



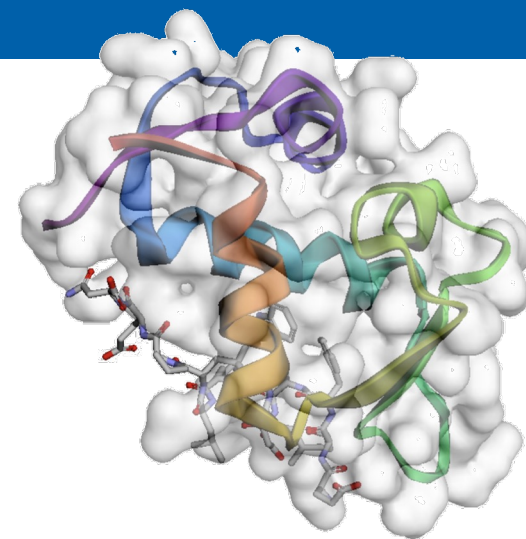
人工智能A – Project1

基于LR、SVM的蛋白质结构分类

复旦大学计算机科学技术学院

陈智能

2024/4/2





任务背景 (Referencing L1)

■ 2021年7月，Nature和Science同时发表基于人工智能的蛋白质结构预测论文

Article | [Open Access](#) | [Published: 15 July 2021](#)

Highly accurate protein structure prediction with AlphaFold



[John Jumper](#) , [Richard Evans](#), ... [Demis Hassabis](#)  [+ Show authors](#)

[Nature](#) **596**, 583–589 (2021) | [Cite this article](#)

582k Accesses | **1102** Citations | **2994** Altmetric | [Metrics](#)

Article | [Open Access](#) | [Published: 22 July 2021](#)

Highly accurate protein structure prediction for the human proteome

[Kathryn Tunyasuvunakool](#) , [Jonas Adler](#), ... [Demis Hassabis](#)  [+ Show authors](#)

[Nature](#) **596**, 590–596 (2021) | [Cite this article](#)

188k Accesses | **229** Citations | **1409** Altmetric | [Metrics](#)

高效预测几乎所有人类蛋白质结构。诺贝尔奖得主

Venki Ramakrishnan、中科院院士施一公等给出高度评价，认为这是**本世纪最重要的科学突破之一**



发展生物计算恰逢其时（百图生科董事长李彦宏2021年5月观点）

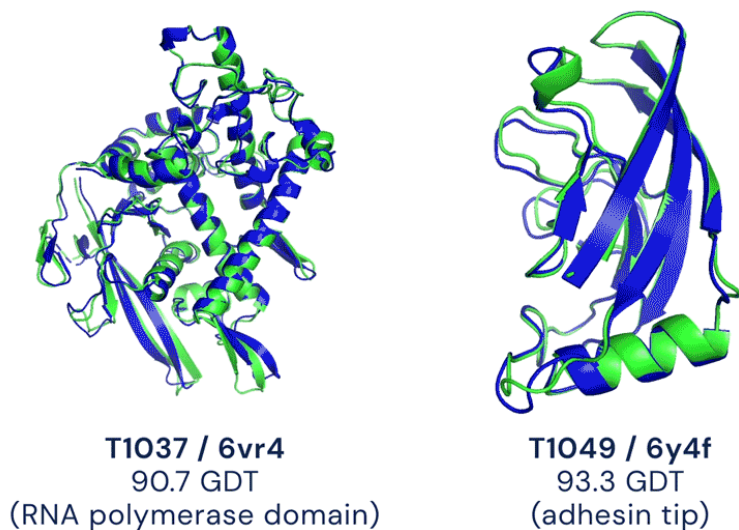
- 基因组学研究带来的人体数据在快速增长
- 新药研发过程当中所积累的知识在快速增长甚至是爆发
- 各类机器学习的算法在快速地变化、在提升和迭代

任务背景

为什么要对蛋白质进行分类?

