Machine Learning Package

Portfolio de algoritmos de Machine Learning

Sumário

- Feature extraction consiste em calcular ou inferir novas variáveis a partir do dataset.
- No caso de sequências biológicas é comum usar descritores da composição nucleotídica (DNA) ou peptídica (aminoácidos)
- O *k-mer* é um método normalmente utilizado para calcular composições nucleotídica (DNA) ou peptídica (aminoácidos)
- No nosso portefólio, métodos de feature extraction podem seguir a estrutura de um ***Transformer***.

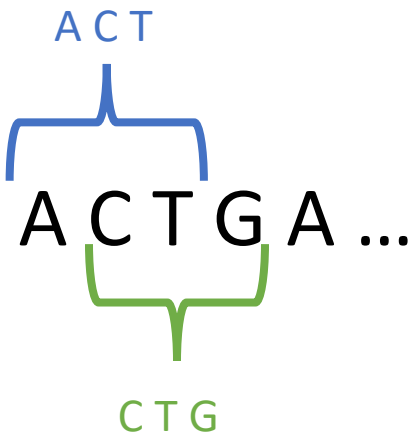
Datasets

- Os datasets estão disponíveis em:
  - https://www.dropbox.com/sh/oas4yru2r9n61hk/AADpRunbqES44W49gx9deRN5a?dl=0

k-mer

- O *k-mer* consiste no conjunto de substrings de comprimento *k* contidas numa sequência.

Por exemplo, para k=3



Objeto *KMer*

- Adiciona o sub-package *feature_extraction*, com o módulo *k_mer.py* que deve conter o objeto *KMer.*
- O *KMer* a implementa é específico para DNA (alfabeto: ACTG)
- *class KMer:*
  - Parâmetros:
    - k – tamanho da substring
  - Parâmetros estimados:
    - k_mers – todos os k-mers possíveis
  - Métodos:
    - fit – estima todos os k-mers possíveis; retorna o self (ele próprio)
    - transform – calcula a frequência normalizada de cada k-mer em cada sequência
    - fit_transform – corre o fit e depois o transform

Teste *KMer*

- *KMer* para sequências de nucleótidos:
  1. Usa o dataset *tfbs.csv*. Inspeciona o conteúdo do dataset.
  2. Usa o *KMer* para obter a frequência de cada substring em cada sequência do dataset. Tamanho da substring (k): 3
  3. Usa o *sklearn.preprocessing.StandardScaler* para standardizar o dataset da composição nucleotídica.
     dataset.X = StandardScaler().fit_transform(dataset.X)
  4. Divide o dataset em treino e teste.
  5. Treina o modelo *LogisticRegression* no dataset de composição nucleotídica.
  6. Qual o score obtido?

Avaliação

- Exercício 9: Adapta o *KMer* para calcular a composição peptídica
  - 9.1) O *KMer* deve ser capaz de calcular a composição nucleotídica e peptídica. Podes adicionar um novo parâmetro chamado *alphabet* onde o utilizador fornece o alfabeto da sequência biológica.
  - 9.2) Testa o novo *KMer* para sequências de aminoácidos:
    1. Usa o dataset ***transporters.csv***. Inspeciona o conteúdo do dataset.
    2. Usa o *KMer* para obter a frequência de cada substring em cada sequência do dataset. **Tamanho da substring (k): 2**
    3. Usa o *sklearn.preprocessing.StandardScaler* para standardizar o dataset da composição pepetídica.
       dataset.X = StandardScaler().fit_transform(dataset.X)
    4. Divide o dataset em treino e teste.
    5. Treina o modelo *LogisticRegression* no dataset de composição pepetídica.
    6. Qual o score obtido?