Federal State Autonomous Educational Institution
for the Higher Education
National Research University "Higher School of Economics"
Faculty of Mathematics

Grachev Denis Vadimovich

Clustering of Multidimensional Random Variables to Improve HMM Sequence Alignment Accuracy

Project proposal

Scientific supervisor:
Prodanov Timofey Petrovich

Moscow 2022

Contents

1 Introduction

1.1 Clsutering

Given $X = \{x_i | x_i \in \mathbb{R}^d, i \in (1 \dots n)\}$ and $m \in \mathbb{N}$, where n is the number of points, m - number of clusters.
Clustering algorithm takes X and m and outputs $C = \{c_i | c_i \in (1 \dots m), i \in (1 \dots n)\}$.

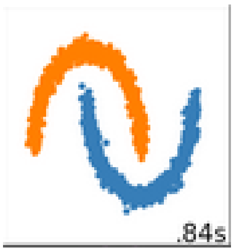

Figure 1: Example of clustering for $d = 2$, $m = 2$, color represents class.

1.2 Strings

Definition 1.1. String of length l over alphabet $A = \{1 \dots m\}$ is a map $s : \{1 \dots l\} \to A$. Usually elements of A are denoted as characters for convenience.

Definition 1.2. Alignment of strings s_1 and s_2 of lengths l_1 and l_2 respectivly, over alphabet A is a pair of strings $\hat{s}_1$ and $\hat{s}_2$ of length l over alphabet $A \sqcup \{-\}$, such that there exists increasing functions $f_i : \{1 \dots l_i\} \to \{1 \dots l\}$ such that $\hat{s}_i|_{\hat{s}_i^{-1}(A)} \circ f_i = s_i$.

Remark. $\mathrm{Im}(f_i) = \hat{s}_i^{-1}(A)$

Example 1.1. Alignment of strings $s_1 = CABCAABA$ and $s_2 = ABADBBAD$ over alphabet $\{A, B, C, D\}$.

s_1	C	A	B	C	A	A	B	A	
s_2	A	B	A	D	B	B	A		
$\hat{s}_1$	C	A	B	C	-	A	A	B	A
$\hat{s}_2$	-	A	B	-	A	D	B	B	A

Definition 1.3. For given matrix $G \in \mathbb{R}^{|A| \times |A|}$ and $p \in \mathbb{R}$ score of alignment $\hat{s}_1, \hat{s}_2$ is

$$S(\hat{s}_1, \hat{s}_2) = \sum_{i=1}^{l} \delta_i, \text{ where } \delta_i = \begin{cases} g_{\hat{s}_1(i)\hat{s}_2(i)}, & \hat{s}_1(i) \neq - \text{ and } \hat{s}_2(i) \neq - \\ p, & \end{cases}$$

Theorem 1. If G is symmetric and $g_{ij} = \begin{cases} 0, & i = j \\ > 0, & \end{cases}$ and $p > 0$, then we can define metric for strings over alphabet A as

$$d(s_1, s_2) = \min\{S(\hat{s}_1, \hat{s}_2)\}$$

Proof. $\square$

Definition 1.4. For a string s of length l, substring s_s is a string of length l_s, such that there exists an function

$$f : \{1 \dots l_s\} \to \{1 \dots l\}$$

$$f(i) = i + d$$

$$s \circ f = s_s$$

Definition 1.5. For a string s_1 and s_2 define string-substring score as

$$S_s(s_1, s_2) = \min\{S(s_s, s_2) | s_s \text{ is a substring of } s\}$$

Definition 1.6. Set of reads R for string s of length l and rate r is

$$R = \{s_s | s_s isasubstringofs,\}$$