Classifying Codon Frequencies using a Neural Network Classifier

Shannon Stoehr

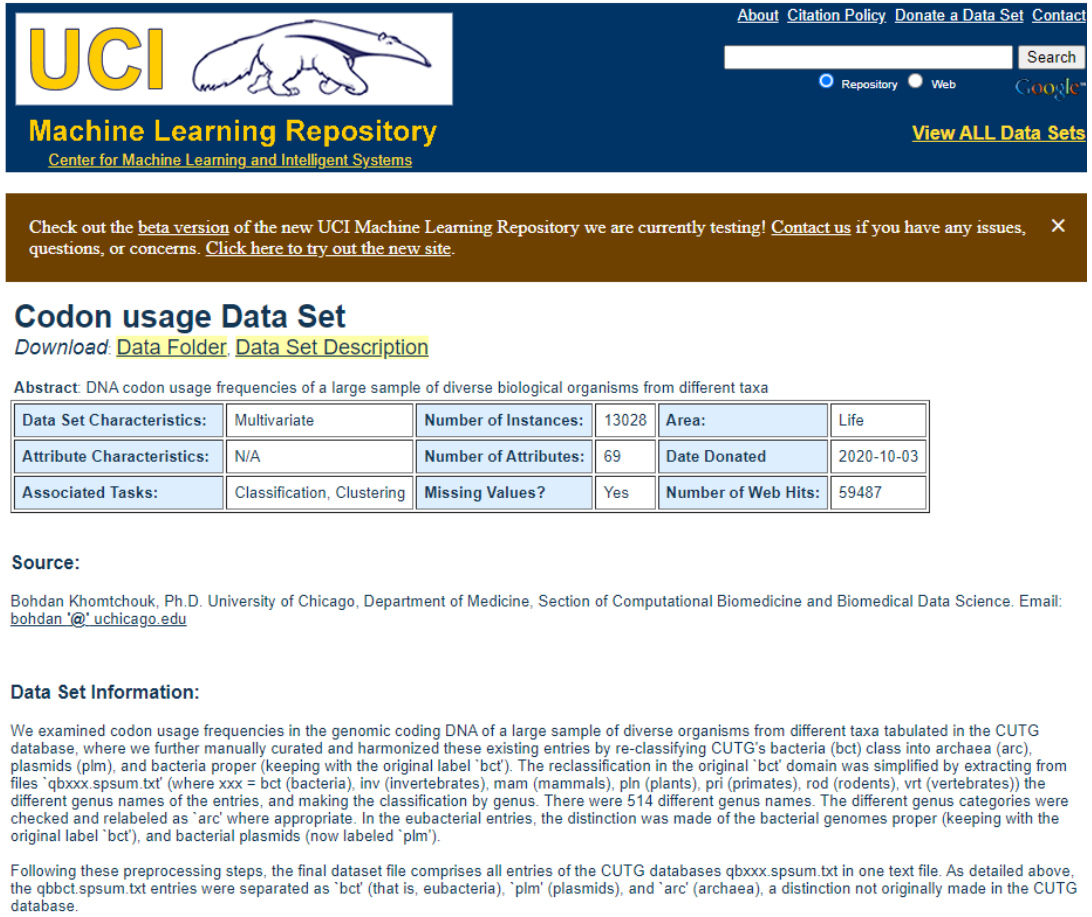
CS 445 Final Project

WHAT ARE CODONS?

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr **UAA Stop** **UAG Stop**	UGU, UGC } Cys **UGA Stop** UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile **AUG Met**	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

- Central Dogma:
  - DNA ☐ RNA ☐ protein
- Nucleotide "alphabet":
  - adenine (A)
  - cytosine (C)
  - guanine (G)
  - thymine (T, DNA) or uracil (U, RNA)
- Codon: nucleotide triplet
  - $4^3 = 64$ combinations
  - Each codon is translated into an amino acid or a release factor (stop)

DATA SET



About Citation Policy Donate a Data Set Contact

Search

Repository Web

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Codon usage Data Set

Download: Data Folder, Data Set Description

Abstract: DNA codon usage frequencies of a large sample of diverse biological organisms from different taxa

Data Set Characteristics:	Multivariate	**Number of Instances:**	13028	**Area:**	Life
Attribute Characteristics:	N/A	**Number of Attributes:**	69	**Date Donated**	2020-10-03
Associated Tasks:	Classification, Clustering	**Missing Values?**	Yes	**Number of Web Hits:**	59487

Source:

Bohdan Khomtchouk, Ph.D. University of Chicago, Department of Medicine, Section of Computational Biomedicine and Biomedical Data Science. Email: bohdan '@' uchicago.edu

Data Set Information:

We examined codon usage frequencies in the genomic coding DNA of a large sample of diverse organisms from different taxa tabulated in the CUTG database, where we further manually curated and harmonized these existing entries by re-classifying CUTG's bacteria (bct) class into archaea (arc), plasmids (plm), and bacteria proper (keeping with the original label `bct'). The reclassification in the original `bct' domain was simplified by extracting from files `qbxxx.spsum.txt' (where xxx = bct (bacteria), inv (invertebrates), mam (mammals), pln (plants), pri (primates), rod (rodents), vrt (vertebrates)) the different genus names of the entries, and making the classification by genus. There were 514 different genus names. The different genus categories were checked and relabeled as `arc' where appropriate. In the eubacterial entries, the distinction was made of the bacterial genomes proper (keeping with the original label `bct'), and bacterial plasmids (now labeled `plm').

