Задачи оценивания значимости выравнивания при помощи скрытых марковских моделей

Власенко Даниил Владимирович, гр.19.Б04-мм

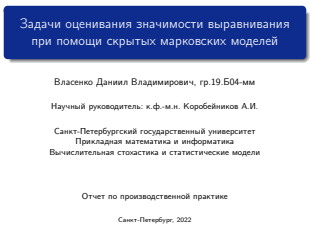
Научный руководитель: к.ф.-м.н. Коробейников А.И.

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

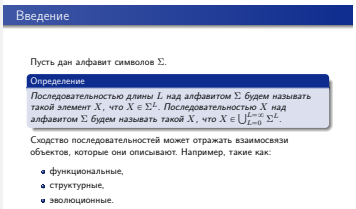
Последовательностью длины L над алфавитом Σ будем называть такой элемент X, что $X \in \Sigma^L$. Последовательностью X над алфавитом Σ будем называть такой X, что $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.

Сходство последовательностей может отражать взаимосвязи объектов, которые они описывают. Например, такие как:

- функциональные,
- структурные,
- эволюционные.

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 $X \in \bigcup_{L=0}^{L=\infty} \Sigma^L$.

Сходство последовательностей может отражать функциональные, структурные или эволюционные взаимосвязи объектов, которые описывают эти последовательности. Таким образом умение находить взаимосвязи в строках может быть приложимо 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

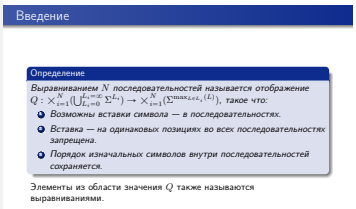
Выравниванием N последовательностей называется отображение $Q : \times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty} \Sigma^{L_i}) \to \times_{i=1}^{N}(\Sigma^{\max_{L \in L_i}(L)})$, такое что:

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

Элементы из области значения Q также называются выравниваниями.

Оценивание значимости выравнивания

└Введение



Определение

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

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

Элементы из области значения Q также называются выравниваниями.

Введение

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 выравнивания.

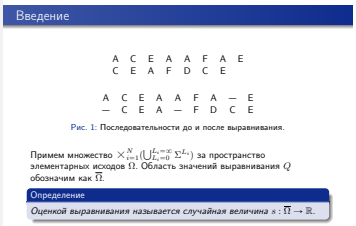
Примем множество $\times_{i=1}^{N}(\bigcup_{L_i=0}^{L_i=\infty}\Sigma^{L_i})$ за пространство элементарных исходов Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

Определение

Оценкой 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 Ω. Область значений выравнивания Q обозначим как $\overline{\Omega}$.

Определение

Оценкой 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 на одинаковых позициях в последовательностях, и уменьшение на $\frac{1}{2}$ при несовпадении. Тогда оценка s приведенного на слайде выравнивания будет равна 3.

Определить оценку выравнивания можно разными способами, но смысл будет иметь такое определение, чтобы оценка была мерой того, насколько сильно строки выравнивания похожи друга на друга.

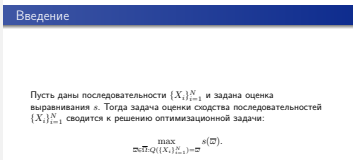
Введение

Пусть даны последовательности $\{X_i\}_{i=1}^N$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательностей $\{X_i\}_{i=1}^N$ сводится к решению оптимизационной задачи:

$$\max_{\overline{\omega}\in\overline{\Omega}:Q(\{X_i\}_{i=1}^N)=\overline{\omega}} s(\overline{\omega}).$$

Оценивание значимости выравнивания

└Введение



Пусть даны последовательности $\{X_i\}_{i=1}^N$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательностей $\{X_i\}_{i=1}^N$ сводится к решению оптимизационной задачи:

$$\max_{\overline{\omega}\in\overline{\Omega}:Q(\{X_i\}_{i=1}^N)=\overline{\omega}} s(\overline{\omega}).$$

Введение

Предположим, что даны строка X и $\omega \in \overline{\Omega}$ из N строк.

Определение

Выравниванием последовательности X к выравниванию w называется отображение $\boldsymbol{Q} : (X, \overline{\Omega}) \to \times_{i=1}^{N+1}(\Sigma^{\max_{L \in L_i}(L)})$, такое что:

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

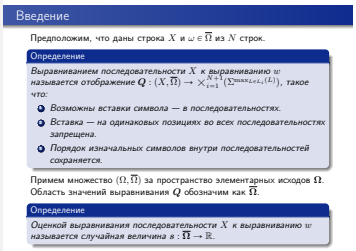
Примем множество $(\Omega, \overline{\Omega})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\boldsymbol{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой выравнивания последовательности X к выравниванию w называется случайная величина $\boldsymbol{s} : \overline{\boldsymbol{\Omega}} \to \mathbb{R}$.

