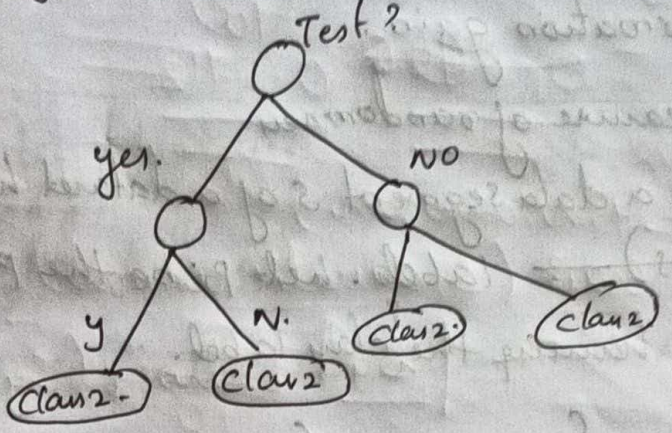
7/3/23
Tuesday

Module-3.

clustering: unsupervised

## Decision tree

Supervised learning method.
It is a hierarchy representation of data.
Simplest type of classification method.




Nodes: Test on attributes.
Branches: outcome for the test.
leaves - final classes.
Examples of decision tree algorithm.
ID3, C4.5, CART

Two types of decision tree:
1) classification → o/p are classes.
2) Regression tree (predicting new value) - leaf nodes are numbers.

consider the following data set.

| Name | Features | | | | class label. |
|---|---|---|---|---|---|
| | give birth | aquatic animal | aerial animal | has legs | |
| human | yes | no | no | yes | mammal |
| Python | no | no | no | no | reptile |
| Salmon | no | yes | no | no | fish |
| frog | yes/no | semi | yes/no | yes | amphibians |
| bat | yes | no | yes | yes | bird |
| Pigeon | no | no | yes | yes | bird |
| cat | yes | no | yes/no | yes | mammal |
| shark | yes | yes | no | no | fish |
| turtle | no | semi | yes/no | yes | amphibians |
| salamander | no | semi | yes/no | yes | amphibians |

Feature selection measures

Two of the popular selection measures are Information gain & Gini index.

CART → Gini index.

ID3 → Information gain.

Entropy - measure of randomness.

Consider a data segment S of a dataset having c no: of class labels. Let $p_i$ be the proportion of examples in S having the $i^{th}$ class label.

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

Problem(1) Total no: of samples $|S| = 10$

Probability of mamal = $\frac{2}{10}$

" " amph = $\frac{3}{10}$

" " bird = $\frac{2}{10}$

" " fish = $\frac{2}{10}$

" " reptile = $\frac{1}{10}$

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

$$= -p_{mam} \log_2(p_{mam}) - p_{amp} \log_2(p_{amp}) - p_{bird} \log_2(p_{bird}) - p_{fish} \log_2(p_{fish}) - p_{reptile} \log_2(p_{reptile})$$

$$= \frac{-2}{10} \log_2\left(\frac{2}{10}\right) - \frac{3}{10} \log_2\left(\frac{3}{10}\right) - \frac{2}{10} \log_2\left(\frac{2}{10}\right) - \frac{2}{10} \log_2\left(\frac{2}{10}\right) - \frac{1}{10} \log_2\left(\frac{1}{10}\right)$$

$= 0.46 + 0.52 + 0.46 + 0.46 + 0.33$

$= 2.232$

Information Gain$(S, A) :=$ Entropy$(S) - \sum_{v \in \text{value}(A)} \frac{|S_v|}{S} \times$ Entropy$(S_v)$

A - attribute / features

V - values of A

S - complete data set

$S_v$ - subtable contains value v.

let A = "Gives birth"

values of A = {yes, no}.

Information Gain("gives birth") = Entropy (S) $- \frac{|S_y|}{|S|}$ Entropy$(S_y)$

$- \frac{|S_{no}|}{|S|}$ Entropy $(S_{no})$.

