COM S 573: Machine Learning

Homework #3

Abdurahman Mohammed

March 14, 2022

Question 1

For this question we will calculate the entropy for each attribute we have. But before that, we will calculate the entropy before splitting.

$$\begin{aligned} Entropy &= -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) \\ &= 0.94 \end{aligned}$$

Now we will compute entropy for each feature. Entropy for **Outlook** feature

$$\begin{aligned} Entropy &= \frac{5}{14}(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})) + \frac{5}{14}(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})) + \frac{4}{14}(-\frac{4}{4}\log_2(\frac{4}{4})) \\ &= \frac{5}{14}(0.971) + \frac{5}{14}(0.971) + \frac{4}{14}(0)) \\ &= 0.693 \end{aligned}$$

Entropy for **Temperature** feature

$$\begin{aligned} Entropy &= \frac{4}{14}(-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})) + \frac{6}{14}(-\frac{2}{6}\log_2(\frac{2}{6}) - \frac{4}{6}\log_2(\frac{4}{6})) + \frac{4}{14}(-\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})) \\ &= 0.911 \end{aligned}$$

Entropy for **Humidity** feature

$$\begin{aligned} Entropy &= \frac{7}{14}(-\frac{3}{7}\log_2(\frac{3}{7}) - \frac{4}{7}\log_2(\frac{4}{7})) + \frac{7}{14}(-\frac{6}{7}\log_2(\frac{6}{7}) - \frac{1}{7}\log_2(\frac{1}{7})) \\ &= 0.0.788 \end{aligned}$$

Entropy for **Wind** feature

$$\begin{aligned} Entropy &= \frac{8}{14}(-\frac{2}{8}\log_2(\frac{2}{8}) - \frac{6}{8}\log_2(\frac{6}{8})) + \frac{6}{14}(-\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2(\frac{3}{6})) \\ &= 0.892 \end{aligned}$$

Now that we have the entropies for each attribute calculated, we can calculate and compare the information gain.

$$IG(Outlook) = E_{before} - E_{after} = 0.94 - 0.63 = 0.247$$

$$IG(Temperature) = E_{before} - E_{after} = 0.94 - 0.911 = 0.029$$

$$IG(Humidity) = E_{before} - E_{after} = 0.94 - 0.788 = 0.152$$

$$IG(Wind) = E_{before} - E_{after} = 0.94 - 0.892 = 0.048$$

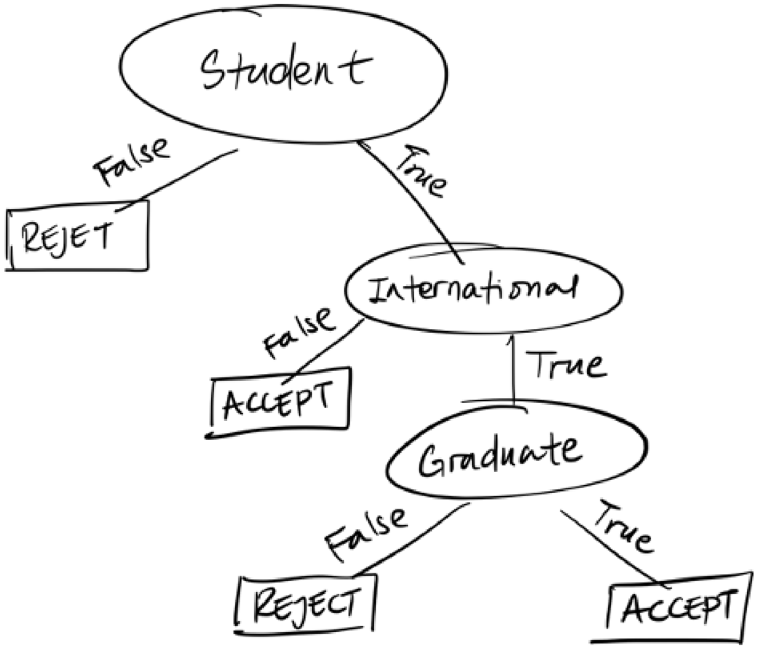
Looking at the information gains calculated, The Outlook feature has the highest information gain. Hence, we will pick it.

Question 2

Yes, it is possible to convert set of rules to a decision tree. We will show this using an example. Let's say we have the following set of rules.

1. If student ==”FALSE” then REJECT
2. If student ==”TRUE” and International==”FALSE” then ACCEPT
3. If student ==”TRUE” and International==”TRUE” and Graduate==”FALSE” then REJECT
4. If student ==”TRUE” and International==”TRUE” and Graduate==”TRUE” then ACCEPT

An equivalent decision tree for the given set of rules can be constructed as follows.



Question 3

1. We want to prove the following where the right hand side represents the standard euclidean distance.

$$(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i - \frac{1}{\sqrt{d}}\sum_{i=1}^{d} z_i) \leq \sum_{i=1}^{d}(x_i - z_i)^2$$

.

We will use the hint given, Jensen's inequality.

$$f(E[X]) \leq E[f(X)]$$

Considering X to be a random number and f is a convex function, let's substitute X by $x - z$.

$$f(E[x-z]) \leq E[f(x-z)]$$
$$(E[x-z])^2 \leq E[(x-z)^2]$$

$$(\frac{1}{d}\sum_{i=1}^{d} x_i - \frac{1}{d}\sum_{i=1}^{d} z_i)^2 \leq \frac{1}{d}\sum_{i=1}^{d}(x_i - z_i)^2$$

$$(\frac{1}{\sqrt{d}}\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i - \frac{1}{\sqrt{d}}\frac{1}{\sqrt{d}}\sum_{i=1}^{d} z_i)^2 \leq \frac{1}{d}\sum_{i=1}^{d}(x_i - z_i)^2$$

$$\frac{1}{d}(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i - \frac{1}{\sqrt{d}}\sum_{i=1}^{d} z_i)^2 \leq \frac{1}{d}\sum_{i=1}^{d}(x_i - z_i)^2$$

We can cancel $\frac{1}{d}$ from both sides to get the following equation.

$$(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i - \frac{1}{\sqrt{d}}\sum_{i=1}^{d} z_i) \leq \sum_{i=1}^{d}(x_i - z_i)^2$$

.

Which is the desired equation.

2. We know that the choice of distance metrics affects the runtime performance of the KNN algorithm. specially when features are in higher dimensional spaces. By using the equation we proved using Jensen's inequality, we can ensure that for every feature we can calculate the sum of all d points and divide it by $\sqrt{d}$ once and use it whenever we want. On the other hand, in the standard euclidean distance, we have to calculate $(x_i - z_i)^2$ for d-dimensions. In other words, it tries to find the d-dimensional distance on a 2-dimension line which is quiet expensive. Therefore, using Jensen's inequality, We can use the left hand side equation as our distance metric as it is less expensive computationally.

Question 4

After completing the methods for the ID3 algorithm, the following results were obtained.

```
Validate accuracy on tree without pruning ========> 0.6615
Validate accuracy on tree with pruning ===========> 0.7846
Test accuracy on tree without pruning ============> 0.6733
Test accuracy on tree with pruning ===============> 0.7426
Tree size without pruning =======================>    273
Tree size with pruning ==========================>      3
Tree depth without pruning ======================>    119
Tree depth with pruning =========================>      2
```

After pruning the decision tree, the validation accuracy shows a slight improvement. The reason behind this is that pruning reduces the complextity of the decision tree and prevent overfitting. The other things that we notice here are the size and depth of the decision tree. In comparison to the normal decision tree, the pruned decision tree has significantly smaller size and depth. This is because pruning removes parts of the tree that are redundant and non critical.