

MATHEMATICS DEPARTMENT

BIostatISTICS

FALL 2020

Classifying Iris - Lab 3

Kerry-Ann Bartley Donaldson

Email: kbartleydona2019@fau.edu

25th November 2020

Contents

1	Introduction	2
2	The Dataset	2
3	Python Codes for Analysis	2

1 Introduction

Turning Biology into Mathematics

In this lab demonstration we will access the Iris data set. This dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

2 The Dataset

The one-letter and three-letter codes for amino acids used in the knowledgebase are those adopted by the commission on Biochemical Nomenclature of the IUPAC-IUB

One-letter code	Three-letter code	Amino-acid name
-----------------	-------------------	-----------------

A	Ala	Alanine
---	-----	---------

R	Arg	Arginine
---	-----	----------

N	Asn	Asparagine
---	-----	------------

D	Asp	Aspartic acid
---	-----	---------------

C	Cys	Cysteine
---	-----	----------

Q	Gln	Glutamine
---	-----	-----------

E	Glu	Glutamic acid
---	-----	---------------

G	Gly	Glycine
---	-----	---------

H	His	Histidine
---	-----	-----------

I	Ile	Isoleucine
---	-----	------------

L	Leu	Leucine
---	-----	---------

K	Lys	Lysine
---	-----	--------

M	Met	Methionine
---	-----	------------

F	Phe	Phenylalanine
---	-----	---------------

P	Pro	Proline
---	-----	---------

S	Ser	Serine
---	-----	--------

T	Thr	Threonine
---	-----	-----------

W	Trp	Tryptophan
---	-----	------------

Y	Tyr	Tyrosine
---	-----	----------

V	Val	Valine
---	-----	--------

O	Pyl	Pyrrolysine
---	-----	-------------

U	Sec	Selenocysteine
---	-----	----------------

B	Asx	Aspartic acid or Asparagine
---	-----	-----------------------------

Z	Glx	Glutamic acid or Glutamine
---	-----	----------------------------

X	Xaa	Any amino acid
---	-----	----------------

3 Python Codes for Analysis

```

1 !pip install git+https://github.com/williamwardhahn/mpcr #
  installing pip from instructor github
2 from mpcr import *

1 # This code will create a dataset from the uniprot database
2 X, Y = get_uniprot_data('antibody', '!antibody', 2000)
3 # create dataset with 2000 samples

1 number_X = len(X)          #Assigning name to the length of X and Y
  from the dataset
2 number_Y = len(Y)

1 print(number_X)            #printing out the length
2 print(number_Y)

1 print(len(X[0]), print(len(Y[0]))    #printing out the length of the 1
  st protein in X (antibodies) and Y (antibodies)
2 print(len(X[1]), print(len(Y[1]))
3 print(len(X[2]), print(len(Y[2]))
4 print(len(X[3]), print(len(Y[3]))
5 print(len(X[1999]), print(len(Y[1999]))

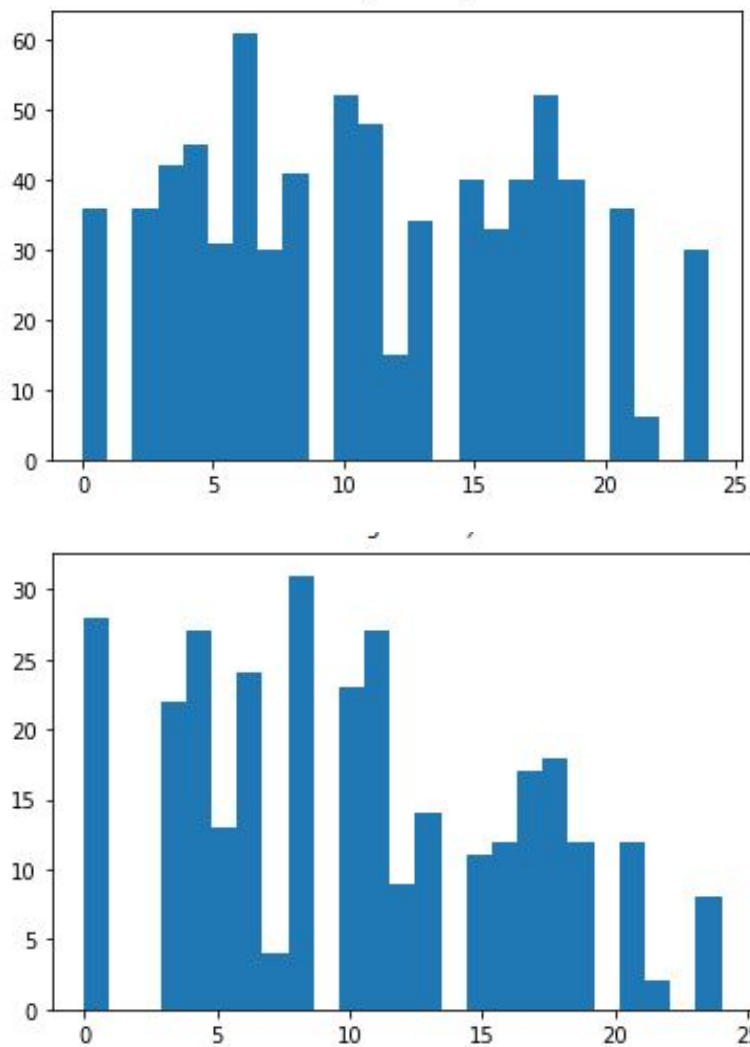
1 X[0] #Amino acid sequence of the first protein on the list of proteins
  associated with 'antibody'

1 def process_strings(c):
2     '''Takes in a list of sequences 'c' and turns each one
3     into a list of numbers.'''
4
5     X = []
6
7     for m, seq in enumerate(c):
8         x = []
9         for letter in seq:
10            x.append(max(ord(letter)-97, 0))
11
12        X.append(x)
13
14    return X

1 X = process_strings(X)      #Assigning names
2 Y = process_strings(Y)

1 plt.hist(X[0],25)          #plotting histogram of the 1st X protein, 25 of
  them
2 plt.hist(Y[0],25)          #plotting histogram of the 1st Y protein, 25 of them

