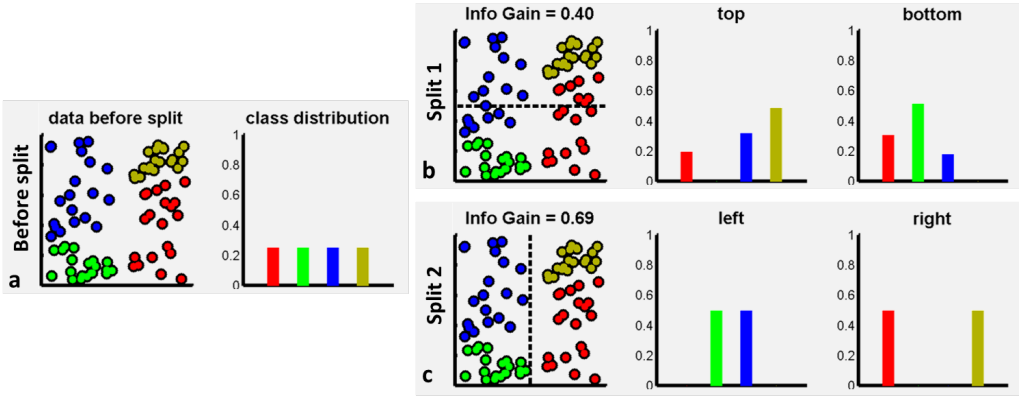

Fig. 2.3: **Information gain for discrete, non-parametric distributions. (a)** Dataset $\mathcal{S}$ before a split. **(b)** After a horizontal split. **(c)** After a vertical split.

fig. 2.4.

Figure 2.3a shows a number of data points on a 2D space. Different colours indicate different classes/groups of points. In fig. 2.3a the distribution over classes is uniform because we have exactly the same number of points in each class. If we split the data horizontally (as shown in fig. 2.3b) this produces two sets of data. Each set is associated with a lower entropy (higher information, peakier histograms). The gain of information achieved by splitting the data is computed as

$$I = H(\mathcal{S}) - \sum_{i \in \{1,2\}} \frac{|\mathcal{S}^i|}{|\mathcal{S}|} H(\mathcal{S}^i)$$

with the Shannon entropy defined mathematically as: $H(\mathcal{S}) = -\sum_{c \in \mathcal{C}} p(c) \log(p(c))$. In our example a horizontal split does not separate the data well, and yields an information gain of $I = 0.4$. When using a vertical split (such as the one in fig. 2.3c) we achieve better class separation, corresponding to lower entropy of the two resulting sets and a higher information gain ($I = 0.69$). This simple example shows how we can use information gain to select the split which produces the highest information (or confidence) in the final distributions. This concept is at the basis of the forest training algorithm.