Homogeneity and Completeness

One common task is to estimate a partition of objects. For example, clustering documents to infer the topics of the document is trying to estimate the partition of the documents by topics. The topic labels of the documents then serve as the ground truth and is used for the evaluation of the clustering.

Homogeneity and **completeness** describe how well the clustering results align with the ground truth labels from two aspects. Homogeneity measures the overall consistency of ground truth labels within the clusters; completeness measures that of cluster labels within each of the classes.

Interpretation of homogeneity score and completeness score

Information entropy

The definition of homogeneity and completeness scores "borrows" the concept of information entropy, so a little introduction to information entropy is needed. The introduction will be limited to the properties of the entropy function. The information theory behind and beyond would require a lecture to a course to elaborate.

> The information entropy H of a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ and probability mass function $P(X)$ as
>
> $$H(X) = -\sum_{i=1}^{n} P(x_i) \log\left(P(x_i)\right)$$

As a example, let's look at the binary case: $X \in \{x_1, x_2\}$, $P(x_1) = p$, $P(x_2) = 1 - p$. Then $H(X) = -p \log p - (1-p) \log(1-p)$ is a function of p. The curve is plotted below:

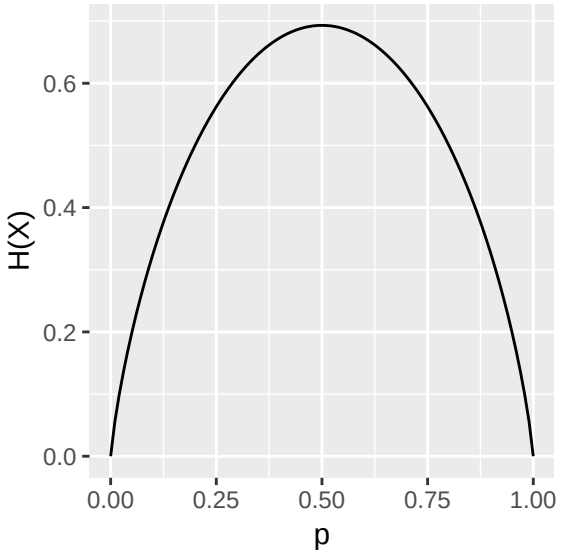

Figure 1: Entropy curve for the binary case

The entropy is maximized at $p = \frac{1}{2} = 1 - p$. Below is another plot of the entropy surface when $X \in \{x_1, x_2, x_3\}$.

Again, the entropy is maximized when $p_1 = p_2 = p_3 = 1 - p_1 - p_2$. Actually, it can be shown that the maxima of the entropy is always achieved when all possible values of X equally share the possibility, i.e.

```
ParametricPlot3D[{u, v, -u Log[u] - v Log[v] - (1 - u - v) Log[1 - u - v]}, {u, 0, 1},
 {v, 0, 1}, AxesLabel → {"p₁", "p₂", "H(X)"}]
```