$= 2.232$

Entropy $(S,$ "gives birth = yes") $= -P_{mam} \log_2(P_{mam}) - P_{bird} \log_2 (P_{bird}) - P_{fish} \log_2 (P_{fish})$

$= -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right)$

$= 1.5$

Entropy $(S,$ "gives birth = no") $= P_{mam} \log_2(P_{mam}) - P_{bird} \log_2 (P_{bird}) - P_{fish} \log_2 (P_{fish}) - P_{rep} \log_2 (P_{rep}) - P_{amp} \log_2 (P_{amp})$

$$= -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1.792$$

Information gain $= 2.24 - \left(\frac{4}{10}\right) \times 1.5 - \left(\frac{6}{10}\right) \times 1.792$

$= 0.5712$

A = aquatic
Values of A = {y, N, semi}
Information Gain (S, aquatic animal) = Entropy (S)

$$- \frac{|S_y|}{S} \text{Entropy}(S_y) - \frac{|S_N|}{S}$$

$$\text{Entropy}(S_N) - \frac{|S_s|}{S} \text{Entropy}(S_s)$$

Entropy (S, "aquatic animal = yes") $= -P_{fish} \log_2(P_{fish})$

$= 1 \log_2(1)$

$= 0$ (pure class)

Entropy (S, "aquatic animal = no") $= -P_{mam} \log_2(P_{mam}) - P_{rep} \log_2 P(rep) - P_{bird} \log_2(P_{bird})$

$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.0643 \; 1.5$

Entropy (S, "aquatic animal = semi") $= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) 0$

Information gain = 2.24 − 0 − $\frac{5}{10}$ × 1.5 − 0.

= 1.49.

A = aerial

Values of A = {Y, N}.

Entropy (S, "aerial animal = yes") = 0.

Entropy (S, "aerial animal = no") = $-\frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{1}{8}\log_2(\frac{1}{8}) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{3}{8}\log_2(\frac{3}{8})$

= 1.905

Information gain (S, aerial animal) = Entropy (S) − $\frac{|S_Y|}{|S|}$ Entropy ($S_Y$) − $\frac{|S_N|}{|S|}$

= 2.24 − $\frac{2}{10}$ × 0 − $\frac{8}{10}$ × 1.905.

= 0.716

A = has legs.

Value of A = {Y, N}

Entropy (S, "has legs = yes") = $-\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{3}{7}\log_2(\frac{3}{7}) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) -$

= 1.5

Entropy(S, "has legs = N") = $-P_{no} \log_2(P_{no}) - P_{yes} \log_2(P_{yes})$.

$= \frac{1}{3} \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \log_2\left(\frac{2}{3}\right)$

$= 0.918$

Information gain = Entropy(S) − $\frac{|S_v|}{|S|}$ Entropy

$= 2.24 - \frac{7}{10} \times 1.5 - \frac{3}{10} \times 0.918$

$= 0.8856$

Aquatic animal is the root node because information gain of aquatic animal is high.

2. Find the root node of Decision tree using ID3 algorithm for the following data

| Age | competetion | Type | class (profit) |
|---|---|---|---|
| old | yes | software | Down. |
| old | NO | " 2 | Down. |
| old | NO | Hardware | Down. |
| mid | yes | software 3 | Down. |
| mid | yes | Hardware | Down. |
| mid | NO | Hardware | UP. |
| mid | NO | software | U.P. |
| new | yes | s/w s | UP. |
| new | NO | H/w | UP |
| new | NO | s/w 5 | UP |

Total Entropy (S) = $\Sigma - P_i \log_2 P_i$

$= -P_{down} \log_2(P_{down}) - P_{up} \log_2(P_{up})$

$= \frac{-5}{10} \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \log_2\left(\frac{5}{10}\right)$

$= 1$

A = ~~down~~ competetion ~~agree~~ age.

values of A = {old, mid, new}.