Оценивание значимости выравнивания

└Введение


Введение

Предположим, что даны строка X и $\omega \in \overline{\Omega}$ из N строк.

Определение

Выравниванием последовательности X к выравниванию w называется отображение $\boldsymbol{Q} : (X, \overline{\Omega}) \to \times_{i=1}^{N+1}(\Sigma^{\max_{L \in L_i}(L)})$, такое что:

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

Примем множество $(\Omega, \overline{\Omega})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\boldsymbol{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой выравнивания последовательности X к выравниванию w называется случайная величина $\boldsymbol{s} : \overline{\boldsymbol{\Omega}} \to \mathbb{R}$.

Предположим, что даны строка X и $\omega \in \overline{\Omega}$ из N строк.

Определение

Выравниванием последовательности X к выравниванию w называется отображение $\boldsymbol{Q} : (X, \overline{\Omega}) \to \times_{i=1}^{N+1}(\Sigma^{\max_{L \in L_i}(L)})$, 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 $(\Omega, \overline{\Omega})$ за пространство элементарных исходов $\boldsymbol{\Omega}$. Область значений выравнивания $\boldsymbol{Q}$ обозначим как $\overline{\boldsymbol{\Omega}}$.

Определение

Оценкой выравнивания последовательности X к выравниванию w называется случайная величина $\boldsymbol{s} : \overline{\boldsymbol{\Omega}} \to \mathbb{R}$.

Введение

Пусть даны строка X, выравнивание $\overline{\omega} \in \overline{\Omega}$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательности X и множества, описываемого $\overline{\omega}$, сводится к решению оптимизационной задачи:

$$\max_{\overline{\boldsymbol{\omega}} \in \overline{\boldsymbol{\Omega}} : \boldsymbol{Q}(X, \overline{\omega}) = \overline{\boldsymbol{\omega}}} s(\overline{\boldsymbol{\omega}}).$$

Оценивание значимости выравнивания

└Введение

Введение

Пусть даны строка X, выравнивание $\overline{\omega} \in \overline{\Omega}$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательности X и множества, описываемого $\overline{\omega}$, сводится к решению оптимизационной задачи:

$$\max_{\overline{\boldsymbol{\omega}} \in \overline{\boldsymbol{\Omega}} : \boldsymbol{Q}(X, \overline{\omega}) = \overline{\boldsymbol{\omega}}} s(\overline{\boldsymbol{\omega}}).$$

Пусть даны строка X, выравнивание $\overline{\omega} \in \overline{\Omega}$ и задана оценка выравнивания s. Тогда задача оценки сходства последовательности X и последовательностей, описываемых $\overline{\omega}$, сводится к решению оптимизационной задачи:

$$\max_{\overline{\boldsymbol{\omega}} \in \overline{\boldsymbol{\Omega}} : \boldsymbol{Q}(X, \overline{\omega}) = \overline{\boldsymbol{\omega}}} s(\overline{\boldsymbol{\omega}}).$$

Так как последовательности описывают некоторые объекты, то введенную на предыдущем слайде задачу можно интерпретировать как оценку того, насколько сильно новая последовательность X близка к множеству последовательностей, на которых было построено выравнивание ω.

Введение

Пусть дана $X \in \Sigma$, $\overline{\omega} \in \overline{\Omega}$, s и известно, что $\overline{\omega}$ построено на последовательностях, описывающих взаимосвязанные объекты.

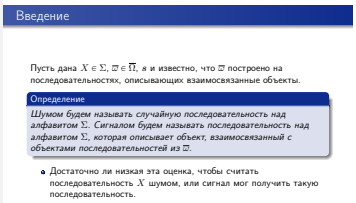
Определение

Шумом будем называть случайную последовательность над алфавитом Σ. Сигналом будем называть последовательность над алфавитом Σ, которая описывает объект, взаимосвязанный с объектами последовательностей из $\overline{\omega}$.

- Достаточно ли низкая эта оценка, чтобы считать последовательность X шумом, или сигнал мог получить такую последовательность.

Оценивание значимости выравнивания

└Введение



Встает вопрос того, как интерпретировать решение этой задачи:
Пусть дана $X \in \Sigma$, $\overline{\omega} \in \overline{\Omega}$, s и известно, что $\overline{\omega}$ построено на последовательностях, описывающих взаимосвязанные объекты.

Определение

Шумом будем называть случайную последовательность над алфавитом Σ. Сигналом будем называть последовательность над алфавитом Σ, которая описывает объект, взаимосвязанный с объектами последовательностей из $\overline{\omega}$.

- Достаточно ли низкая оценка s, чтобы считать последовательность X шумом, или сигнал мог получить такую последовательность.