

**COMP 3710 - 3**

**Applied Artificial Intelligence (3,1,0)**

**Fall 2017**

**Seminar/Lab 7.**

**Decision tree, and k-Nearest Neighbor (kNN) algorithm**

**Instructor:** Mahnhoon Lee

**Student Name:** ZHENYU WANG

**Student Number:**  T00059541

1. **Decision tree**

Here is the training data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Film** | ***Country*** | ***Big Star*** | ***Genre*** | **Success** |
| Film 1 | *USA* | *Yes* | *SF* | False |
| Film 2 | *USA* | *No* | *Comedy* | False |
| Film 3 | *USA* | *Yes* | *Comedy* | True |
| Film 4 | *Europe* | *No* | *Comedy* | True |
| Film 5 | *Europe* | *Yes* | *SF* | True |
| Film 6 | *Europe* | *Yes* | *Romance* | False |
| Film 7 | *Other* | *Yes* | *Comedy* | True |
| Film 8 | *Other* | *No* | *SF* | False |
| Film 9 | *Europe* | *Yes* | *Comedy* | False |
| Film 10 | *USA* | *Yes* | *Comedy* | True |

1. Construct a decision tree with the above table. You should show how in the tree is constructed, by computing information gains and entropies.

Gain(Country) = 1 – w-entropy(USA) – w-entropy(Europe) – w-entropy(Other)

= 1 – 4/10 \* H(USA) – 4/10 \* H(Europe) – 2/10 \* H(Other)

= 1 – 4/10 \* (-2/4 log2(2/4) – 2/4 log2(2/4)) – 4/10 \* H(Europe) – 2/10 \* H(Other)

= 1 – 4/10 \* (-2/4 log2(2/4) – 2/4 log2(2/4)) – 4/10 \* (-1/2 log2(1/2) – 1/2 log2(1/2))- 2/10 \* (-1/2 log2(1/2) – 1/2 log2(1/2))

= 1 – (4/10 \* 1) – (4/10 \* 1) – (2/10 \*1)

= 1 – 0.4 – 0.4 – 0.2

= 0

Gain(Big Star) = 1 – w-entropy(YES) – w-entropy(NO)

= 1 – 7/10 \* H(YES) – 3/10 \* H(NO)

= 1 – 7/10 \* (-4/7 log2(4/7) – 3/7 log2(3/7)) – 3/10 \* (-2/3 log2(2/3) – 1/3 log2(1/3))

= 1 – (0.7 \* 0.985) – (0.3 \* 0.918)

= 0.0351

Gain(Genre) = 1 – w-entropy(SF) – w-entropy(Comedy) – w-entropy(Romance)

= 1 – 3/10 \* H(SF) – 6/10 \* H(Comedy) – 1/10 \* H(Romance)

= 1 – 3/10 \* (-1/3 log2(1/3) – 2/3 log2(2/3)) – 6/10 \* (-4/6 log2(4/6) – 2/6 log2(2/6)) - 0

= 1 – (0.3 \* 0.1245) – (0.6 \* 0.918) – 0

= 0.41185

* The information gain for *Country*= 0
* The information gain for *Big Star* = 0.0351
* The information gain for *Genre* = 0.41185
* Therefore, the attribute *Genre* provides the greatest information gain and so is placed at the top of the decision tree.

Romance

Comedy

SF

SF:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Film 1 | *USA* | *Yes* | *SF* | False |
| Film 5 | *Europe* | *Yes* | *SF* | True |
| Film 8 | *Other* | *No* | *SF* | False |

Gain(Country) = 1 – w-entropy(USA) – w-entropy(Europe) – w-entropy(Other)

= 1 – 1/3 \* H(USA) – 1/3 \* H(Europe) – 1/3 \* H(Other)

= 1 – 1/3 \* 0 – 1/3 \* 0- 1/3 \* 0

= 1

Gain(Big Star) = 1 – w-entropy(YES) – w-entropy(NO)

= 1 – 2/3 \* H(YES) – 1/3 \* H(NO)

= 1 – 2/3 \* (-1/2 log2(1/2) – 1/2 log2(1/2)) – 1/3 \* 0

= 1 – (2/3 \* 1)

= 1/3 = 0.33

* The information gain for *Country*= 1
* The information gain for *Big Star* = 0.33
* Therefore, the attribute *Country* provides the greatest information gain and so is placed under the SF.