Entropy (S, "age = old") = ~~-0.~~

Entropy (S, "age = ~~mid~~ new") = $\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$

$= 1$

Entropy (S, "age = ~~new~~") =

Entropy (S, "age = mid") = $\frac{-2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$

$= 1$

Entropy (S, "age = new") = 0

Information gain = $1 - 0 - \frac{4}{10} \times 1 - \frac{3}{10} \times 0$

$= 0.6$

A = competetion.

values of A = {yes, no}.

Entropy (S, "competetion = yes") = $\frac{-3}{4} \log_2\left(\frac{3}{4}\right)$

$- \frac{1}{4} \log_2\left(\frac{1}{4}\right)$.

$= 0.8112$

Entropy $(S, \text{"Comp = no"}) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$

$= 0.918$

Information gain $(S, comp) = 1 - \frac{4}{10} \times 0.8112 - \frac{6}{10} \times 0.918$

$= 0.124$

Ex A = software Type.

Values of A = {S/w, H/w}.

Entropy $(S, \text{Type} = \text{"S/w"}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$= 1$

Entropy $(S, \text{"Type = H/w"}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1$

$(S, \text{Type})$

Info gain $= 1 - \frac{6}{10} \times 1 - \frac{4}{10} \times 1$

$= 1 - 0.6 - 0.4 = 0$

Age is the root node.

| Instance no. | class label | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 1 | T | T |
| 2 | 1 | T | T |
| 3 | 0 | T | F |
| 4 | 1 | F | F |
| 5 | 0 | F | T |
| 6 | 0 | F | T |

Entropy $(S) = \sum p_i \log_2 (p_i)$

$= p_1 \log_2 (p_1) - p_0 \log_2 (p_0)$.

$= -\frac{3}{6} \log_2 \left(\frac{3}{6}\right) - \frac{3}{6} \log_2 \left(\frac{3}{6}\right)$

$= 1$.

$A = x_1$.

Values of $A = \{T, F\}$.

Entropy $(S, "x_1 = T") = -\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right)$

$= 0.918$

Entropy $(S, "x_1 = F") = -\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \left(\frac{2}{3}\right)$.

$= 0.918$.

Information gain $(S, x_1) = 1 - \frac{3}{6} \times 0.918 - \frac{3}{6} \times 0.918$

$= 0.082$.

$A = x_2$

Values of $A = \{T, F\}$.

$$= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\times\log_2\left(\frac{1}{2}\right).$$

$$= 1$$

Information $(S, x_2) = 1 - \frac{1\times 3}{6} - \frac{1\times 3}{6}$

$$= 0$$

$$= 1 - \frac{4}{6}\times 1 - \frac{2}{6}\times 1.$$

$$= 1 - 1 = 0$$

13/3/23 Monday

## Gini Index

It's a degree or probability of random a particular class being wrongly classified.

Gini index is used in CART algorithm.

Consider a dataset 'S' having class labels $C_1, C_2, \ldots C_r$

Let $P_i$ is the probability of examples having class label $C_i$

Then Gini index is calculated by

$$Gini(S) = 1 - \sum_{i=1}^{c} (P_i)^2$$

$$P_{amp} = \frac{3}{10}$$

$$P_{rep} = \frac{1}{10}$$

$$P_{bird} = \frac{2}{10}$$

$$P_{fish} = \frac{2}{10}$$

$$P_{mamal} = \frac{2}{10}$$

$$Gini(S) = 1 - \left[\left(\frac{3}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$$

$$= 0.78$$

## Gini split index

Let S be a set of examples, A - attribute/feature.

$S_v$ - Subtable with $A = v$

Values(A) - set of all possible values of A.

Gini split index of an attribute A is defined by,

~~Gini split (S, A) =~~

$$Gini_{split}(S, A) = \sum_{v \in value(A)} \frac{|S_v|}{|S|} \times Gini(S_v)$$

$Gini_{split}(S, A = \text{"gnes birth"})$.