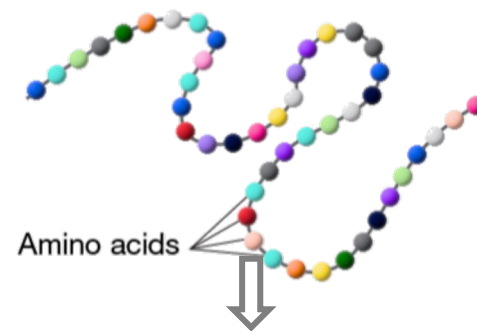
蛋白质功能预测 – 已知蛋白质序列/结构进行分类

蛋白质结构预测 – 已知蛋白质序列预测空间结构

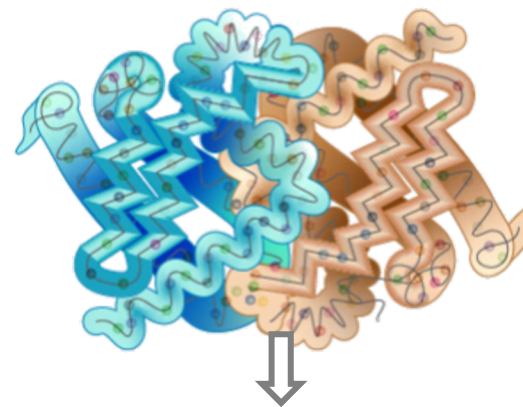


- Experimental result
- Computational prediction

➤ Sequence of Amino Acids



➤ 3-D Structure



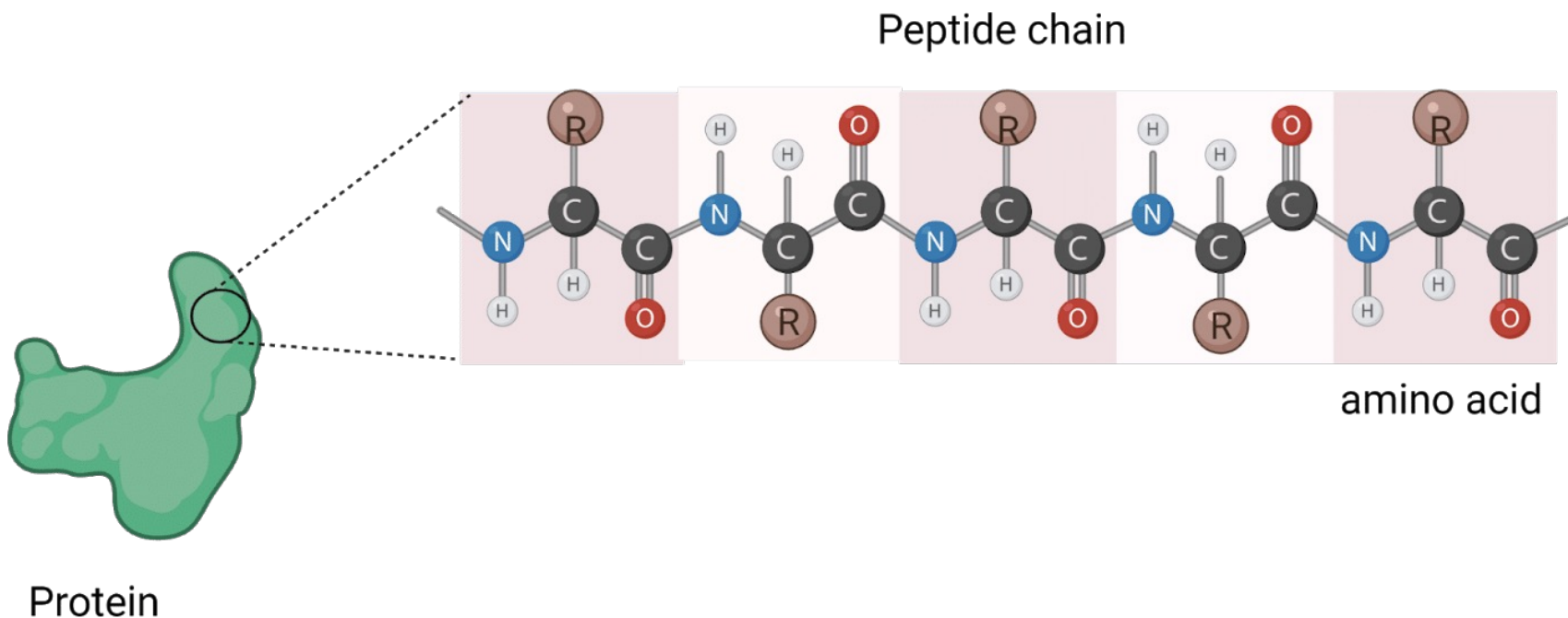