```



```
1 np.mean(X[0]),np.std(X[0]) #meaningful? Finding the mean and standard
  deviation of the 1st proteins
```

```
1 np.array(X[0]).shape      #size of the first protein in X
```

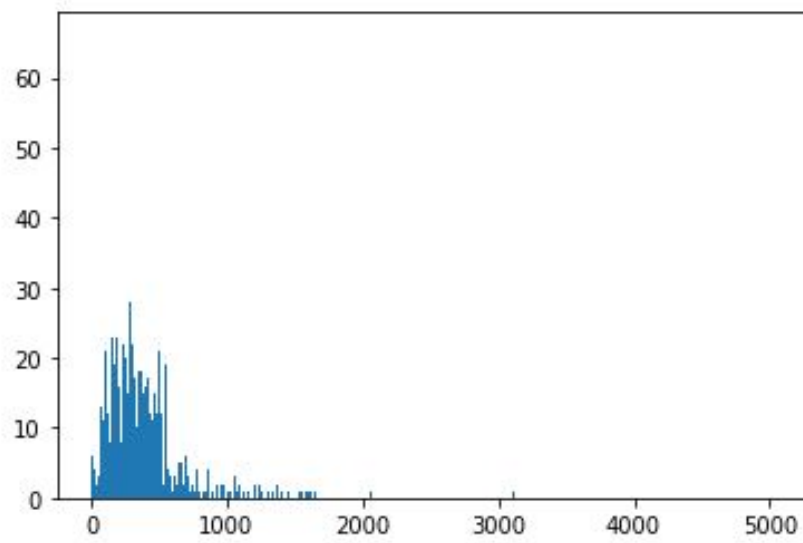
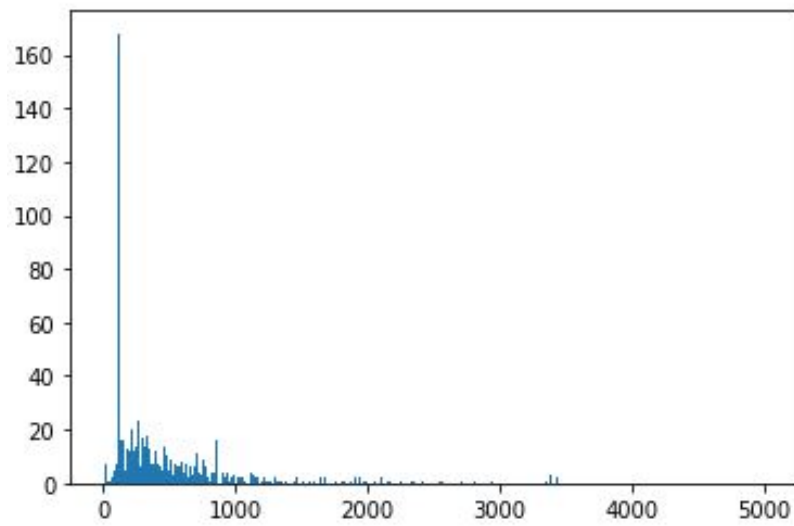
Find lengths of all proteins:

```
1 X_lengths = [len(s) for s in X]
2 Y_lengths = [len(s) for s in Y]
```

```
1 np.max(X_lengths) #Max length in X
2 np.max(Y_lengths) #Max length in Y
```

```
1 np.min(X_lengths) #min length in X
2 np.min(Y_lengths) #min length in Y
```

```
1 plt.hist(X_lengths,bins=1000,range=(0,5000)); #plot a histogram
  of X lengths with intervals 1000 from 0-5000
2 plt.hist(Y_lengths,bins=1000,range=(0,5000)); #plot a histogram
  of Y lengths with intervals 1000 from 0-5000
```



K. Bartley Donaldson