Machine Learning In Action — Chapter 3

3.1 — Explaining Shannon's Entropy

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

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

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: the group whose last letter ends with a consonant (Group 0) and 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.