DATA WAREHOUSING & MINING

Unit-4: Classification

Parameshwar R Hegde, PhD

Basic Concepts

- **Classification**
  - Classification in data mining is a systematic approach that separates data points into different classes
  - It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones
  - It primarily involves using algorithms that can be easily modified to improve the data quality
  - The primary goal of classification is to connect a variable of interest with the required variables

Eg., Students list with their class labels (pass/ fail information)

Basic Concepts

- **Classification**
  - ➢ There are two steps
    - I. Model construction
    - II. Model usage

- **Predication**
  - ➢ Process of identifying the missing or unavailable numerical data for a new object
  - ➢ Algorithm which we use training dataset to derive a model, that model is predictor when a new data is given this model this model should find the output

  Eg., Predict the class label for students based on their marks (use the constructed model for reference)

Decision Tree Induction

- **It is a classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given data set**

- **The set of records available for developing classification methods is generally divided into two disjoint subsets - a training set and a test set**

- **The “training set” is used to derive the classifier and The “test set” is used to measure the accuracy of the classifier**

- **Common properties**
  - ➢ An inner node represents an attribute
  - ➢ An edge represents a test on the attribute of the father node
  - ➢ A leaf represents one of the classes

Decision Tree Induction

- **Construction of a tree**
  - ➢ Based on the training data
  - ➢ Top-Down strategy

Example

Training Data Set

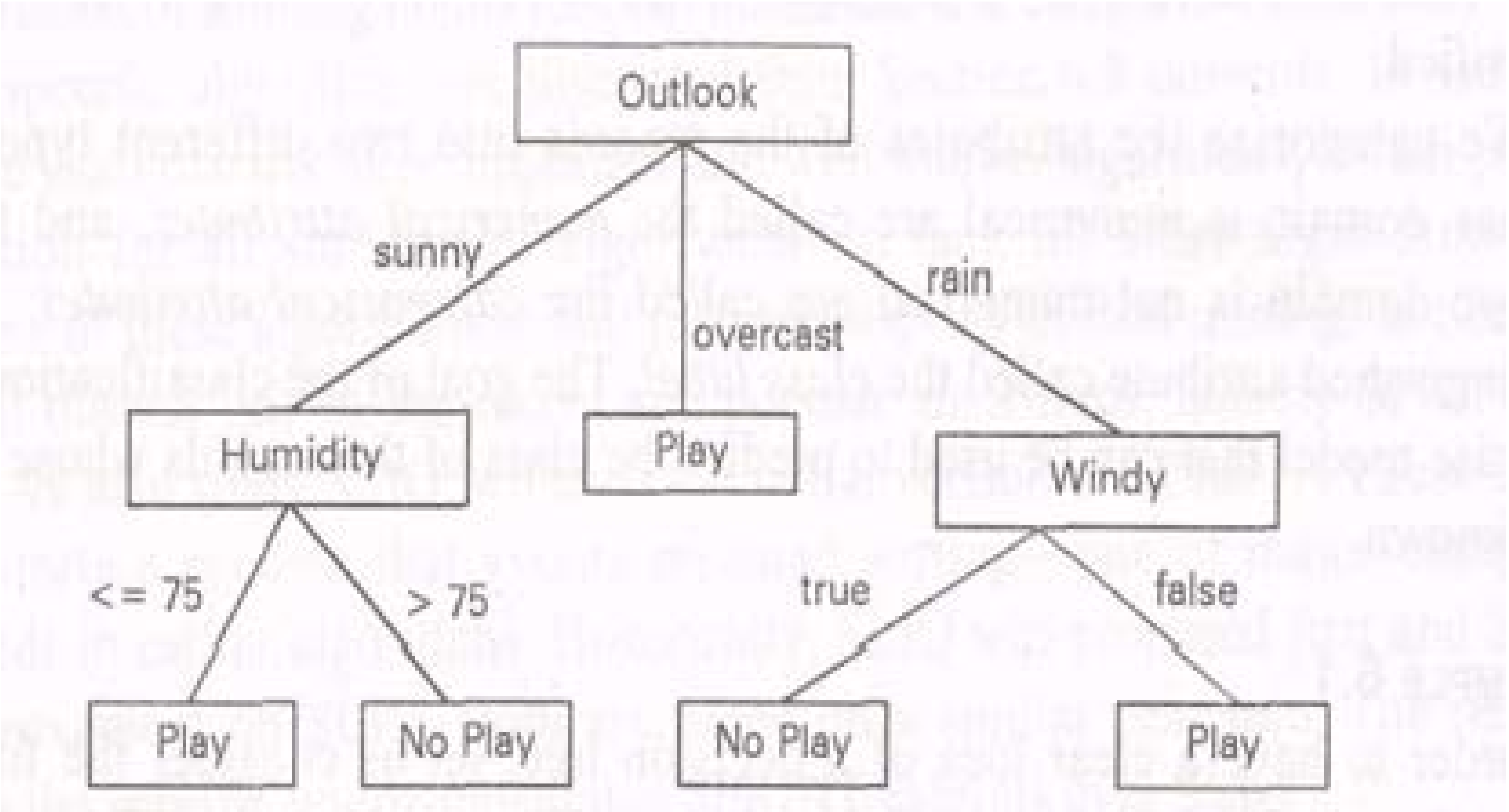
OUTLOOK	TEMP(F)	HUMIDITY(%)	WINDY	CLASS
sunny	79	90	true	no play
sunny	56	70	false	play
sunny	79	75	true	play
sunny	60	90	true	no play
overcast	88	88	false	no play
overcast	63	75	true	play
overcast	88	95	false	play
rain	78	60	false	play
rain	66	70	false	no play
rain	68	60	true	no play