Following these preprocessing steps, the final dataset file comprises all entries of the CUTG databases qbxxx.spsum.txt in one text file. As detailed above, the qbbct.spsum.txt entries were separated as `bct' (that is, eubacteria), `plm' (plasmids), and `arc' (archaea), a distinction not originally made in the CUTG database.

- Codon usage Data Set
  - Bohdan Khomtchouk, Ph.D. University of Chicago, Department of Medicine, Section of Computational Biomedicine and Biomedical Data Science. Email: bohdan@uchicago.edu
- Includes…
  - Kingdom
  - DNA type
  - Species (approx. 13,000 total)
  - Total number of codons
  - Frequency of each of 64 codons

NEURAL NETWORK CLASSIFIER

A4 Classification of Hand-Drawn Digits

In this assignment, you will define a new class named `NeuralNetworkClassifier` that extends the `NeuralNetwork` class provided here and is the solution to Assignment A2. You will use `NeuralNetworkClassifier` to train a classifier of hand-drawn digits.

You will also define the function `confusion_matrix`.

`NeuralNetwork` class

In [1]:
```
import matplotlib.pyplot as plt
```

The following code cell will write its contents to `optimizers.py` so the `import optimizers` statement in the code cell after it will work correctly.

In [2]:
```
%%writefile optimizers.py
import numpy as np

######################################################################
## class Optimizers()
######################################################################

class Optimizers():

    def __init__(self, all_weights):
        '''all_weights is a vector of all of a neural networks weights concatena

        self.all_weights = all_weights

        # The following initializations are only used by adam.
        # Only initializing m, v, beta1t and beta2t here allows multiple calls t
        # with multiple subsets (batches) of training data.
        self.mt = np.zeros_like(all_weights)
        self.vt = np.zeros_like(all_weights)
        self.beta1 = 0.9
        self.beta2 = 0.999
        self.beta1t = 1
        self.beta2t = 1
```

- Based off Assignment 4
- Experimenting with different hidden layer structures
- Two tests based on codon frequencies:
  - How well does the classifier predict kingdom?
  - How well does the classifier predict DNA type?
- Questions:
  - Is there a particular codon that is statistically significant in predicting a species' kingdom and/or DNA type?

PROGRESS

jupyter Stoehr CS445 Final Project (unsaved changes)