➤ Measure protein function in the lab

什么是蛋白质？

- 蛋白质由多个**氨基酸**组成

氨基酸种类不多，只有 20 种标准氨基酸。

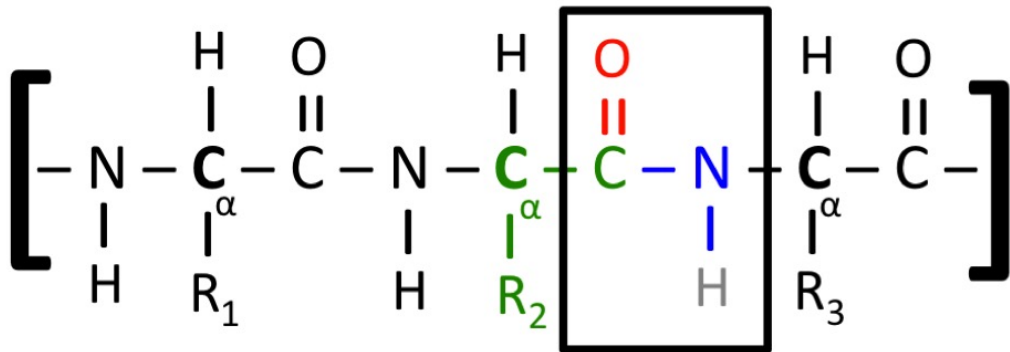
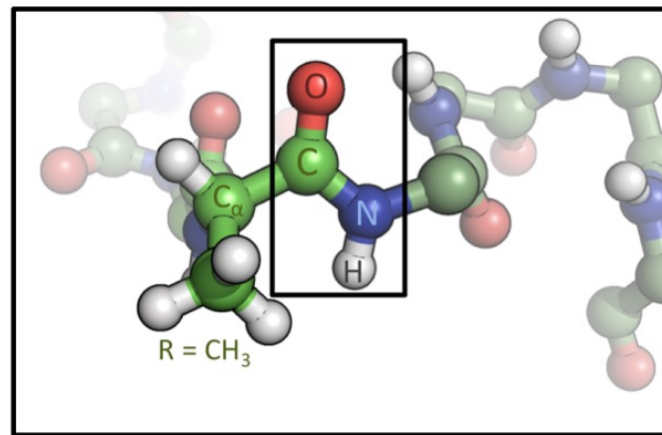
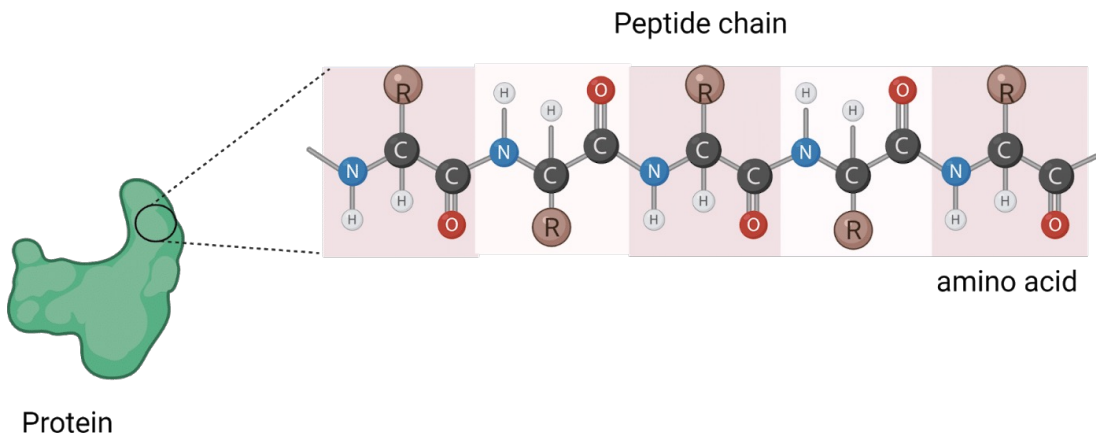
导致蛋白质巨大多样性的原因是这些氨基酸可以按任何顺序组合，而由此产生的蛋白质链可以具有截然不同的形状和功能，因为链的不同部分会粘连以及彼此折叠。