Decision Tree Induction

- **Example**
  - In this, Five leaf nodes are present
  - In a decision tree, each leaf node represents a rule
    - **Rule 1** - If it is sunny and the humidity is not above 75%, then play
    - **Rule 2** - If it is sunny and the humidity is above 75%, then do not play
    - **Rule 3** - If it is overcast, then play
    - **Rule 4** - If it is rainy and not windy, then play
    - **Rule 5** - If it is rainy and windy, then don't play

Decision Tree Induction

- **Example**



Decision Tree Induction

- **Advantages**
  - Able to generate understandable rules
  - Able to handle both numerical and categorical attributes
  - Provide clear indication of which fields are most important for prediction or classification

- **Disadvantages**
  - Process of growing a decision tree is computationally expensive. At each node, each candidate splitting field is examined before its best split can be found
  - Some decision tree can only deal with binary-valued target classes

Splitting Indices

- **To find out best split, there are 3 measures**
  I. Entropy
  II. Gini
  III. Classification error

Splitting Indices

- **Entropy**
  - It measures the randomness in the information being processed
  - It measures the purity of the split
  - The higher the Entropy, the harder it is to draw any conclusions from that information

$$\text{Entropy (t)} = -\sum_{i=0}^{c-1} p\left(\frac{i}{t}\right) \log_2 p\left(\frac{i}{t}\right)$$

Splitting Indices

- **Gini**
  - Gini impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set

$$\text{Gini (t)} = 1 - \sum_{i=0}^{c-1} \left[p\left(\frac{i}{t}\right)\right]^2$$

Splitting Indices

- **Classification error**

$$\text{Classification error (t)} = 1 - \max[\mathrm{p}(\frac{i}{t})]$$

Splitting Indices

- **Find Entropy, Gini and classification error**
  - **Example 1**

Node	Count
Class=0	0
Class=1	6

Splitting Indices

- **Solution**

$$\text{Entropy (t)} = -\sum_{i=0}^{c-1} p\left(\frac{i}{t}\right)\log_2 p\left(\frac{i}{t}\right)$$

$$\text{Entropy (t)} = -\left[\frac{0}{6}\log_2\frac{0}{6} - \frac{6}{6}\log_2\frac{6}{6}\right]$$

$$\text{Entropy (t)} = -[0.\log_2 0 - 0.\log_2 1]$$

$$\text{Entropy (t)} = -[0 - 1]$$

$$\text{Entropy (t)} = 1$$

Splitting Indices

- **Solution**

$$\text{Gini (t)} = 1 - \sum_{i=0}^{c-1} \left[p\left(\frac{i}{t}\right)\right]^2$$

$$\text{Gini (t)} = 1 - \left(\frac{0}{6}\right)^2 - \left(\frac{6}{6}\right)^2$$

$$\text{Gini (t)} = 1 - 0 - 1$$

$$\text{Gini (t)} = 0$$

Splitting Indices

- **Solution**

$$\text{Classification error (t)} = 1 - \max[\mathrm{p}(\frac{i}{t})]$$

$$\text{Classification error (t)} = 1 - \max[\left(\frac{0}{6}\right), \left(\frac{6}{6}\right)]$$

$$\text{Classification error (t)} = 1 - \left(\frac{6}{6}\right)$$

$$\text{Classification error (t)} = 1 - 1$$

$$\text{Classification error (t)} = 0$$

Splitting Indices

- **Find Entropy, Gini and classification error for A and B**
  - **Class activity**

Instance	A	B	Target
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-

Splitting Indices

- **Solution**

Entropy (A) =

$$-\sum_{i=0}^{c-1} p\left(\frac{A_T}{A_T}\right)\log_2 p\left(\frac{A_F}{A_T}\right)$$

Entropy (A) = $-\left[\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}\right]$

<u>Entropy (A) = 0.99</u>

Entropy (A_{True}) =

$$-\sum_{i=0}^{c-1} p\left(\frac{A_+}{A_{True}}\right)\log_2 p\left(\frac{A_-}{A_{True}}\right)$$

Entropy (A_{True}) = $-\left[\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right]$

