Computational Science – Machine Learning for Physiscists
Project: Unsupervised and supervised analysis of protein sequences

Martin Weigt, Sorbonne Université
(Dated: October 14, 2021)

The aim of the project is the application of some of the basic ML methods discussed during the lectures, using real data. The data are protein-sequence data, which have been recently elaborated by my team together with a number of experimental biologists (no worry, no prior biological knowledge is needed). The data and some overlapping analysis of this project are published in

- Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M. and Ranganathan, R., 2020. An evolution-based model for designing chorismate mutase enzymes. Science, 369(6502), pp.440-445.

The paper is provided together with the data in the usual DropSU folder.

Data are provided as multiple-sequence alignments (MSA), *i.e.* as rectangular arrays, where each row is a protein sequence, each column an aligned position. The entries are either one of the 20 natural amino acids $\{\mathtt{A},\mathtt{C},\mathtt{D},\mathtt{E},\mathtt{F},\mathtt{G},\mathtt{H},\mathtt{I},\mathtt{K},\mathtt{L},\mathtt{M},\mathtt{N},\mathtt{P},\mathtt{Q},\mathtt{R},\mathtt{S},\mathtt{T},\mathtt{V},\mathtt{W},\mathtt{Y}\}$ or the alignment gap "−", *i.e.* the variables are 21-state categorical variables. The format of the datafiles is the so-called Fasta file format:

```
> sequence_1 functional_true
−TSENPLLALREKISALDEKLLALLAERRELAVEVGKAKLLSHRPVRDE...
> sequence_2 functional_false
VENNDKINKLRTQIDPLDHKIIEDLGKRMKIADEIGELKKEQNVAVLQAK...
```

The line starting with ">" is a comment line, containing arbitrary information about the following sequence. In our case it is just a protein identifier (simplified in our files) and, more importantly, an information if the protein is functioning or not in an experimental screen. The next line(s) until the next ">" contain(s) one aligned amino-acid sequence.

There are three data files. The first consists of natural sequences (cf. my presentation), which have been annotated experimentally. The second contains artificial sequences, which were generated by a generative model learned on the natural MSA, and which are also annotated. The third contains a larger number of natural sequences without any annotation.

Task 1: One-hot encoding of protein sequence data
As discussed in the lectures, categorical variables are frequently represented in one-hot encoding, *i.e.* as vectors containing one entry equal to 1, and all the other equal to 0. In the case of protein data, a little variant is useful: You may use a 20-dimensional representation with $\mathtt{A}\to(1,0,...,0)$, $\mathtt{C}\to(0,1,0,...,0)$, ..., $\mathtt{Y}\to(0,...,0,1)$, while the gap is mapped to the zero-vector, $-\to(0,...,0)$. Note that the one-hot encoding blows up the feature vectors from $L=96$ categorical variables to $20L=2920$ binary variables, but the numerical treatment is easier.

Task 2: Dimensional reduction and visualization of sequence space
Use PCA of the natural data (first dataset) in one-hot encoding, to determine the first few principle components (PCs) of the dataset. Project the sequences onto PCs and represent graphically the dimensionally reduced data. What do you observe? Color sequences according to their functionality. Are functional and non-functional sequences well sperarated in PCA space? Project also the generated sequences and the natural sequences from the third dataset onto the PCs determined from the natural data. Do they occupy a similar region in (dimensionally reduced) sequence space? Discuss your observations!

Task 3: Unsupervised analysis of data
Use unsupervised machine learning (e.g. clustering, generative modeling...) to explore the organization of the sequences in the three datasets. Possible questions might be if functional and non-functional sequences are well separated in sequence space, if natural and generated sequences are well separated etc.

Task 4: Supervised analysis of data
Use a classifier of your choice (e.g. logistic regression, random forest, neural networks etc.) to learn a functional

mapping sequences to functionality, according to the label provided in the first (and/or second) dataset. Explore the predictive capacities of the classifier, use them to annotate non-annotated sequences. Discuss the results! What can we learn from the labels, which was impossible to discover in Task 3?

Solving these tasks may strongly benefit from the provided Jupyter notebooks. Please prepare a notebook with your implementations, analyses and results. You can work in pairs of two students. Examinations are individual, you will have 10min to present your work (be concise and respect the time, please) followed by questions about the underlying methods and your findings. The notebooks with your code, and the pdf of you presentation have to be sent the day before the examination to me by email.