什么是蛋白质？

• 蛋白质的表示形式

所有氨基酸都包含碳 - 碳 - 氮 (C-C-N) 序列。当氨基酸融合到蛋白质中时，这种重复模式将贯穿始终，我们称为蛋白质的“骨架”。然而，氨基酸的不同之处在于它们的“侧链”，侧链指的是附着在 C-C-N 主链上的原子。



丙氨酸 (A)



什么是蛋白质?

Protein Data Bank (PDB)

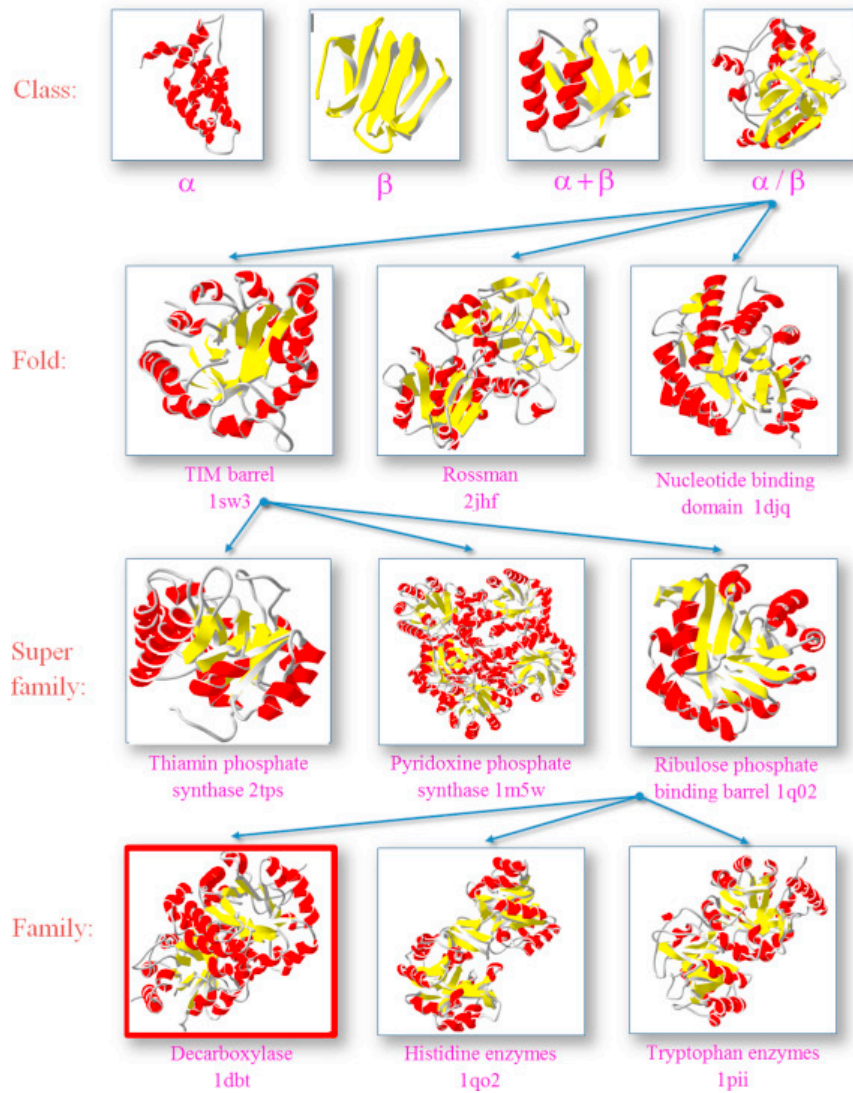
- ATOM: 表示原子条目的关键字
- 1894: 原子序列号
- C: 原子名称
- PRO: 残基名称 (在本例中为脯氨酸)
- 241: 残基序列号
- 8.855、-19.591、45.301: 分别为原子的x、y、z坐标
- 1.00: 占位
- 28.08: 温度系数
- C: 元素符号