<u>Entropy (A_{True}) = 0.81</u>

<u>Entropy (A_{False}) = 0.72</u>

Splitting Indices

- **Solution**

$$\text{Gini (A)} = 1 - \sum_{i=0}^{c-1} \left[p\left(\frac{A}{A_{Total}}\right)\right]^2$$

$$\text{Gini (A)} = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2$$

$$\text{Gini (A)} = 1 - 0.19 - 0.30$$

<u>Gini (A) = 0.51</u>

$$\text{Gini }(A_{True}) = 1 - \sum_{i=0}^{c-1} \left[p\left(\frac{A_{+\&}}{A_{True}}\right)\right]^2$$

$$\text{Gini }(A_{True}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$\text{Gini }(A_{True}) = 1 - 0.56 - 0.06$$

<u>Gini (A_{True}) = 0.38</u>

Splitting Indices

- **Solution**

Classification error (A) = $1 - \max[p(\frac{A}{A_{Total}})]$

Classification error (A) = $1 - \max[\left(\frac{4}{9}\right), \left(\frac{5}{9}\right)]$

Classification error (A) = $1 - \left(\frac{5}{9}\right)$

Classification error (t) = $1 - 0.55$

Classification error (t) = 0.45

Classification error (A_T) = $1 - \max[p(\frac{A}{A_{True}})]$

Classification error (A_T) = $1 - \max[\left(\frac{3}{4}\right), \left(\frac{1}{4}\right)]$

Classification error (A_T) = $1 - \left(\frac{3}{4}\right)$

Classification error (A_T) = $1 - 0.75$

Classification error (A_T) = 0.25

Bayes Classification

- **Statistical classifier depends on the Bayes theorem**

- **They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class**

- **Advantages**
  - Bayes classification are as efficient as decision tree and neural network classifier
  - They are very accurate
  - They exhibit high speed
  - They make the process of computation simple

Bayes Classification

- **Bayes theorem**
  - Bayes theorem plays a critical role in probabilistic learning and classification
  - It is a mathematical formula used in calculating conditional probability
  - It is based on conditional probability
  - It predicts the occurrence of any event
  - It depends on conditional probability, that describe occurrence of an event X with respect to condition Y
  - **Conditional probability** refers to the chances that some outcome occurs given that another event has also occurred
  - It is often stated as the probability of B given A and is written as P(B|A), where the probability of B depends on that of A happening

Bayes Classification

$$p\left(\frac{A}{B}\right) = \frac{p(A \cap B)}{p(B)} \qquad (1)$$

$$p\left(\frac{B}{A}\right) = \frac{p(A \cap B)}{p(A)} \qquad (2)$$

(1) Can be written as, P(A/B) . p(B) = p(A ∩B)

(2) Can be written as, P(B/A) . p(A) = p(A ∩B)

Now RHS became the same

Bayes Classification

Therefore, $p\left(\frac{A}{B}\right).p(B) = p\left(\frac{B}{A}\right).p(A)$

$$p\left(\frac{A}{B}\right) = \frac{p\left(\frac{B}{A}\right).p(A)}{p(B)}$$

This is known as Bayes theorem

Where, A = Hypothesis

B = Evidence (Data tuple)

Naïve Bayes Classification

- **Example**

Outlook – Sunny

Temp. – Cool

Humidity – High

Wind – Strong

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Bayes Classification

Step 1: Calculate the prior probabilities

P(yes) = 9/14 = 0.64

P(no) = 5/14 = 0.36

Step 2: Calculate the conditional probability for each attributes

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0
Rain	3/9	2/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Temp	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Bayes Classification

Step 3: Calculate the new instance in to yes or no class

i.e., Outlook=sunny; Temp.=cool; Humidity=high; Wind=Strong

$$V_{NB} = argmax_{v_j \in \{yes,no\}} p(v_j) \prod_i p(a_i|v_j)$$

$$V_{NB} = argmax_{v_j \in \{yes,no\}} p(v_j)\, \mathrm{p}(\mathrm{Outlook} = \mathrm{sunny}|v_j)\mathrm{p}(\mathrm{Temp.} = \mathrm{cool}|v_j)$$
$$\mathrm{p}(\mathrm{Humidity} = \mathrm{high}|v_j)\, \mathrm{p}(\mathrm{Wind} = \mathrm{Strong}|v_j)$$

$$V_{NB}(yes) = p(yes)p(sunny|yes)p(cool|yes)p(high|yes)p(strong|yes) = 0.005$$
$$V_{NB}(no) = p(no)p(sunny|no)p(cool|no)p(high|no)p(strong|no) = 0.206$$

Since the $V_{NB}(no)$ value is high, the new instance will be classified as ‘no’ class

Bayes Classification

Step 4: Normalize the calculated probability values

$$V_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = \frac{0.053}{0.053+0.206} = 0.205$$

$$V_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = \frac{0.206}{0.053+0.206} = 0.795$$

Naïve Bayes Classification

- **Class activity**

Outlook – Rain

Temp. – Mild

Humidity – High

Wind – Weak

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No