Comedy:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Film 2 | *USA* | *No* | *Comedy* | False |
| Film 3 | *USA* | *Yes* | *Comedy* | True |
| Film 4 | *Europe* | *No* | *Comedy* | True |
| Film 7 | *Other* | *Yes* | *Comedy* | True |
| Film 9 | *Europe* | *Yes* | *Comedy* | False |
| Film 10 | *USA* | *Yes* | *Comedy* | True |

Gain(Country) = 1 – w-entropy(USA) – w-entropy(Europe) – w-entropy(Other)

= 1 – 3/6 \* H(USA) – 2/6 \* H(Europe) – 1/6 \* H(Other)

= 1 – 3/6 \* (-2/3 log2(2/3) – 1/3 log2(1/3))– 2/6 \* (-1/2 log2(1/2) – 1/2 log2(1/2))- 1/6\* (-1 log2(1) – 1 log2(1))

= 1 – (3/6 \*0.918) – (2/6 \* 1) – (1/6 \*0)

= 1 – 0.459 – 0.33

= 0.211

Gain(Big Star) = 1 – w-entropy(YES) – w-entropy(NO)

= 1 – 4/6 \* H(YES) – 2/6 \* H(NO)

= 1 – 4/6 \* (-3/4 log2(3/4) – 1/4 log2(1/4)) –2/6 \* (-1/2 log2(1/2) – 1/2 log2(1/2))

= 1 – (4/6 \*0.811) - (2/6 \* 1)

= 0.126

* The information gain for *Country*= 0.211
* The information gain for *Big Star* = 0.126
* Therefore, the attribute *Country* provides the greatest information gain and so is placed under the comedy.

Romance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Film 6 | *Europe* | *Yes* | *Romance* | False |

Gain(Country) = 1 – w-entropy(USA) – w-entropy(Europe) – w-entropy(Other)

= 1 –0– 1 \* H(Europe) –0

= 1 – 1 \* 0

= 1

Gain(Big Star) = 1 – w-entropy(YES) – w-entropy(NO)

= 1 – 1 \* H(YES) – 0 \* H(NO)

= 1 –1\* (-1 log2(1) –0

= 1 – (1 \* 0)

= 1

* The information gain for *Country*= 1
* The information gain for *Big Star* = 1
* And there is only one sample for Romance, the result only can be false.

Romance

Comedy

SF

USA

Europe

Other

USA

Other

Europe

Yes

No

Yes

Yes

No

1. Answer the next two queries using the above decision tree.

* (China, Yes, SF)

From above decision tree, we can know other country in SF and has Big Star will be False

* (USA, No, Comedy)

From above decision tree, we can see USA in Comedy and doesn’t have Big Star will be False

1. **k-Nearest Neighbor (kNN) algorithm**

Here is the training data set for class grades.

|  |  |  |  |
| --- | --- | --- | --- |
| MATH 1650 | COMP 2230 | COMP 3710 | Grade level |
| 4 | 3 | 4 | Excellent |
| 3 | 4 | 3 | Good |
| 3 | 4 | 4 | Excellent |
| 2 | 3 | 3 | Okay |
| 3 | 3 | 2 | Good |
| 2 | 3 | 2 | Okay |

* 1. Find the grade levels for the next two queries using the 3NN algorithm.
* (3, 2, 4)
* (3, 4, 3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (3,2,4) | (4,3,4) |  |  | Excellent |
|  | (3,4,3) |  |  | Good |
|  | (3,4,4) |  | 2 |  |
|  | (2,3,3) |  |  | Okay |
|  | (3,3,2) |  |  |  |
|  | (2,3,2) |  |  |  |

To decide the grade level for (3,2,4), we can see from above table, the minimum distance is , which means Excellent.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (3,4,3) | (4,3,4) |  |  |  |
|  | (3,4,3) |  | 0 | Good |
|  | (3,4,4) |  | 1 | Excellent |
|  | (2,3,3) |  |  | Okay |
|  | (3,3,2) |  |  |  |
|  | (2,3,2) |  |  |  |

To decide the grade level for (3,4,3), we can see from above table, the minimum distance is 0, which means Good.