Задачи оценивания значимости выравнивания при помощи скрытых марковских моделей

Власенко Даниил Владимирович, гр.19.Б04-мм

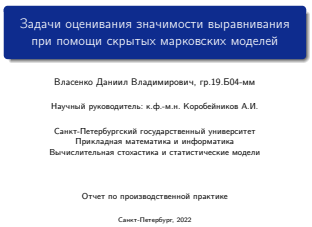
Научный руководитель: к.ф.-м.н. Коробейников А.И.

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Отчет по производственной практике

Санкт-Петербург, 2022

Оценивание значимости выравнивания



Научный руководитель к.ф.-м.н., Коробейников А.И.,
кафедра статистического моделирования

Введение

Пусть дан алфавит символов Σ.

Определение

Последовательностью длины L над алфавитом Σ будем называть такой элемент X, что $X \in \Sigma^L$. Последовательностью X над алфавитом Σ будем называть такой X, что $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.

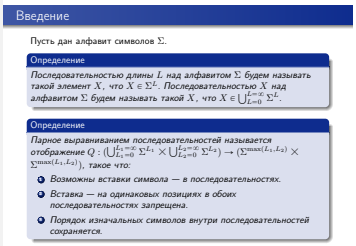
Определение

Парное выравниванием последовательностей называется отображение $Q : (\bigcup_{L_1=0}^{L_1=\infty} \Sigma^{L_1} \times \bigcup_{L_2=0}^{L_2=\infty} \Sigma^{L_2}) \to (\Sigma^{\max(L_1,L_2)} \times \Sigma^{\max(L_1,L_2)})$, такое что:

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях в обоих последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

Оценивание значимости выравнивания

└Введение



Пусть дан алфавит символов Σ.

Определение

Последовательностью длины L над алфавитом Σ будем называть такой элемент X, что $X \in \Sigma^L$. Последовательностью X над алфавитом Σ будем называть такой X, что $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.

Определение

Парное выравниванием последовательностей называется отображение $Q : (\bigcup_{L_1=0}^{L_1=\infty} \Sigma^{L_1} \times \bigcup_{L_2=0}^{L_2=\infty} \Sigma^{L_2}) \to (\Sigma^{\max(L_1,L_2)} \times \Sigma^{\max(L_1,L_2)})$, такое что:

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях в обоих последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

A C E A A F A E
C E A F D C E

A C E A A F A — E
— C E A — F D C E

Рис. 1: Последовательности до и после парного выравнивания.

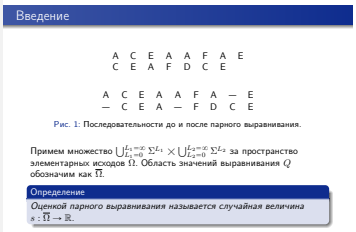
Примем множество $\bigcup_{L_1=0}^{L_1=\infty} \Sigma^{L_1} \times \bigcup_{L_2=0}^{L_2=\infty} \Sigma^{L_2}$ за пространство элементарных исходов Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

Определение

Оценкой парного выравнивания называется случайная величина $s : \overline{\Omega} \to \mathbb{R}$.

Оценивание значимости выравнивания

└Введение



Примем множество $\bigcup_{L_1=0}^{L_1=\infty} \Sigma^{L_1} \times \bigcup_{L_2=0}^{L_2=\infty} \Sigma^{L_2}$ за пространство элементарных исходов Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

Определение

Оценкой парного выравнивания называется случайная величина $s : \overline{\Omega} \to \mathbb{R}$.

Способом вычисления оценки выравнивания s может быть, например, увеличение оценки на 1 при совпадении символов, стоящих друг под другом, и уменьшение на $\frac{1}{2}$ при несовпадении. Тогда оценка s приведенного на слайде выравнивания будет равна 3.

Сходство последовательностей может отражать функциональные, структурные или эволюционные взаимосвязи объектов, которые описывают эти последовательности. Таким образом вычисление оценки выравнивания последовательностей может быть полезно в задаче определения степени родства биологических организмов путем сравнения их ДНК или РНК, нуклеотидных последовательностей, задаче анализа свойств белков, аминокислотных последовательностей, задаче распознавания речи человека или письменного языка и многих других приложениях.

Введение

Определение

Множественным выравниванием N последовательностей называется отображение $\boldsymbol{Q}: \times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty}\Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^{\max_{L\in L_i}(L)})$, такое что:

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях во всех последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

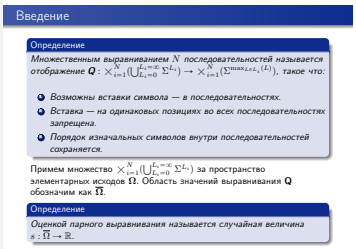
Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty}\Sigma^{L_i})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\mathbf{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой парного выравнивания называется случайная величина $s: \overline{\Omega} \to \mathbb{R}$.

Оценивание значимости выравнивания

└Введение



На предыдущем слайде приведен пример попарного выравнивания двух строк, но если сходство последовательностей слабое, то через такое выравнивание может не выйти идентифицировать взаимосвязь описываемых последовательностями объектов. В таких случаях может помочь множественное выравнивание, обобщим имеющиеся определения.

Определение

Множественным выравниванием N последовательностей называется отображение $\boldsymbol{Q}: \times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty}\Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^{\max_{L\in L_i}(L)})$, такое что:

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях во всех последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

Введение

Определение

Множественным выравниванием N последовательностей называется отображение $\boldsymbol{Q}: \times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^{\max_{L \in L_i}(L)})$, такое что:

1. *Возможны вставки символа — в последовательностях.*
2. *Вставка — на одинаковых позициях во всех последовательностях запрещена.*
3. *Порядок изначальных символов внутри последовательностей сохраняется.*

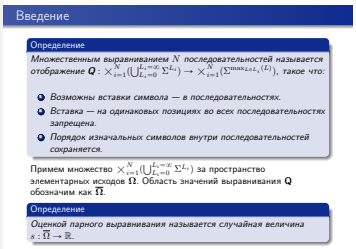
Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\mathbf{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой парного выравнивания называется случайная величина $s: \overline{\Omega} \to \mathbb{R}$.

Оценивание значимости выравнивания

└Введение



Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\mathbf{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой парного выравнивания называется случайная величина $s: \overline{\Omega} \to \mathbb{R}$.