Machine Learning Homework 1

Andrew Sanders

Due 1/30/2020

Training Set

Example	Type	Price	Buy
CD1	HipHop	Expensive	Yes
CD2	Rock	Cheap	Yes
CD3	Rock	Expensive	Yes
CD4	HipHop	Cheap	Yes
CD5	Jazz	Cheap	Yes
CD6	Rock	Expensive	No
CD7	Jazz	Expensive	No
CD8	Jazz	Cheap	No
CD9	HipHop	Expensive	Yes
CD10	Jazz	Expensive	No
CD11	Rock	Expensive	No
CD12	Jazz	Cheap	Yes
CD13	Rock	Expensive	No

Test Set

Example	Type	Price	Buy
CD1	Rock	Cheap	Yes

CD2	Jazz	Cheap	No
CD3	Jazz	Expensive	No
CD4	Rock	Expensive	Yes
CD5	HipHop	Expensive	Yes

Using ID3 to create Decision Tree from Training Set

$$TrainingSet$$

$$S : [7^+, 6^-]$$

$$E(entropy) : -\frac{7}{13}\log_2\frac{7}{13} - \frac{6}{13}\log_2\frac{6}{13} = .996$$

> **Note**: We need to find the “best” attribute. To find this attribute, we must compare their information gains.
> Information Gain is found by finding each attribute’s entropy and subtracting it from the entire set’s entropy.

Potential Attribute: Type

		Buy	
		Yes	**No**
	HipHop	3	0
Type	**Rock**	2	3
	Jazz	2	3

$$HipHop : [3^+, 0^-]$$

$$E : -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$Rock : [2^+, 3^-]$$

$$E : -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = .971$$

$$Jazz : [2^+, 3^-]$$

$$E : -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = .971$$

Potential Attribute: Price

		Buy	
		Yes	**No**
	Cheap	4	1
Type	**Expensive**	3	5

$$Cheap : [4^+, 1^-]$$

$$E : -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = .722$$

$$Expensive : [3^+, 5^-]$$

$$E : -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = .954$$

Comparing Information Gain

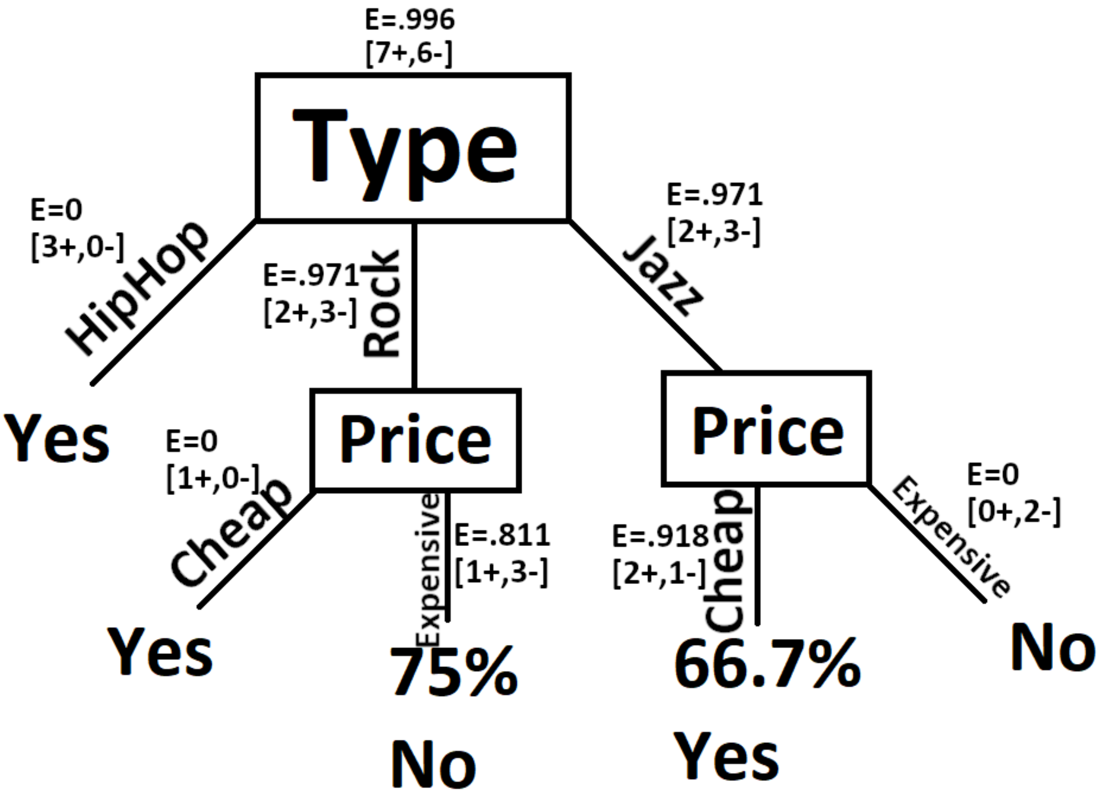
$$Gain(S, Type) = .996 - (3/13)0 - (5/13).971 - (5/13).971 = .25$$

$$Gain(S, Price) = .996 - (5/13).722 - (8/13).954 = .131$$

> Type has a higher information gain (.25>.131) so we use it as the best attribute.
> There are only two attributes, so Price acts as the next best attribute if the entropy is not zero.

Graph

> Entropy of paths are found using the conditional entropy formula
> **i.e.**
>
> $$Expensive|Rock : [1^+, 3^-]$$
>
> $$E(Expensive|Rock) : -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = .811$$



Evaluation of Test Set

> ✓ means correct output
>
> ✕ means incorrect output

$$< Type = Rock, Price = Cheap >= Yes\checkmark$$

$$< Type = Jazz, Price = Cheap >= 66.7\% Yes\times$$

$$< Type = Jazz, Price = Expensive >= No\checkmark$$

$$< Type = Rock, Price = Expensive >= 75\% No\times$$

$$< Type = HipHop, Price = Expensive >= Yes\checkmark$$