http://pongor.itk.ppke.hu/benchmark/#/Benchmark_data_formats

data > SCOP40mini > d1a0i_1.ent

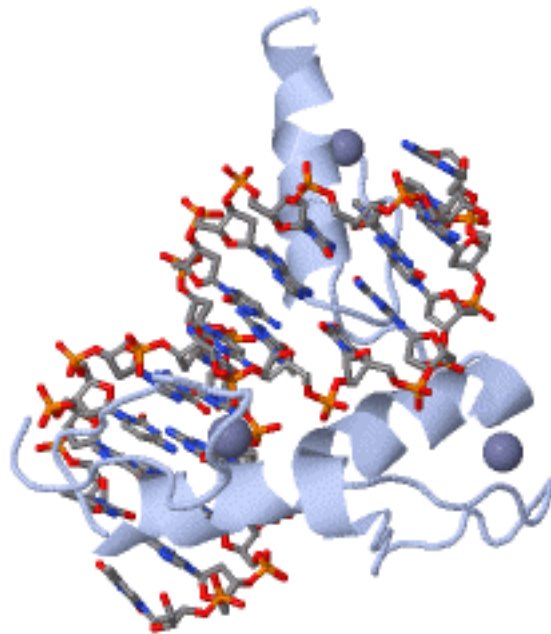
```
1  HEADER      SCOP/ASTRAL domain d1a0i_1  [25363]      28-JUL-05      0000
2  REMARK      99
3  REMARK      99 ASTRAL ASTRAL-version: 1.69
4  REMARK      99 ASTRAL SCOP-sid: d1a0i_1
5  REMARK      99 ASTRAL SCOP-sun: 25363
6  REMARK      99 ASTRAL SCOP-sccs: b.40.4.6
7  REMARK      99 ASTRAL Source-PDB: 1a0i
8  REMARK      99 ASTRAL Source-PDB-REVDAT: 25-MAR-98
9  REMARK      99 ASTRAL Region: 241-349
10 REMARK      99 ASTRAL ASTRAL-SPACI: 0.25
11 REMARK      99 ASTRAL ASTRAL-AEROSPACI: 0.25
12 REMARK      99 ASTRAL Data-updated-release: 1.61
13 ATOM      1892  N   PRO   241      10.397 -20.558  46.870  1.00 26.69      N
14 ATOM      1893  CA  PRO   241      10.311 -19.941  45.544  1.00 23.21      C
15 ATOM      1894  C   PRO   241      8.855 -19.591  45.301  1.00 28.08      C
16 ATOM      1895  O   PRO   241      8.001 -20.485  45.296  1.00 38.05      O
17 ATOM      1896  CB  PRO   241      10.841 -20.985  44.635  1.00 16.07      C
18 ATOM      1897  CG  PRO   241      11.754 -21.768  45.523  1.00 23.48      C
19 ATOM      1898  CD  PRO   241      10.862 -21.922  46.732  1.00 20.94      C
20 ATOM      1899  N   GLU   242      8.623 -18.280  45.353  1.00 17.26      N
21 ATOM      1900  CA  GLU   242      7.302 -17.700  45.248  1.00 24.00      C
22 ATOM      1901  C   GLU   242      7.146 -16.864  43.992  1.00 31.89      C
23 ATOM      1902  O   GLU   242      6.052 -16.394  43.670  1.00 34.42      O
24 ATOM      1903  CB  GLU   242      7.025 -16.857  46.506  1.00 22.55      C
25 ATOM      1904  CG  GLU   242      6.668 -17.726  47.732  1.00 28.97      C
26 ATOM      1905  CD  GLU   242      6.179 -16.952  48.949  1.00 25.42      C
27 ATOM      1906  OE1 GLU   242      6.953 -16.217  49.562  1.00 38.18      O
28 ATOM      1907  OE2 GLU   242      5.015 -17.076  49.329  1.00 24.03      O
29 ATOM      1908  N   ASN   243      8.249 -16.576  43.297  1.00 34.46      N
30 ATOM      1909  CA  ASN   243      8.160 -15.923  41.997  1.00 34.99      C
```


