UiO : Universitetet i Oslo

Machine Learning in Computational Biology: Data Representation

IN-BIOS5000/IN-BIOS9000

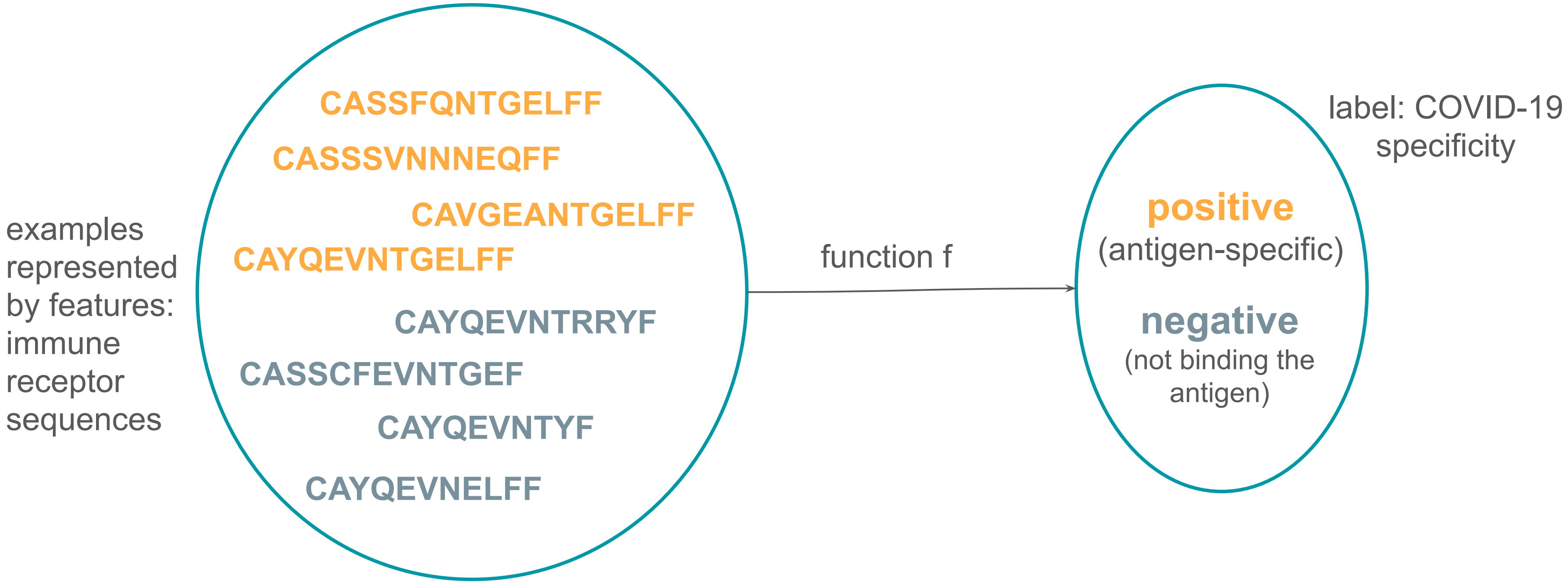
Milena Pavlović
Biomedical Informatics Research Group
Department of Informatics

