Ben Simpson

HW2

CODE: https://github.com/bmsr56/cs5402/tree/master/hw2

***Task 1***

***Describe the difference between classification and clustering?***

Classification is used in supervised learning. Using the properties of the data, these labels are assigned to instances of the data.

Clustering is used in unsupervised learning. In clustering, similar instances are grouped based on their features.

***Task 2***

***Describe what is entropy?***

Entropy is a measure of uncertainty in the data.

***Task 3***

**Information Gain:**

Better to use when the partitions are the values are great in number and distinct. Entropy is used for the base calculations, resulting in wide ranging results.

**Gain ratio:**

Measure of statistical dispersion. GR can reduce a bias towards multi-valued attributes. This is because it takes the number and size of branches into account when choosing an attribute.

**Gini Index:**

Gini Index = 0 when things are perfectly classified. Larger data partitions preferred.

**Task 4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Yes** | **No** | **Tot.** | **Gini Index** | **Weighted Gini index** |
| **Outlook** |  |  |  |  |  |
| **Sunny** | **2** | **3** | **5** | **1 - (2/5)^2 - (3/5)^2 = 0.48** |  |
| **Overcast** | **4** | **0** | **4** | **1 - (4/4)^2 - (0/4)^2 = 0** |  |
| **Rainy** | **3** | **2** | **5** | **1 - (3/5)^2 - (2/5)^2 = 0.48** |  |
|  |  |  |  |  | **(5/14)(.48) + (4/14)(0) + (5/14)(.48) = 0.342** |
| **Temp** |  |  |  |  |  |
| **Hot** | **2** | **2** | **4** | **1 - (2/4)^2 - (2/4)^2 = 0.5** |  |
| **Cool** | **3** | **1** | **4** | **1 - (3/4)^2 - (1/4)^2 = 0.375** |  |
| **Mild** | **4** | **2** | **6** | **1 - (4/6)^2 - (2/6)^2 = 0.4444** |  |
|  |  |  |  |  | **(4/14)(0.5) + (4/14)(0.375) + (6/14)(.445) = 0.439** |
| **Humidity** |  |  |  |  |  |
| **High** | **3** | **4** | **7** | **1 - (3/7)^2 - (4/7)^2 = 0.4898** |  |
| **Normal** | **6** | **1** | **7** | **1 - (6/7)^2 - (1/7)^2 = 0.2449** |  |
|  |
|  |  |  |  |  | **(7/14)(0.4898) + (7/14)(0.244) = 0.367** |
| **Windy** |  |  |  |  |  |
| **Weak** | **6** | **2** | **8** | **1 - (6/8)^2 - (2/8)^2 = 0.375** |  |
| **Strong** | **3** | **3** | **6** | **1 - (3/6)^2 - (3/6)^2 = 0.5** |  |
|  |  |  |  |  | **(8/14)(0.375) + (6/14)(0.5) = 0.428** |
| **Result: Outlook, 0.342** |  |  |  |  |  |
| ***Outlook = Sunny*** |  |  |  |  |  |
| **Temp** |  |  |  |  |  |
| **Hot** | **0** | **2** | **2** | **1 – (0/2)^2 – (2/2)^2 = 0** |  |
| **Cool** | **1** | **0** | **1** | **1 – (1/1)^2 – (0/1)^2 = 0** |  |
| **Mild** | **1** | **1** | **2** | **1 – (1/2)^2 – (1/2)^2 = 0.5** |  |
|  |  |  |  |  | **(2/5)(0)+(1/5)(0)+(2/5)(0.5) = 0.2** |
|  |  |  |  |  |  |
| **Humidity** |  |  |  |  |  |
| **High** | **0** | **3** | **3** | **1 – (0/3)^2 – (3/3)^2 = 0** |  |
| **Normal** | **2** | **0** | **2** | **1 – (2/2)^2 – (0/2)^2 = 0** |  |
|  |  |  |  |  | **0** |
| **Windy** |  |  |  |  |  |
| **Weak** | **1** | **2** | **3** | **1 – (1/3)^2 – (2/3)^2 = .266** |  |
| **Strong** | **1** | **1** | **2** | **1 – (1/2)^2 – (1/2)^2 = .2** |  |
|  |  |  |  |  |  |
|  |  |  |  |  | **(3/5)(0.266)+(2/5)(0.2) = 0.466** |
|  |  |  |  |  |  |
| **Result: Humidity, 0** |  |  |  |  |  |
|  |  |  |  |  |  |
| **Temp** |  |  |  |  |  |
| **Hot** | **0** | **2** | **2** | **1 – (0/2)^2 – (2/2)^2 = 0** |  |
| **Cool** | **1** | **0** | **1** | **1 – (1/1)^2 – (0/1)^2 = 0** |  |
| **Mild** | **1** | **1** | **2** | **1 – (1/2)^2 – (1/2)^2 = 0.5** |  |
|  |  |  |  |  | **(2/5)(0)+(1/5)(0)+(2/5)(0.5) = 0.2** |
|  |  |  |  |  |  |
| **Humidity** |  |  |  |  |  |
| **High** | **0** | **3** | **3** | **1 – (0/3)^2 – (3/3)^2 = 0** |  |
| **Normal** | **2** | **0** | **2** | **1 – (2/2)^2 – (0/2)^2 = 0** |  |
|  |  |  |  |  | **0** |
| **Windy** |  |  |  |  |  |
| **Weak** | **1** | **2** | **3** | **1 – (1/3)^2 – (2/3)^2 = .266** |  |
| **Strong** | **1** | **1** | **2** | **1 – (1/2)^2 – (1/2)^2 = .2** |  |
|  |  |  |  |  |  |
|  |  |  |  |  | **(3/5)(0.266)+(2/5)(0.2) = 0.466** |

**Task 5**

Code was written to create a decision tree and classify data. It works.

**See code at** <https://github.com/bmsr56/cs5402/tree/master/hw2>