蛋白质分类



使用逻辑回归、SVM

根据蛋白质结构信息，完成正例/负例的分类



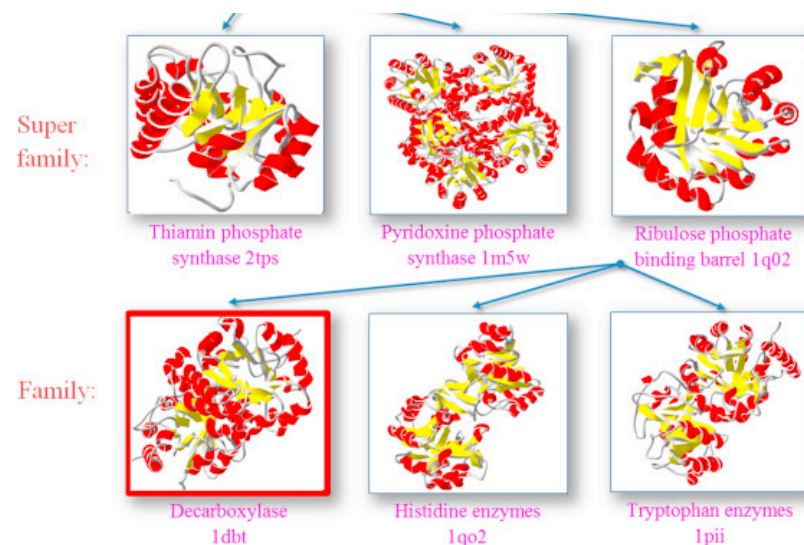
数据集介绍

PCB00019[1]

<http://pongor.itk.ppke.hu/benchmark>

SCOP40 蛋白质数据库的子集，在Family的基础上将蛋白质结构域序列和结构分类为Super Family
总共1357个蛋白质，完成55个分类任务

	Min	Max	Average
Positive Train	10	168	89
Positive Test	5	52	28
Negative Train	592	670	631
Negative test	592	671	632



[1] Sonego, Paolo, et al. "A protein classification benchmark collection for machine learning." *Nucleic acids research* 35.suppl_1 (2007): D232-D236.



代码结构

1. 数据处理
2. 模型设计
3. 模型训练
4. 验证/测试



代码结构

1. 数据处理

```
import pandas as pd

cast = pd.read_table('./SCOP40mini_sequence_minidatabase_19.cast')
```

	protein	a.118.1._a.118.1.14.	a.3.1._a.3.1.1.	a.39.1._a.39.1.2.	a.39.1._a.39.1.5.	a.4.1._a.4.1.1.	a.4.1._a.4.1.3.
0	d1c75a_	4	3	4	4	4	4
1	d1ctj_	2	3	2	2	2	2
2	d1c53_	4	3	4	4	4	4
3	d1c52_	2	3	2	2	2	2
4	d1ql3a_	4	3	4	4	4	4

"1"= +train; "2"= -train;
"3"= +test; "4"= -test



代码结构

1. 数据处理

```
import numpy as np
```

```
diagrams = np.load('./diagrams.npy')
```

```
array([[2.11734134e+03, 2.11734134e+03, 2.11734134e+03, ...,  
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],  
       [1.94871102e+03, 1.94871102e+03, 1.94871102e+03, ...,  
        0.00000000e+00, 6.87837601e-02, 0.00000000e+00],  
       [3.14334294e+03, 3.14334294e+03, 3.14334294e+03, ...,  
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],  
       ...,  
       [2.93193603e+03, 2.93193603e+03, 2.93193603e+03, ...,  
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],  
       [2.58570815e+03, 2.58570815e+03, 2.58570815e+03, ...,  
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],  
       [2.09656826e+03, 2.09656826e+03, 2.09656826e+03, ...,  
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00]])
```

