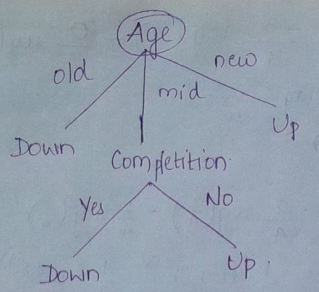
Probability = $\frac{4}{10}$ ⇒ $E(mid) = \frac{2}{5}$

IG(new) = 0
E(new) = 0

E(Age) = E(old) + E(mid) + E(new) = 0.4

Gain = IG − E(A) = 0.6

Gain(Age) = 0.6 → root node ①
Gain(Type) = 0 ③
Gain(completion) = 0.124 ②



∴ There is no use of type attribute.

* Bayesian Classification:

→ Bayes Theorem:

Total probability theorem: $P(B) = \sum_{i=1}^{M} P(B|A_i)\, P(A_i)$

Bayes formula: $P(H|X) = \frac{P(X|H)\, P(H)}{P(X)}$

$= P(X|H) \times P(H) / P(X)$

$P(X)$ = probability that sample data is observed

$P(H)$ = prior probability

$P(X|H)$ = probability of observing sample X, given hypothesis holds.

→ maximum posteriori ⇒

$P(C_i|X)$

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

→ $P(X)$ is constant for all classes.

$\Rightarrow P(C_i|X) = P(X|C_i)\, P(C_i)$

EX: $P(\text{King}|\text{Face})$

$= \frac{P(\text{Face}|\text{king}) \cdot P(\text{king})}{P(\text{Face})}$

$= \frac{1 \times \frac{4}{52}}{\frac{12}{52}} = \frac{\frac{1}{13}}{\frac{3}{13}} = \frac{1}{3}$

→ Naive Bayes Classifier:

$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) * P(x_2|C_i) \cdots P(x_n|C_i)$

* gaussian distribution:

$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Ex:

Fruit	Yellow	Sweet	long	Total
Orange	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	750
Total	800	850	400	1200

$$P(\text{Yellow}|\text{Orange}) = \frac{P(\text{Orange}|\text{Yellow})\,P(\text{Yellow})}{P(\text{Orange})}$$

$$= \frac{\frac{350}{800} \times \frac{800}{1200}}{\frac{650}{1200}} = 0.53$$

$$P(\text{Sweet}|\text{Orange}) = \frac{P(\text{Orange}|\text{sweet})\,P(\text{Sweet})}{P(\text{Orange})}$$

$$= \frac{\frac{450}{850} \times \frac{850}{1200}}{\frac{650}{1200}} = 0.69$$

$P(\text{Long}|\text{Orange}) = 0$

$P(\text{Fruit}|\text{orange}) = P(Y|O) * P(S|O) * P(L|O)$

$= 0$

$P(\text{Fruit}|\text{Banana}) = P(Y|B) * P(S|B) * P(L|B)$

$= 0.65$

$P(F|\text{others}) = 0.072$

→ Given: Fruit = {Yellow, Sweet, Long}

∴ $P(F|B) = 0.65$

$P(F|\text{others}) = 0.072$

∴ $P(F|B) = 0.65$

* Advantages:
  - Easy to implement
  - Good results obtained in most of the cases.

* Disadvantages:
  - loss of accuracy
  - dependencies among variables.

* Rule-based classification:

→ using IF-THEN rules:

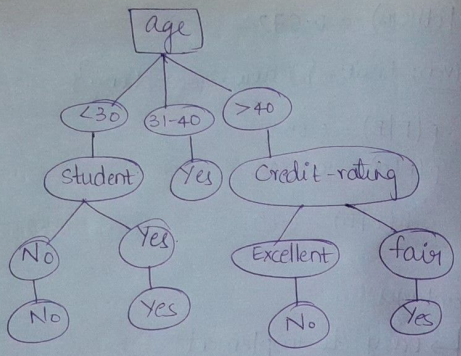
R: if age = youth and student = yes
then buys_computer = yes

* Rule extraction from a decision tree:

→ One rule is created for each path from the root to a leaf.

→ Leaf holds the class label.

Ex:



if age = young and student = no
then buys_computer = no

if age = old and credit_rating = fair
then buys_computer = yes

*Rule Induction:

Sequential covering method:

Steps:

- Rules are learned one at a time
- Each time a rule is learned, tuples covered are removed
- Repeat until quality of rule is below threshold

* Model Evaluation and Selection:

→ Accuracy is assessed using validation test set.

→ Methods to estimate classifier's accuracy:

- Holdout method, random subsampling
- Cross-validation
- Bootstrap.

* Confusion Matrix:

Actual \ Predicted	C_1	$\neg C_1$	
C_1	True Positives	False Neg.	P
$\neg C_1$	False Positives	True Neg.	N
	P'	N'	All

Metrics:

Accuracy: $\frac{TP+TN}{All}$

Error Rate $= \frac{FP+FN}{All}$

Sensitivity $= \frac{TP}{P}$

Specificity $= \frac{TN}{N}$

Precision $= \frac{TP}{TP+FP}$

Recall $= \frac{TP}{TP+FN}$

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

* Evaluating Classifier Accuracy:

→ Holdout method:

Given data is partitioned into 2 sets.

→ Training set ⇒ 2/3 for construction

→ Test set ⇒ 1/3 for accuracy estimation

Random sampling:

Repeat holdout k times

accuracy = avg of accuracies obtained

→ Cross-validation:

↳ k-fold where k = 10.

→ Randomly partition the data into k mutually exclusive subsets each of equal size.

* t-test:

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{Var(M_1 - M_2)/k}}$$

$$Var(M_1 - M_2) = \frac{1}{2} \sum_{i=1}^{k} \left[err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2)) \right]^2$$

⇒ non-paired t-test

$$Var(M_1 - M_2) = \sqrt{\frac{Var(M_1)}{k_1} + \frac{Var(M_2)}{k_2}}$$

M_1 & M_2 → Null hypothesis.

→ select a significance level (sig)

confidence limit = $\frac{sig}{2} = z$

→ $t > z$ or $t < -z$ ⇒ Reject null hypothesis.

⇒ statistically significant diff.

else: any diff. is chance.

→ Receiver Operating Characteristics (ROC) curve is used for visual comparison of classification models.

* Issues effecting model selection:

1. Accuracy
2. Speed
3. Robustness
4. Scalability
5. Interpretability

* Techniques to improve classification accuracy.

→ Ensemble Methods.

* Use a combination of models.
* combine a series of k learned models to create a improved M* model.

→ Popular methods:
- Bagging
- Boosting
- Ensemble

* Bagging: averaging the prediction over a collection of classifiers

→ Analogy: diagnosis based on majority vote.

→ Training
→ Classification
→ Prediction
→ Accuracy:
- Often significantly better than single classifier.
- for noise data.
- improved accuracy in prediction

* Boosting: weighted vote with a collection of classifiers.

Steps:
→ Weights are assigned to each training tuple
→ A series of k classifiers is iteratively learned
→ After a classifier Mi is learned, weights are updated.
→ final M* combines the votes of each classifier

** greater accuracy compared to bagging.
** But risks overfitting the model to misclassify data.

* Classification of Class-Imbalanced data.

Methods:
- Oversampling: resampling of data from +ve class.
- Undersampling: eliminate tuples from -ve class.
- Threshold moving: moves decision threshold; so that rare class tuples are easier to classify.