A = gnes birth

Values of A = {"yes", "no"}

$$Gini_{split}(S, A = gnes\ birth) = \frac{|S_{yes}|}{|S|} \times Gini(S_{gnesbirth=yes})$$

$$= \frac{|S_{no}|}{|S|} \times Gini(S_{gnesbirth=no})$$

$$Gini(S_{gnesbirth=yes}) = 1 - \sum P_i^2 \quad \text{Total no.}$$

$$P(mam) = \frac{2}{4}$$

$$P(bird) = \frac{1}{4}$$

$$P(fish) = \frac{1}{4}$$

$$= 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$$

$$= 0.625$$

$$Gini(S_{gnes\ birth=no}) = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{3}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right]$$

$$= 0.667$$

$$Gini\ split(S, gnes\ birth) = \frac{4}{10} \times 0.625 + \frac{6}{10} \times 0.667$$

$$= 0.6502$$

A = aquatic animal

values of A = {"yes", "no", "semi"}

Gini:

Gini(S, aquatic animal = yes) = 0

Gini(S, aquatic = no) = $1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$

$= 0.64$

Gini(S, aquatic = semi) = 0

Gini split(S, aquatic animal) = $\frac{|S_{yes}|}{|S|} \times Gini(S, aq = yes) + \frac{|S_{no}|}{|S|} \times Gini(S, aq = no) + \frac{|S_{semi}|}{|S|} \times Gini(S, aq = semi)$

$= 0.64 \times \frac{5}{10} = 0.32$

A = aerial animal

values of A = {"yes", "no"}

Gini(S, aerial = yes) = $1 - [0]$

Gini(S, aerial = no) = $1 - \left[\left(\frac{2}{8}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{2}{8}\right)^2 + \left(\frac{3}{8}\right)^2\right]$

$= 0.7187$

Gini split(S, aerial animal) = $\frac{|S_{yes}|}{|S|} \times Gini(S, aerial = y) + \frac{|S_{no}|}{|S|} \times Gini(S, aerial = no)$

$$0 + \frac{8}{10} \times 0.71875$$

$$= 0.575$$

A = has legs.

Value of A = { "yes", "no" }.

$$Gini(S, \text{has legs} = yes) = 1 - \left[\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right].$$

$$= 0.653$$

$$Gini(S, \text{has legs} = no) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right].$$

$$= 0.44$$

$$Gini_{split}(S, \text{has legs}) = \frac{|S_{yes}|}{|S|} \times Gini(S, \text{has legs} = yes) + \frac{|S_{no}|}{|S|} \times Gini(S, \text{has legs} = no)$$

$$= \frac{7}{10} \times 0.653 + \frac{3}{10} \times 0.44.$$

$$= 0.589$$

split attribute

Root node (split node) is selected as the attribute with lowest $Gini_{split}$ index. Here, aquatic animal is the lowest $Gini_{split}$ index.

$Gini_{split}$ index is used in CART algorithm.

$P_1 = \frac{3}{6}$

$P_2 = \frac{3}{6}$

$Gini(S) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$

$= \underline{0.5}$

$A = x_1$

values of $A = \{T, F\}$

$Gini(S, x_1 = T) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right]$

$= \underline{0.44}$

$Gini(S, x_1 = F) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$

$= \underline{0.44}$

$Gini(S, x_1) = \frac{|S_T|}{|S|} \times Gini(S, x_1 = T) + \frac{|S_F|}{|S|} \times Gini(S, x_2 = F)$

$= \frac{3}{6} \times 0.44 + \frac{3}{6} \times 0.44$

$= \underline{0.44}$

$A = x_2$

values of $A = \{T, F\}$

$Gain\ Gini(S, x_2 = T) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$

$= \underline{0.5}$

$Gini(S, x_2 = F) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$

$= 0.5$

$$Gini(S, X_2) = \frac{4}{6} \times 0.5 + \frac{2}{6} \times 0.5$$

$$= 0.5$$

$X_1$ is the root node.

14/3/23 Tuesday

## Measuring Performance of a classifier Algorithm

- confusion matrix

| predicted class | Actual class: yes | Actual class: NO |
|---|---|---|
| yes | TP | FP |
| NO | FN | TN |

TP → True positive
FP → False positive
FN - False negative
TN - True negative

2 class classification problem.

