Consensus clustering: proof outline

September 14, 2020

Let $X = (x_1, \ldots, x_N)$ be the observed data where X is composed of data from K distinct sets for some $N, K \in \mathbb{N}$. Let $c = (c_1, \ldots, c_N), c_i \in [1, K]$ indicate to which set the i^{th} data point belongs. c represents a partition of the data and is referred to as the true allocation vector.

We let $c^* = (c_1^*, \ldots, c_N^*)$ be the allocation vector for some random K^*-partition of the data with $K^* \in \mathbb{N}, K^* \geq K$. If we randomly select some index $i \in \mathbb{I} = (1, \ldots, N)$ and then another index $j \in J = \{j' : j' \in \mathbb{I} - \{i\} \wedge c_{j'}^* = c_i^*\}$. If it is not already the case, we then update c^* such that $c_j^* = c_i^*$.

Consider some random pair of indices, (i, j) such that $c_i \neq c_j$. In this case the probability of these being allocated a common label in any c^* is

$$p(c_i^* = c_j^*) = \frac{1}{K^*}. \tag{1}$$

Consider some random pair of indices, (i, j) such that $c_i = c_j$. In this case the probability of these being allocated a common label in any c^* is

$$p(c_i^* = c_j^*) = 2\left(1 - \frac{1}{K^*}\right)\frac{1}{N}\frac{1}{N_i - 1} + \frac{1}{K^*} \tag{2}$$

where N_i is the number of data points generated from the same function as x_i. Here $\frac{1}{N}$ is the probability of i being the first index selected, $N_i - 1$ is the size of the set $J = \{j' : j' \in \mathbb{I} - \{i\} \wedge c_{j'}^* = c_i^*\}$ and thus $\frac{1}{N_i - 1}$ is the probability of picking j from J. $(1 - \frac{1}{K^*})$ is the probability that $c_i^* \neq c_j^*$ by chance (i.e. the probability that i, j do not already have a common label by chance) and the 2 arises as i, j are exchangeable. $\frac{1}{K^*}$ is the probability of any two items being allocated together in the initial random partition (as per equation 1).

This means that if one generates R partitions using an algorithm with 1 allocation better than random, then the consensus matrix, C, in the limit as $R \to \infty$ will have entries of

$$C(i, j) = \begin{cases} 2(1 - \frac{1}{K^*})\frac{1}{N(N_i - 1)} + \frac{1}{K^*} & \text{if } c_i = c_j \\ \frac{1}{K^*} & \text{otherwise} \end{cases} \tag{3}$$

and thus a sufficiently large ensemble of learners that are better than random will uncover the true structure.