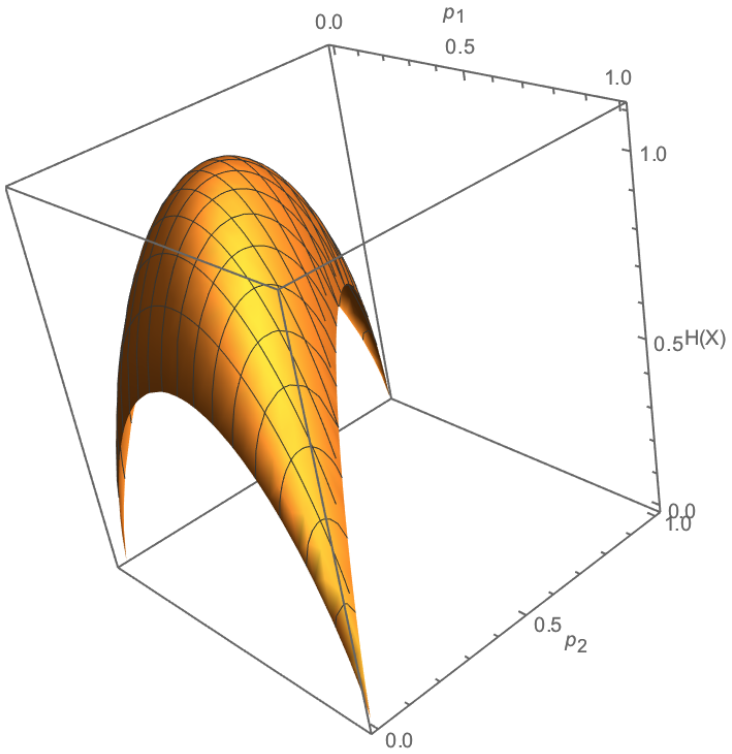

Figure 2: Entropy surface

$P(x_1) = P(x_2) = \cdots = P(x_n) = \frac{1}{n}$. On the other hand, the minimum of the entropy is obtained when one of the $P(x_i)$'s takes up all possibility and is 1, in which case the entropy is 0 (if we define $0 \times \log 0$ to be 0). This also tells us that the entropy is always non-negative.

Entropy of a partition

By extending the concept of information entropy, we can define the "entropy" of a partition of a number of objects. For example, consider N data points $\mathcal{U} = \{u_1, \dots, u_N\}$ divided into non-overlapping groups $\mathcal{C} = \{C_1, C_2, \dots\}$, i.e. $\bigcup_{i=1}^{|\mathcal{C}|} C_i = \mathcal{U}$ and $C_i \cap C_j = \varnothing, \forall i \neq j$. $\mathcal{C}$ gives a partition of the data points. We define its entropy to be

$$H(\mathcal{C}) = -\sum_{i=1}^{|\mathcal{C}|} \frac{|C_i|}{N} \log\left(\frac{|C_i|}{N}\right)$$

The properties discussed in the last section still applies to this "entropy": it is maximized when the sizes of the groups $\{C_i\}_{i=1}^{|\mathcal{C}|}$ are equal and minimized when some $C_i = \mathcal{U}$.

Conditional entropy

Suppose there is another set of non-overlapping groups $\mathcal{K} = \{K_1, K_2, \dots\}$ that divides the data points $\mathcal{U}$, s.t. $\bigcup_{i=1}^{|\mathcal{K}|} K_i = \mathcal{U}$ and $K_i \cap K_j = \varnothing, \forall i \neq j$ (we can relate this to the "communities" in the project 2). Of course we know how to calculate its entropy $H(\mathcal{K})$ already, but further, we can define "conditional entropy" using $\mathcal{C}$ and $\mathcal{K}$.

Obviously, each group K_i is divided by $\mathcal{C}$: some of the data points in K_i belongs to C_1, some to C_2, and so on. If we use $A_{ij} = |K_i \cap C_j|$ to denote the number of data points shared by K_i and C_j, we can express the entropy of the $\mathcal{C}$ *within* K_i:

$$H(\mathcal{C} \mid K_i) = -\sum_{j=1}^{|\mathcal{C}|} \frac{A_{ij}}{|K_i|} \log\left(\frac{A_{ij}}{|K_i|}\right)$$

The entropy of $\mathcal{C}$ given $\mathcal{K}$, denoted as $H(\mathcal{C} \mid \mathcal{K})$, is then defined to be the weighted sum of $H(\mathcal{C} \mid K_i)$ over i.

$$\begin{aligned} H(\mathcal{C} \mid \mathcal{K}) &= \sum_{i=1}^{|\mathcal{K}|} \frac{|K_i|}{N} H(\mathcal{C} \mid K_i) \\ &= -\sum_{i=1}^{|\mathcal{K}|} \sum_{j=1}^{|\mathcal{C}|} \frac{A_{ij}}{N} \log\left(\frac{A_{ij}}{|K_i|}\right) \end{aligned}$$

We can similarly define

$$H(\mathcal{K} \mid \mathcal{C}) = -\sum_{j=1}^{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{K}|} \frac{A_{ij}}{N} \log\left(\frac{A_{ij}}{|C_i|}\right)$$

Homogeneity and completeness score

If we use $\mathcal{C}$ as the ground truth, and want to evaluate how well $\mathcal{K}$ estimates $\mathcal{C}$, we can use homogeneity score and completeness score as measures.