(1357, 300)



```
# TODO
# 利用diagrams的特征维度数据，处理得到train_data和test_data
# 利用cast的标签数据，处理train_targets和test_targets

data_list.append((train_data, test_data))
target_list.append((train_targets, test_targets))
```

```
[[1.94871102e+03 1.94871102e+03 1.94871102e+03 ... 0.
  6.87837601e-02 0.00000000e+00]
...
[3474.29834223 3474.29834223 3474.29834223 ... 0.
  0. 0. ]
[2585.70815122 2585.70815122 2585.70815122 ... 0.
  0. 0. ]]
```

[illegible]



代码结构

2. 模型设计 – 模型训练/测试

```
from sklearn.svm import SVC

class SVMModel:
    def __init__(self, C=1.0):
        self.model = SVC(kernel=kernel, C=C, max_iter=10000)

    def train(self, train_data, train_targets):
        self.model.fit(train_data, train_targets)

    def evaluate(self, data, targets):
        return self.model.score(data, targets)

model = SVMModel(kernel=args.kernel, C=args.C)
model.train(train_data, train_targets)
test_accuracy = model.evaluate(test_data, test_targets)
```




代码结构

2. 模型设计 – 模型训练/测试

```
class LinearSVMModel:  
# TODO  
# 实现线性 SVM  
  
class LRModel:  
# TODO  
# 实现Logistic Regression
```

```
# TODO  
# 分析讨论SVM的核函数、正则化系数的影响
```



实验要求

1. 完善代码实现蛋白质分类（数据读取） - 4分
2. 增加实验分析线性SVM和其他机器学习方法（如LR） - 2分
3. 分析讨论SVM的核函数、正则化系数的影响 - 2分 枚举，控制变量
4. *特征工程：引入一些基本的特征工程技巧，例如如何从蛋白质结构数据中考虑使用距离，删除负样本
提取有用的特征，或者如何使用特征选择来减少维度 - 2分

PJ1 截止日期: 2024年4月7日13:30 !!

总分: 10分 提交: 【代码+实验报告（上限4页）】至elearning



*特征工程

1. 序列特征

- **氨基酸组成**: 计算蛋白质序列中各氨基酸的比例或出现频次。
- **物理化学属性**: 根据氨基酸的物理化学属性 (如亲水性、疏水性、电荷) 进行编码。
- **序列模体 (Motifs) 和保守区域**: 识别序列中的功能性模体或保守序列, 这些区域可能与蛋白质的功能紧密相关。

2. 结构特征

- **二级结构**: 蛋白质的二级结构 (如 α -螺旋、 β -折叠) 可以通过各种预测工具获得, 并作为特征使用。
- **接触图**: 基于蛋白质三维结构, 计算氨基酸残基之间的距离, 生成接触图或距离矩阵。
- **溶剂可及表面积 (SASA)**: 计算蛋白质中每个氨基酸残基的溶剂可及表面积, 反映其暴露程度。
- **折叠能量**: 使用各种工具计算蛋白质的理论折叠能量, 这可以反映蛋白质稳定性的不同方面。



*特征工程

3. 功能域和位点

- **功能域**：通过数据库如Pfam识别蛋白质中已知的功能域，这些域通常与特定的生物学功能相关联。
- **活性位点**：识别蛋白质中的活性位点或结合位点，这些位点通常对蛋白质的功能至关重要。

4. 特征选择和降维

- **Permutation Importance**：通过随机打乱每个特征的值，并观察这对模型性能的影响来衡量特征的重要性。
- **主成分分析 (PCA)**：用于减少特征维度，同时尽可能保留数据的变异性。
- **t-分布随机邻域嵌入 (t-SNE)**：用于高维数据的可视化，可以帮助识别数据中的模式或群组。
- **自动编码器**：深度学习方法，可以用于学习数据的低维表示。
- **特征选择技术**：例如递归特征消除 (RFE) 或基于模型的特征选择方法（如使用随机森林的特征重要性），可以用来识别最有预测力的特征。

THANKS

陈智能

2024/4/2