milenpa@student.matnat.uio.no

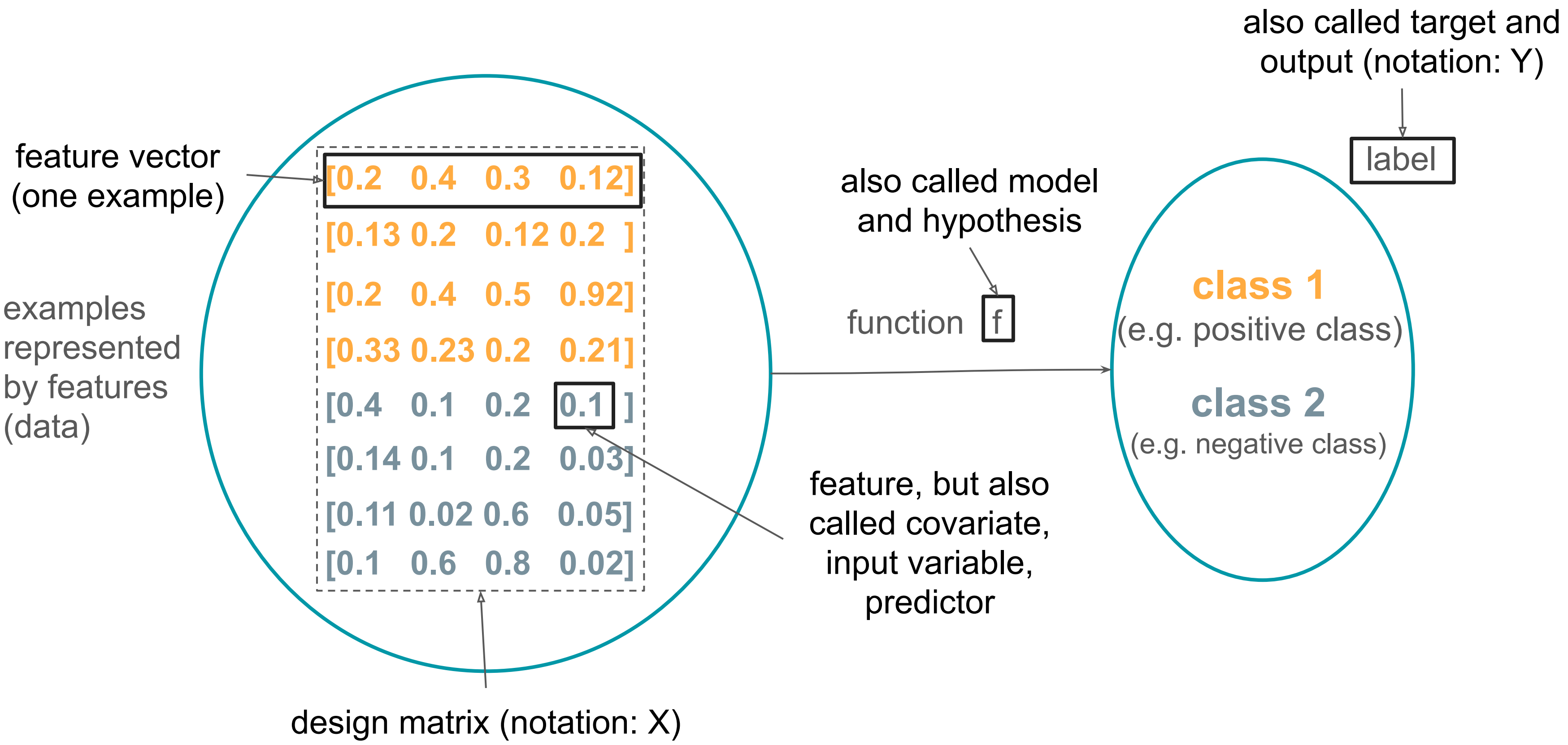
Machine learning in computational biology - outline

- Introduction to machine learning:
  - What is machine learning, types of problems, assumptions, workflow, generalization
- Machine learning models and algorithms:
  - Discriminative vs generative models, supervised models (logistic and linear regression, kNN, neural networks), unsupervised models (dimensionality reduction, clustering)
- **Data representation:**
  - **Considerations and examples, one-hot encoding, feature engineering, representation learning**
- Model comparison and uncertainty:
  - Model assessment, model selection, uncertainty, cross-validation
- Transparency and reproducibility

We are given a set of sequences… but algorithms only understand numbers!



We can represent sequences by their physicochemical properties, for instance



Some examples of data representation (encoding)

- One-hot encoding
- K-mer frequencies

Data representation heavily depends on data, so in different domains, there will be different representations:

When classifying images with classical approaches: number of edges, objects

When predicting the length of the trip: number of traffic lights, time of day

When predicting if an email is a spam or not: certain words, presence/absence of personal name

Some examples of data representation (encoding)

- One-hot encoding
- K-mer frequencies

We have to be careful how we choose features - we must not introduce information that is not there!