Logout

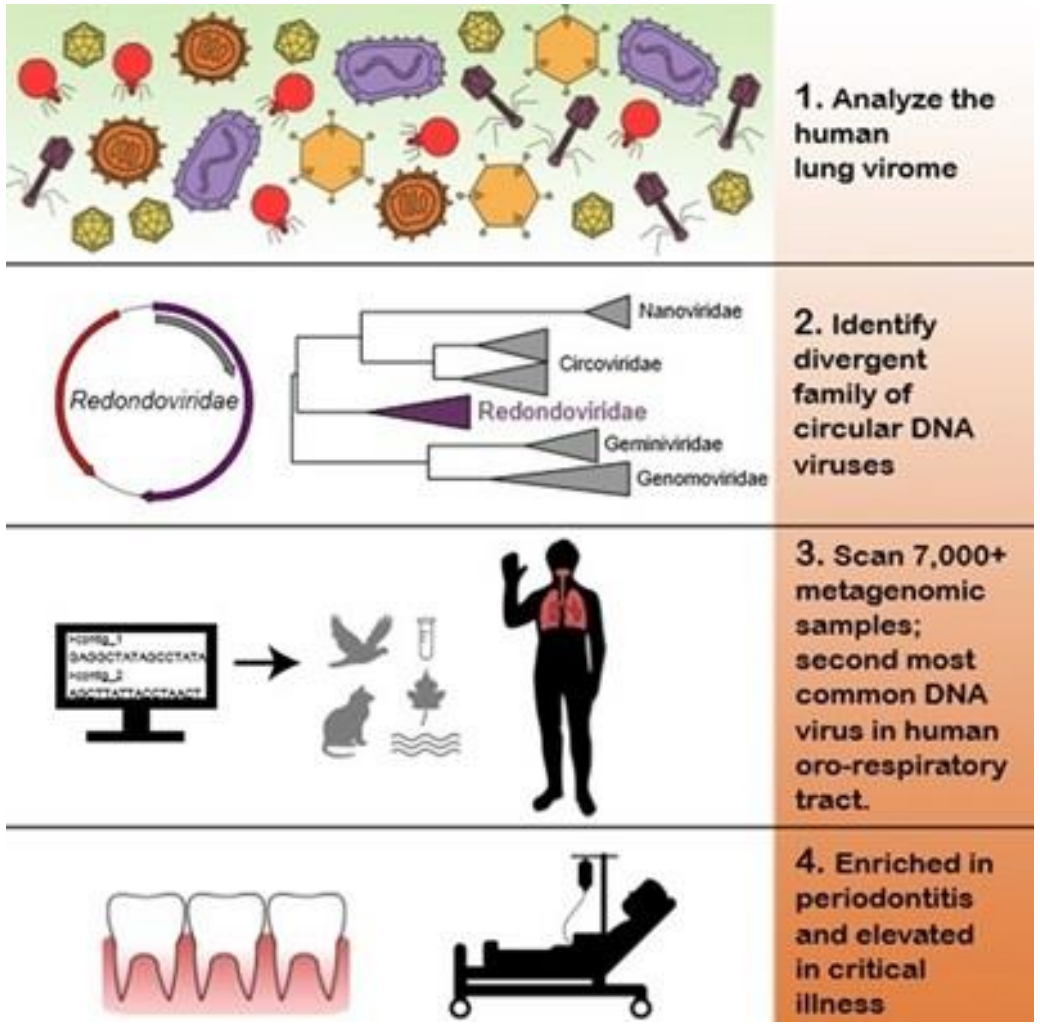
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
Adam: Epoch 100 Error=0.87059
Adam: Epoch 150 Error=0.87669
Adam: Epoch 200 Error=0.87899
Adam: Epoch 250 Error=0.88009
Adam: Epoch 300 Error=0.88078
Adam: Epoch 350 Error=0.88124
Adam: Epoch 400 Error=0.88162
Adam: Epoch 450 Error=0.88199
Adam: Epoch 500 Error=0.88243
Adam: Epoch 50 Error=0.88735
Adam: Epoch 100 Error=0.91203
Adam: Epoch 150 Error=0.92136
Adam: Epoch 200 Error=0.93166
Adam: Epoch 250 Error=0.93713
Adam: Epoch 300 Error=0.93944
Adam: Epoch 350 Error=0.94069
Adam: Epoch 400 Error=0.94144
Adam: Epoch 450 Error=0.94196
Adam: Epoch 500 Error=0.94244
Adam: Epoch 50 Error=0.90989
Adam: Epoch 100 Error=0.93495
Adam: Epoch 150 Error=0.94303
Adam: Epoch 200 Error=0.94776
Adam: Epoch 250 Error=0.95042
Adam: Epoch 300 Error=0.95168
Adam: Epoch 350 Error=0.95249
Adam: Epoch 400 Error=0.95304
Adam: Epoch 450 Error=0.95353
Adam: Epoch 500 Error=0.95403
Adam: Epoch 50 Error=0.93489
Adam: Epoch 100 Error=0.95315
Adam: Epoch 150 Error=0.96082
Adam: Epoch 200 Error=0.96454
Adam: Epoch 250 Error=0.96669
Adam: Epoch 300 Error=0.96822
Adam: Epoch 350 Error=0.96948
Adam: Epoch 400 Error=0.97047
Adam: Epoch 450 Error=0.97131
Adam: Epoch 500 Error=0.97205
Adam: Epoch 50 Error=0.90397
Adam: Epoch 100 Error=0.93796
Adam: Epoch 150 Error=0.95689
Adam: Epoch 200 Error=0.96503
Adam: Epoch 250 Error=0.96888
Adam: Epoch 300 Error=0.97090
Adam: Epoch 350 Error=0.97250
Adam: Epoch 400 Error=0.97391
Adam: Epoch 450 Error=0.97472
Adam: Epoch 500 Error=0.97561
```

- Some say it's still running to this day…
- I'm working on analysis now!

APPLICATIONS



- Classifying unknown DNA
  - Finding close genetic relatives
  - In the case of new viruses:
    - Identifying origin
    - Modifying an existing effective vaccine
- "The team knew they might have discovered a virus because the sequence of DNA building blocks – that eventually form proteins – allowed them to recognize these as distant relatives of known viral molecules, which are important for making the virus particle shell and managing replication."

  —*Penn Medicine, 2019*

RESOURCES

- https://openstax.org/books/biology/pages/15-1-the-genetic-code
- https://archive.ics.uci.edu/ml/datasets/Codon+usage
- Assignment 2: Multilayer Neural Networks for Nonlinear Regression
- Assignment 4: Classification of Hand-Drawn Digits
- Lecture 7.2: Optimizers, Data Partitioning, Finding Good Parameters
- https://www.pennmedicine.org/news/news-releases/2019/may/how-do-you-find-a-virus-thats-completely-unknown-study-says-look-to-the-genome

Thank you!

Shannon Stoehr

CS 445 Final Project