Confusion matrix is a model that is used to describe the performance of a classification model. It is a table that categorises predictions according to whether they match the actual values.

Precision
Recall
accuracy.

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

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

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

1) Suppose a computer program for recognizing dogs in a photograph identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 dogs identified, 5 actually are dogs, while the rest are cats. Compute the precision and recall for the computer program.

| predicted \ Actual | dog | cat | |
|---|---|---|---|
| dog | 5 | 3 | 8 |
| cat | 7 | | |
| | sum=12 | | |

TP = 5
FP = 3
FN = 7

$$\text{precision} = \frac{TP}{TP + FP} = \frac{5}{5+3} = \frac{5}{8} = 0.625$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{5}{7+5} = \frac{5}{12} = 0.416$$

2) Let there be 10 balls (6 white & 4 red balls) in a box and it is desired to pick up red balls from them. Suppose we pick 7 balls as red of which 2 are actually red balls what are the values of precision & recall in picking red ball.

| predicted \ Actual | red | white | predicted total |
|---|---|---|---|
| red | 2 | 5 | 7 |
| white | 2 | 1 | |
| | 4 | 6 | |

TP = 2
FP = 5
FN = 2

$$\text{precision} = \frac{2}{2+5} = \frac{2}{7} = \frac{2}{2+5} = \frac{2}{7}$$

$$\text{Recall} = \frac{2}{2+2} = \frac{2}{4} = \frac{2}{2+2} = \frac{2}{4}$$

3) A database contains 80 records on a particular topic of which 55 are relevant to certain investigations. A search was conducted on that topic and 50 records are retrieved, 40 were relevant. Construct the confusion matrix for the search and calculate the precision, recall and finally the accuracy whole classification problem

| | | Actual: relevant | Actual: not relevant | |
|---|---|---|---|---|
| predicted | relevant | 40 | 10 | 50 |
| | not relevant | 15 | 15 | |
| | | 55 | | |

TP = 40
FP = 10
TN = 15
FN = 15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{40}{40+10} = \frac{40}{50}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{40}{40+15} = \frac{40}{55}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{40+10}{40+15+10+15} = \frac{50}{80}$$

23 Tuesday

## The ID3 Algorithm

- Developed by Ross Quinlan
- Iterative Dichotomiser 3
- Assumptions
  - The algorithm uses information gain to select the

most useful attribute for classification.
•Assume that there are only two class labels namely, "+" and "−". The example with class labels "+" are called positive examples and other negative examples.

Notations

The following notations are used in the algorithm.

| | |
|---|---|
| S | The set of examples. |
| C | The set of class labels. |
| F | The set of features. |
| A | An arbitary feature (attribute). |
| values(A) | The set of values of the feature A. |
| V | An arbitary value of A. |
| $S_v$ | The set of examples with A = V. |

Algorithm ID3 (S, F, C)

1. Create a root node for the tree.
2. if (all examples in S are positive) then.
3. return single node tree Root with label "+"
4. end if.
5. if (all examples are negative) then
6. return single node tree Root with label "−".
7. end if.
8. if (number of feature is 0) then
9. return single nodes root tree.
10. else
11. let A be the feature in F with the highest information gain.
12. Assign A to the root node in decision tree

13. for all (values v of A) do.
14. Add a new tree branch below Root node.
corresponding to v.
15. if (S, is empty) then.
16. Below this branch add a leaf node with label
equal to the most common class label in the set
17. else
18. Below this branch add the subtree formed by applying
the same algorithm ID3 with the values.
ID3 ($S_v$, C, I-{A}).
19. end if
20. end for
21. end if

4) Consider a two class classification problem of predicting whether a photograph contains a man or a woman. Suppose we have a test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm.

| | Expected | predicted. |
|---|---|---|
| 1 | man | woman. |
| 2 | man | man |
| 3 | woman | woman |
| 4 | man | man. |
| 5 | woman | man |
| 6 | woman | woman |
| 7 | woman. | woman |
| 8 | man. | man |
| 9 | man | woman |
| 10 | woman. | women. |

a) Compute the confusion matrix for the data.

