HW 2 - Clustering

Martin Kraus

February 2020

1 Theory

1.1 Given an average intra-cluster distance for clustering level j, W_j, what is the third derivative at W_j, W_j'''?

The second derivative of W_j is the rate of change for the slopes of the average distances. The point at which the greatest rate of change occurs is synonymous with the point at which k clusters fits the data best. The third derivative would be the rate at which the second derivative changes.

$$\begin{aligned}
W_j''' &= \frac{W_{j+1}'' - W_{j-1}''}{2} \\
&= \frac{\frac{W_{j+2}' - W_j'}{2} - \frac{W_{j+1}' - W_{j-1}'}{2}}{2} \\
&= \frac{\frac{\frac{W_{j+3} - W_{j+1}}{2} - \frac{W_{j+1} - W_{j-1}}{2}}{2} - \frac{\frac{W_{j+1} - W_{j-1}}{2} - \frac{W_{j-1} - W_{j-3}}{2}}{2}}{2} \\
&= \frac{\frac{W_{j+3} - 2W_{j+1} + W_{j-1}'}{4} - \frac{W_{j+1} - 2W_{j-1} - W_{j-3}}{4}}{2} \\
&= \frac{W_{j+3} - 3W_{j+1} + 3W_{j-1} - W_{j-3}}{6}
\end{aligned}$$

1.2 Given the output of your clustering algorithm as $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{5, 6, 7, 8\}$, and a hand labeled clustering of $C_1 = \{3, 4\}$, $C2 = \{1, 2, 5, 6, 7, 8\}$, what is the weighed average purity of the clusters created by the clustering algorithm?

$$Purity(C_i) = \frac{1}{|C_i|} max_j N_{ij} \tag{1}$$

$$Purity = \frac{1}{N} \sum_{i=1}^{k} |C_i| Purity(C_i) \tag{2}$$

Using equation 1 to find the cluster purity for clusters C_1 and C_2:

$$
\begin{aligned}
Purity(C_1) &= \frac{1}{4} max(2,2) \\
&= \frac{1}{4}(2) \\
&= 0.5
\end{aligned}
\qquad\qquad
\begin{aligned}
Purity(C_2) &= \frac{1}{4} max(0,4) \\
&= \frac{1}{4}(4) \\
&= 1.0
\end{aligned}
$$

Then using equation 2, the average purity is calculated:

$$
\begin{aligned}
Purity &= \frac{1}{8}(4(0.5) + 4(1.0)) \\
&= \frac{1}{8}(2+4) \\
&= \frac{6}{8} = \mathbf{0.75\%}
\end{aligned}
$$