Data representation heavily depends on data, so in different domains, there will be different representations:

When classifying images with classical approaches: number of edges, objects

When predicting the length of the trip: number of traffic lights, time of day

When predicting if an email is a spam or not: certain words, presence/absence of personal name

One-hot encoding

- A common way to represent categorical data where only one value can be chosen: rows represent the possible values
- Also called *dummy variables* in statistics

nucleotide sequence: AATGC

	A	A	T	G	C
is it A	1	1	0	0	0
is it C	0	0	0	0	1
is it G	0	0	0	1	0
is it T	0	0	1	0	0

one-hot encoding

K-mer frequency

- Often used for sequence representation
- k-mers are (optionally overlapping) subsequences of length k

nucleotide sequence: AATGC →

<u>A A T G C</u>
A A T
  A T G
    T G C

present 3-mers: AAT, ATG, TGC

all possible 3-mers: AAA, AAC, AAG, AAT, ACA, …, TTT
(4^3=64 combinations)

AAA ...			AAT	...	ATG	...	TGC	... TTT
0	0	0	0.33	0 … 0	0.33	0 … 0	0.33	0 … 0

k-mer frequency encoding (k=3)

ML algorithm performance heavily depends on data representation

- Data representation refers to choosing and constructing features
- We don't always know it advance which features are the best for the problem: we have to know the domain:

CASSFQNTGELYF
CASSSVNNNEYFF
CAVGEANTGELFF
CAYQEVNTGELFF
CAYQEVNTRRYF
CASSCFEVNTGEF
CAYQEVNTYF
CAYQEVNELFF

raw data

represent sequence as 1 if it contains Y (tyrosine) and 0 if not:

$[1, 1, 0, 1, 1, 0, 1, 1]^T$

data representation: option 1

represent the sequence by k-mer frequency:

[0, .., 0.02, 0.03, .., 0]

data representation: option 2

Which one is better?

Feature engineering & feature selection

- Feature engineering: together with domain experts, ML researchers would discuss and derive features which they believe could be useful for the model

  Example: for biological sequences, there are a few popular alternatives like k-mer frequencies and physicochemical properties

- This way a lot of features could be constructed and the best ones would be selected as a part of fitting the model (feature selection)

Representation learning

- Most often in context of neural networks: the many layers of the network learn a hierarchical, alternative representation of the (raw) data that was provided as input

Convolutional Networks on Graphs for Learning Molecular Fingerprints

David Duvenaud†, Dougal Maclaurin†, Jorge Aguilera-Iparraguirre
Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, Ryan P. Adams
Harvard University

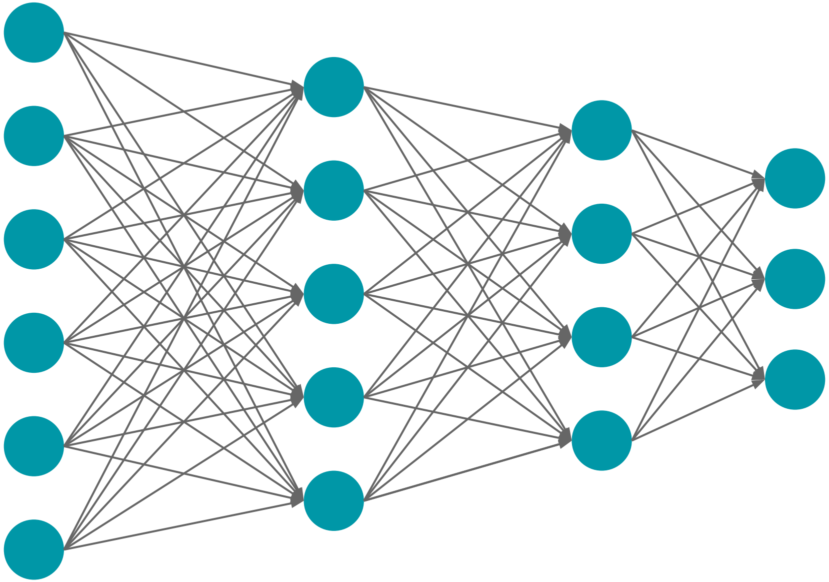
Article | Published: 21 October 2019

Unified rational protein engineering with sequence-based deep representation learning