b) Compute the accuracy, precision recall.

| Predicted \ Actual | Man | woman | |
|---|---|---|---|
| Man | 3 | 1 | 4 |
| woman | 2 | 4 | 6 |
| | 5 | 5 | |

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

$$= \frac{3}{3+1} = 3/4$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+2} = 3/5$$

$$\text{Accuracy} = \frac{3+4}{3+2+1+4} = 7/10$$

5) Suppose 10000 patients tested for flu out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the accuracy, precision and recall for the data.

| Predicted \ Actual | Sick | Healthy | |
|---|---|---|---|
| Sick | 620 | 180 | |
| Healthy | 380 | 8820 | |
| | 1000 | 9000 | |

Precision $= \frac{620}{620+180} = \frac{620}{800} = \frac{31}{40}$

$$\begin{array}{r} 1620 \\ 5800 \\ \hline 170 \end{array}$$

Recall $= \frac{620}{620+380} = \frac{620}{1000} = \frac{31}{51.}$

Accuracy $= \frac{620+8820}{620+180+380+8820} = \frac{9440}{10000} = \frac{944}{1000}$

13/23 Sunday) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) what is the mean of the data? what is the median?

(b) Use smoothing by bin means to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

c) How might you determine outliers in the data?

d) what other methods are there for data smoothing?

e) Use min-max ~~algorithm~~ normalization to transform value 35 for age on to ~~the~~ the range [0.0, 1.0]

f) Comment on which method of normalization you would prefer to use for the given data, giving reasons as to why?

f) Use Z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

g) Use normalization by decimal scaling to transform the value 35 for

a) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25,
30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Mean = 29.96

b) Median = 25.

b) ~~Smoothing by bin means of b~~ =

Bin 1: 13, 15, 16 {
Bin 2: 16, 19, 20
Bin 3: 20, 21, 22
" 4: 22, 25, 25.
" 5: 25, 25, 30.
" 6: 33, 33, 35
" 7: 35, 35, 35
" 8: 36, 40, 45
" 9: 46, 52, 70.

Smoothing by bin means

Bin 1: 14.6, 14.6, 14.6. Replaced by mean value
Bin 2: 18.3, 18.3, 18.3.
Bin 3: 21, 21, 21
" 4: 24, 24, 24.
" 5: 26.6, 26.6, 26.6
" 6: 33.6, 33.6, 33.6
" 7: 35, 35, 35
" 8: 40.3, 40.3, 40.3
" 9: 56, 56, 56

c) Eliminating outliers (values fall outside the data)

d) Smoothing by bin medians
" " " boundaries

(e) new.max = 1.0

new.min = 0.0

V = 35.

minA = 13

maxA = 70.

$$V' = \frac{V - minA}{maxA - minA} (new\_max_A - new\_min_A) + new\_min$$

$$= \frac{35 - 13}{70 - 13} (1.0 - 0) + 0.$$

$$= 0.385$$

(f) $$V' = \frac{V - \bar{A}}{\sigma A}$$

$\bar{A} = 29.96$

$$V' = \frac{35 - 29.96}{12.94}$$

$$= 0.389$$

complete data 2 to 3 digit divided by 10^3 decimal digit scale change value complete data divided by

(g) $$v' = \frac{V}{10^3} = \frac{35}{100} = 0.35$$

h) Z-score normalization because

2) Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

| age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| %fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 | 34.6 | 42.5 | 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

a) calculate the mean, median and SD of age and fat
(b) Normalize the two variables based on z-score normalization
c) Calculate the correlation coefficient (pearson's prdt moment coefficient) Are these two variables positively or negatively correlated

a) mean of age = $\frac{46.4}{}$

Median of age = $\frac{50+52}{2} = \frac{51}{}$

decision tree

7.8, 9.5, 17.8, 25.9, 26.5, 27.

Mean of fat = $\frac{26.9}{}$