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

Martin Weigt, Sorbonne Université
(Dated: October 9, 2023)

The aim of the project is the application of some of the basic ML methods, which we have discussed in our lectures, using real data. The data are protein-sequence data, which have been elaborated by my team together with a number of experimental biologists (no worry, no prior biological knowledge is needed beyond the introduction given in the lecture). 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 a shared DropSU folder. The folder contains also the PDF of the introduction to the subject given during the lectures.

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 two data files. The first consists of natural sequences (cf. my presentation), which we consider to be the training set, and the object of our interest for most of the unsupervised learning. The second contains artificial sequences, which were generated by a generative model learned on the natural MSA. Both files have the same form, both files contain the functionality, which was determined in *in vivo* experiments.

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 in one-hot encoding, to determine the first few principle components (PCs) of the dataset. Project 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 onto the PCs determined from the natural data. Do they occupy a similar region in (dimensionally reduced) sequence space?

Task 3: Clustering sequence data
Use a clustering algorithm of your choice (k-means, hierarchical...) to cluster the natural sequences, try different cluster numbers. Represent clusters graphically using the PCA of Task 2. Discuss your observations. Are functional and nonfunctional sequences separated into distinct clusters?
Unify natural and artificial sequences into one large dataset, and apply again the clustering algorithm. Are the two datasets separated by this procedure, or are clusters mixed in natural and artificial sequences?

Task 4: Predicting protein functionality
Use a classifier of your choice (e.g. logistic regression, but you could use random forest, neural networks etc.) to learn a function mapping sequences to functionality, *i.e.* to a binary output, using the training MSA of natural sequences. Test this classifier using the artificial test data, too. Determine the numbers of predictions which are true positives (TP – functional sequences predicted to be functional), false postives (FP – nonfunctional sequences predicted to be functional), true negatives (TN – nonfunctional sequences predicted to be nonfunctional) and false negatives (FN – functional sequences predicted to be nonfunctional). Compare them for training and test data. Note that for logistic regression, as for any soft classifier, these quantities are functions of a cutoff used to achieve a hard classification.

Task 5: Generating artificial sequences
Learn a generative model $P(a_1, ..., a_L)$ of your choice from the provided data and generate a set of artificial amino-acid sequences by sampling from P. Use your work in Tasks 1-4 to check if your generated sequences are good candidates for functional sequences or not.

Solving these tasks may strongly benefit from the notebooks provided in the exercises and in the original reference (Mehta et al., Physics Reports 2019). You may use the code of the exercises in Tasks 2-4, but the exploration of alternative approaches will be positively evaluated. In Task 5, it is not important that your model is a perfect generative model, but you should assess its qualities and potential limitations.

For the project, please prepare a Colab notebook with your implementations, analyses and results. You can work in grups of two students. Examinations are individual, you will have 12min to present your work (be concise and respect the time, please) followed by questions about the underlying methods and your findings. The focus will rather be on your understanding of the algorithms and interpretation of your results rather than on implementation details.

The notebook and the PDF of your presentation have to be sent the day before the examination to me by email.