Intuitively, homogeneity measures how "pure" each K_i is: ideally, each K_i only contains data points from a single $C_j \in \mathcal{C}$, and the homogeneity score is maximized to 1. Completeness measures another aspect: how complete each K_i covers C_j's. Ideally, each $C_j \in \mathcal{C}$ is a subset of some K_i, and the completeness score is maximized to 1.

Although homogeneity and completeness scores are both maximized when $\mathcal{K}$ is identical to $\mathcal{C}$, they can be respectively maximized trivially: if we assign each data point to a separate K_i, the homogeneity is maximized, although the completeness would be low; if we assign all data points to a single big K_1, the completeness is maximized, although the homogeneity is low. One must combine both homogeneity and completeness scores to evaluate $\mathcal{K}$ based on $\mathcal{C}$ comprehensively.

Homogeneity

The homogeneity score is defined to be

$$h = 1 - \frac{H(\mathcal{C} \mid \mathcal{K})}{H(\mathcal{C})}$$

It can be shown that $H(\mathcal{C} \mid \mathcal{K}) \leq H(\mathcal{C})$, and of course all entropies are non-negative, so $h \in [0, 1]$.

To see why such definition works, we consider two extreme cases: when h is maximized to 1, and when h is minimized to 0.

- When $h = 1$, $H(\mathcal{C} \mid \mathcal{K}) = 0$. Knowing that $H(\mathcal{C} \mid \mathcal{K})$ is the weighted sum of $H(\mathcal{C} \mid K_i)$'s, we have $H(\mathcal{C} \mid K_i) = 0, i = 1, \ldots, |\mathcal{K}|$. From the properties of entropy, we know this can only be obtained when in each K_i, the data points are assigned to only one $C_j \in \mathcal{C}$, which means perfect purity for each K_i.
- When $h = 0$, $H(\mathcal{C} \mid \mathcal{K}) = H(\mathcal{C})$. It is obtained when the ratio between C_j's sizes is preserved in each K_i, i.e. $\frac{A_{ij}}{A_{ik}} = \frac{|C_j|}{|C_k|}$, $i = 1, \ldots, |\mathcal{K}|$.

  Under this case, knowing a data point's assignment in $\mathcal{K}$ doesn't help us determine which $C_i \in \mathcal{C}$ it belongs to at all; in this sense, $\mathcal{K}$ doesn't provide any information about $\mathcal{C}$, and thus is the worst estimation of it.

  Using the fact that $\sum_j A_{ij} = |K_i|$, we can know the exactly value of A_{ij} in this case. $A_{ij} = |K_i| \frac{|C_j|}{N}$.

Completeness

The completeness score is defined to be

$$c = 1 - \frac{H(\mathcal{K} \mid \mathcal{C})}{H(\mathcal{K})}$$

Similarly, we know $c \in [0, 1]$, and we analyze the two extremities: when c is maximized to 1, and when c is minimized to 0.

- When $c = 1$, $H(\mathcal{K} \mid \mathcal{C}) = 0$. Knowing that $H(\mathcal{K} \mid \mathcal{C})$ is the weighted sum of $H(\mathcal{K} \mid C_i)$'s, we have $H(\mathcal{K} \mid C_i) = 0, i = 1, \ldots, |\mathcal{C}|$. From the properties of entropy, we know this can only be obtained when for each C_i, the data points are assigned to only one $K_j \in \mathcal{K}$, i.e. they are *completely* covered by a single $K_j \in \mathcal{K}$.
- When $c = 0$, $H(\mathcal{K} \mid \mathcal{C}) = H(\mathcal{K})$. It is obtained when the ratio between K_i's sizes is retained in each C_j, i.e. $\frac{A_{ij}}{A_{kj}} = \frac{|K_i|}{|K_k|}$, $j = 1, \ldots, |\mathcal{C}|$.

  Similar to the analysis in the last section, we can derive that $A_{ij} = |C_i| \frac{|K_j|}{N}$, and realize that it's exactly the same case as when $h = 0$.