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi & George M. Church

Nature Methods **16**, 1315–1322(2019) | Cite this article

Representation learning - hidden layers in neural networks can be seen as different representations

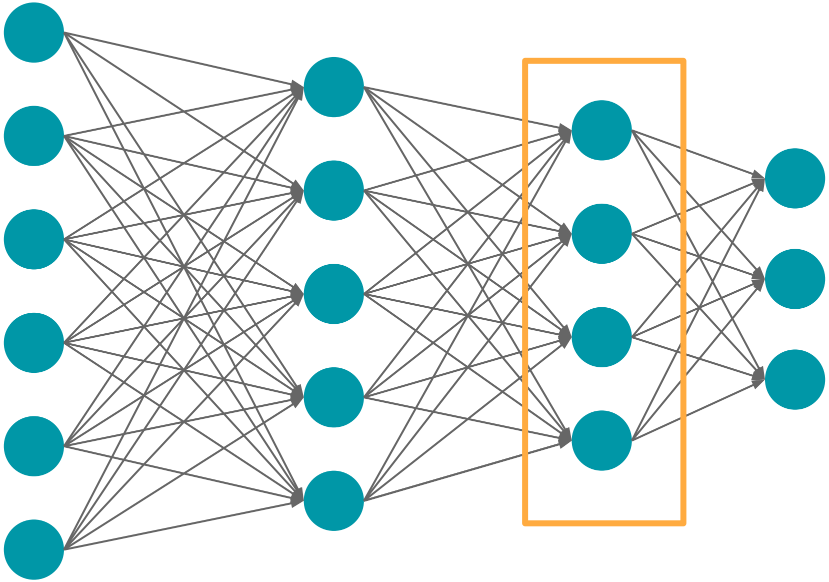


Deep neural network

"A good representation is the one that makes the learning task easier."

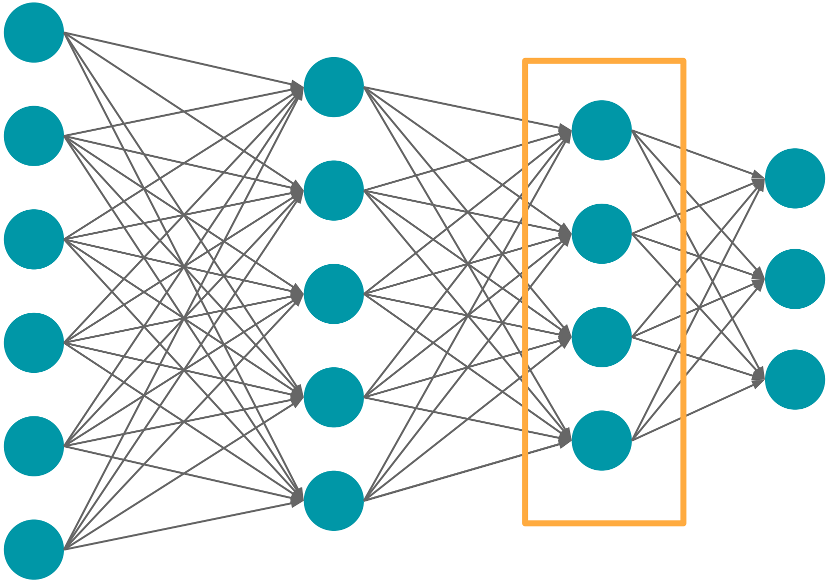
Goodfellow et al. 2016

Representation learning - hidden layers in neural networks can be seen as different representations

