Machine Learning In Action — Chapter 3

3.1 — Explaining Shannon's Entropy

Given a random variable X, the Shannon Entropy, $H(X)$ of the random variable is given by

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) \tag{1}$$

where n is the number of classes. The higher the entropy, the harder it is to predict a value. Since only the size of the Shannon Entropy is relevant for analysis, I will be changing the sign for all values.

Example 3.1: Predicting the Gender of a Child's Name

In an example where you intend to predict the gender of a child by his/her name, consider that you have 14 children names, of which 9 are male and 5 are female. Currently, the entropy, $H_1(Y)$ is

$$\begin{aligned} H_1(Y) &= -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) \\ &= -\left[\tfrac{9}{14} \cdot \log_2\left(\tfrac{9}{14}\right) + \tfrac{5}{14} \cdot \log_2\left(\tfrac{5}{14}\right)\right] \\ &= 0.94029 \end{aligned}$$

After splitting by some property, say last letter, we end up with 2 groups:

1. the group whose last letter ends with a consonant (Group 0)
2. the group whose last letter ends with a vowel (Group 1)

Group 0 has 6 males and 1 female, while Group 1 has 3 males and 4 females. The new entropy of the individual groups, $H_{2,0}(Y)$ and $H_{2,1}(Y)$ are

$$\begin{aligned} H_{2,0}(Y) &= -\left[\tfrac{6}{7} \cdot log_2(\tfrac{6}{7}) + \tfrac{1}{7} \cdot log_2(\tfrac{1}{7})\right] \\ &= 0.59167 \\ H_{2,1}(Y) &= -\left[\tfrac{3}{7} \cdot log_2(\tfrac{3}{7}) + \tfrac{4}{7} \cdot log_2(\tfrac{4}{7})\right] \\ &= 0.98523 \end{aligned}$$

Taking the weighted of the two, the final entropy after the split , $H_2(Y)$ is

$$\begin{aligned} H_2(Y) &= -\left[\tfrac{7}{14} \cdot 0.0.59167 + \tfrac{7}{14} \cdot 0.98523\right] \\ &= 0.78845 \end{aligned}$$

And so the entropy gain from this split is

$$H_2(Y) - H_1(Y) = 0.94029 - 0.78845 = 0.15184$$

By doing this split, we are able to reduce the uncertainty in the outcome by 0.1518.

Textbook Example

Given the example, letting 'Yes' to be denoted as 1 and 'No' to be denoted as 0, we have the matrix:

$$dataset = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

and the base entropy, $H_1(dataset)$ to be

$$\begin{aligned} H(dataset) &= -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) \\ &= -\left[0.4 \cdot \log_2(0.4) + 0.6 \cdot \log_2(0.6)\right] \\ &= 0.97095 \end{aligned}$$

If we split $dataset$ by the first attribute, we will get 2 matrices, $dataset_{1,1}$ and $dataset_{1,2}$ and if we split $dataset$ by the second attribute, we will get $dataset_{2,1}$ and $dataset_{2,2}$ such that

$$dataset_{1,1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} \qquad dataset_{1,2} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$dataset_{2,1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad dataset_{2,2} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

The new Shannon Entropy after the various splits, is

$$
\begin{aligned}
H(dataset_1) &= -0.6\left[\tfrac{2}{3}\cdot\log_2\left(\tfrac{2}{3}\right)+\tfrac{1}{3}\cdot\log_2\left(\tfrac{1}{3}\right)\right]-0.4\overbrace{\left[1\cdot\log_2(1)\right]}^{0}\\
&= 0.55098
\end{aligned}
$$

and

$$
\begin{aligned}
H(dataset_2) &= -0.8\left[\tfrac{2}{4}\cdot\log_2\left(\tfrac{2}{4}\right)+\tfrac{2}{4}\cdot\log_2\left(\tfrac{2}{4}\right)\right]-0.2\overbrace{\left[1\cdot\log_2(1)\right]}^{0}\\
&= 0.8
\end{aligned}
$$

The entropy gain by splitting is

$$
\begin{aligned}
\Delta H_1 &= H(dataset_1) - H(dataset_0)\\
&= -0.55098 - (-0.97095)\\
&= 0.41997
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta H_2 &= H(dataset_2) - H(dataset_0)\\
&= -0.8 - (-0.97095)\\
&= 0.17095
\end{aligned}
$$

Since $\Delta H_1 > \Delta H_2$, the first attribute is the better option of the two.