A General Segmentation Scheme

1 The Problem

We are given, for each nucleotide position k and vector $\vec{x}_k$ of m-dimensional values (or even more generally, just some not feature x_k in an abstract metric space (X, d).

We look for an *incomplete segmentation* of the sequene positions $[1, n]$ into non-overlapping intervals.

Clustering in Metric Space

k-means style problem is well defined.

Question: given a set of points X is there a quantity in a metric space that can take the role of the centroid distance $d(x) = \|x - \bar{x}\|$ where $\bar{x}$ is the average over a cluster $\mathcal{C}$, i.e., $\bar{x} = (1/|\mathcal{C}|) \sum_{y \in \mathcal{C}} y$. On the one hand we have

$$d^2(x) = \frac{1}{|\mathcal{C}|^2} \left(\sum_{y \in \mathcal{C}} (x - y) \right)^2 = \frac{1}{|\mathcal{C}|^2} \sum_{y,z \in \mathcal{C}} (x - y)(x - z)$$

while another short computation shows the identity Consider $\sum_{y,z \in \mathcal{C}} (x - y)^2 + (x - z)^2 - (y - z)^2 = 2 \sum_{y,z \in \mathcal{C}} (x - y)(x - z)$. Hence, in a *Euclidean vector space* we can compute the centroid distance

$$d^2(x) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} d^2(x, y) - \frac{1}{2|\mathcal{C}|} \sum_{y,z \in \mathcal{C}} d^2(y, z)$$

We remark that the squared distances have an intuitive interpretation as variance constributions.

http://www.cs.cornell.edu/johannes/papers/1999/icde1999-clustering.pdf uses the same idea, sort of

2 Solution

First the feature values $\{x_k | 1 \leq k \leq n\}$ are clustered into a set of $N + 1$ clusters, $\mathcal{C}_\alpha$, where $\mathcal{C}_0$ takes the role of a nuissance or noise cluster designating the positions that should not be included in any of the final intervals.

The clustering assigns a cluster number $\alpha_k \in \{0, \dots, N\}$ to each sequence position. Positions with insufficient data are assigned to the nuissance cluster 0.

Next we define a scoring function $s(i, j, \beta)$ measuring how well the cluster β represents the measured cluster assignments $\alpha_i, \alpha_{i+1}, \dots, \alpha_k$. A good candidate is

$$s(i, j, \beta) = -M + \sum_{j=i} D(\alpha_j, \beta) \tag{1}$$

for $\beta \neq 0$ and $s(i,j,0) = 0$, some constant $M > 0$ to enforce a minimum length of an interval and $D(\alpha,\beta) = 1$ if $\alpha = \beta$ and $D(\alpha,\beta) = -a$ for $\alpha \neq \beta$.

The interval assignment that maximes the total score can be computed by dynamic programming:

$$S_{i,\alpha} = \max_{k<i} \max_{\beta\neq\alpha} S_{k-1,\beta} + s(k,i,\alpha) \tag{2}$$

with the initialization $S_{0,-1} = 0$ and $S_{0,\alpha} = -\infty$ for $0 \leq \alpha \leq N$.