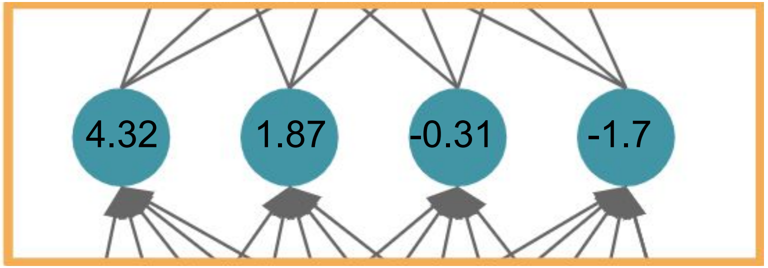


Deep neural network

Representation learning - hidden layers in neural networks can be seen as different representations

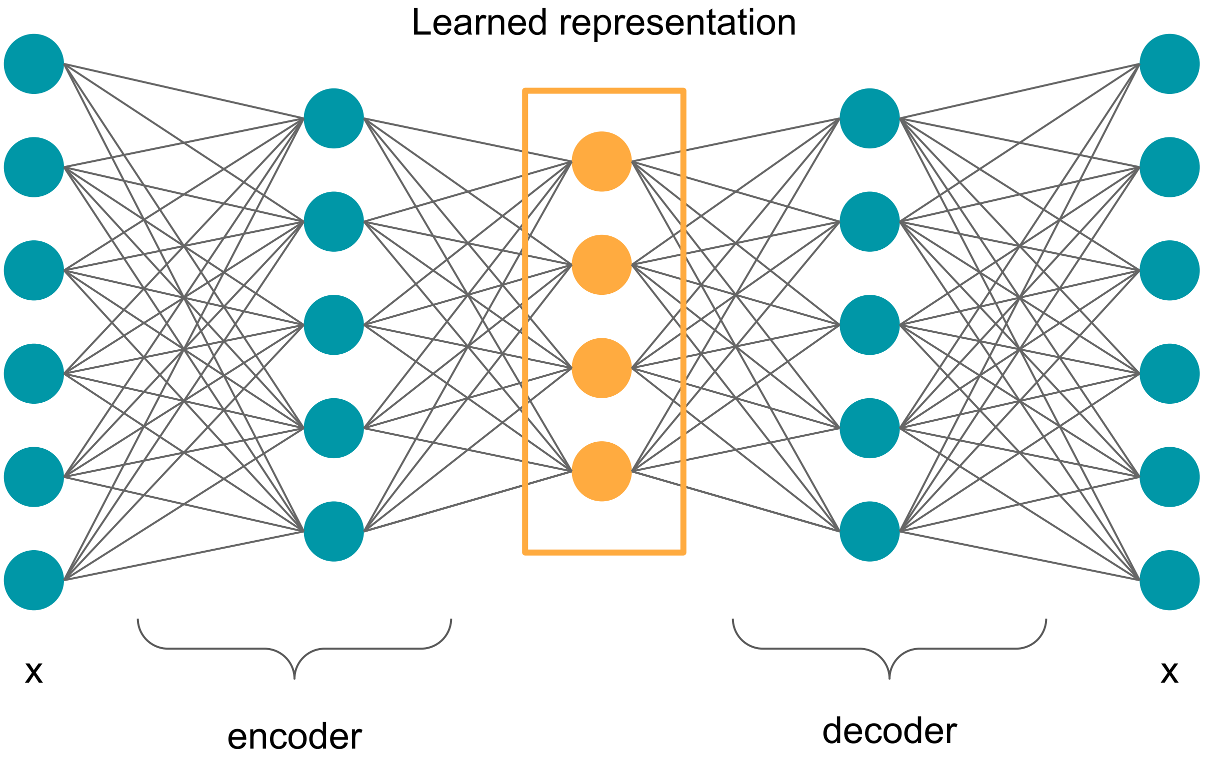


Deep neural network



New data representation

Representation learning with autoencoders



- Also used for dimensionality reduction and visualization
- Different types of autoencoders: denoising, sparse