String Set Metrics

1 Introduction

In this sketch we are interested in studying metric spaces between sets of strings.

2 Preliminaries

Definition 1 (Metric Space)**.** *A metric space* (M, d) *is a set* M *along with a distance function* $d\colon M \times M \to \mathbb{R}_{\geq 0}$ *such that for any* $x, y, z \in M$*:*

(1) $d(x, y) \geq 0$

(2) $d(x, y) = 0 \iff x = y$

(3) $d(x, y) = d(y, x)$

(4) $d(x, z) \leq d(x, y) + d(y, z)$

Definition 2 (Alphabet)**.** *An alphabet* Σ *is a finite set of unique symbols.*

Definition 3 (String)**.** *Given an alphabet* Σ*, a string* σ *is a finite sequence of symbols from* Σ*.*

Definition 4 (Alphabet Strings)**.** *Let* $\Sigma^\star$ *denote the set of all possible strings from* Σ*.*

Definition 5 (String Metric)**.** *A function* $\delta\colon \Sigma^\star \times \Sigma^\star \to \mathbb{R}_{\geq 0}$ *that satisfies metric space axioms.*

3 String Set Metric Spaces

3.1 Merging Sets

We first consider the following problem. Given a single string σ, and a set of string A, how might we calculate a distance from σ to A? Let δ be a string metric, then one idea is as follows:

$$d(\sigma, A) = \inf \{\delta(\sigma, a) \ : a \in A\}$$

The idea here is that we take the string in A that most closely resembles σ with respect to the string metric δ, and consider that the distance between σ and A.

We can take this idea further. Suppose we have two sets of strings A and B. Let us define the merge cost of A into B as follows with the $\gg\colon 2^{\Sigma^\star} \times 2^{\Sigma^\star} \to \mathbb{R}_{\geq 0}$ function:

$$A \gg B = \sum_{a \in A} \inf \{\delta(a, b) \ : b \in B\}$$

Again, for each $a \in A$, we find their individual merge cost into B. Of course one additional possibility is weighting strings by length instead of just purely summing them. But most importantly, it sure feels great to make up notation!

3.2 Bi-Directional Merging

Let's try the following:

Definition 6 (Bidirectional Merge Cost)**.** *For set of strings $A, B \subseteq 2^{\Sigma^\star}$, define the bi-directional merge cost $M\colon 2^{\Sigma^\star} \times 2^{\Sigma^\star} \to \mathbb{R}_{\geq 0}$ as follows:*

$$M(A, B) = (A \gg B) + (B \gg A)$$

For now we are only concerned about when A and B are both finite sets. Later we can try to use probability distribution style weighting to account for when A and B are infinite. Nevertheless:

Theorem 1. *The bi-directional merge cost M is a metric.*

Proof. We prove only the triangle inequality. Consider some $A, B, C \subseteq 2^{\Sigma^\star}$ and:

$$M(A, C) = \underbrace{\sum_{a \in A} \inf \{\delta(a, c) \ : c \in c\}}_{S_{A,C}} + \underbrace{\sum_{c \in C} \inf \{\delta(c, a) \ : a \in A\}}_{S_{C,A}}$$
$$M(A, B) = \underbrace{\sum_{a \in A} \inf \{\delta(a, b) \ : b \in B\}}_{S_{A,B}} + \underbrace{\sum_{b \in B} \inf \{\delta(b, a) \ : a \in A\}}_{S_{B,A}}$$
$$M(B, C) = \underbrace{\sum_{b \in B} \inf \{\delta(b, c) \ : c \in C\}}_{S_{B,C}} + \underbrace{\sum_{c \in C} \inf \{\delta(c, b) \ : b \in B\}}_{S_{C,B}}$$

Consider some pair (a, c) used in the sum $S_{A,C}$. Observe that there then exists some b_1 and b_2 such that (a, b_1) appears in the sum $S_{A,B}$ and (c, b_2) appears in the sum $S_{C,B}$. Since δ is a string metric, this naturally means that:

$$\delta(a, c) \leq \delta(a, b_1) + \delta(b_1, b_2) + \delta(b_2, c)$$

Furthermore, we can uniquely identify the pairs (a, b_1) and (c, b_2) with (a, c), since both a and c are iterated on in $S_{A,B}$ and $S_{C,B}$ respectively. Therefore, every such pair in $S_{A,C}$ can be uniquely identified with two such sets in $S_{A,B}$ and $S_{B,C}$. The consequence is then:

$$\underbrace{\sum_{a\in A} \inf \left\{\delta\left(a,c\right)\ : c\in C\right\}}_{S_{A,C}} \leq \underbrace{\sum_{a\in A} \inf \left\{\delta\left(a,b\right)\ : b\in B\right\}}_{S_{A,B}} + \underbrace{\sum_{c\in A} \inf \left\{\delta\left(c,b\right)\ : b\in B\right\}}_{S_{C,B}}$$

We mirror the argument for $S_{C,A}$, to get another inequality:

$$\underbrace{\sum_{c\in C} \inf \left\{\delta\left(c,a\right)\ : a\in A\right\}}_{S_{C,A}} \leq \underbrace{\sum_{c\in C} \inf \left\{\delta\left(c,b\right)\ : b\in B\right\}}_{S_{C,B}} + \underbrace{\sum_{a\in A} \inf \left\{\delta\left(a,b\right)\ : a\in A\right\}}_{S_{A,B}}$$

Collectively this results in the triangle inequality:

$$M\left(A,C\right) \leq M\left(A,B\right) + M\left(B,C\right)$$

$\square$