11: Catchup II
Machine Learning and Real-world Data (MLRD)

Ann Copestake

Lent 2020
Last session: HMM in a biological application

- In the last session, we used an HMM as a way of approximating some aspects of protein structure.
- Today: catchup session 2.
- Bit more about cell membranes and proteins.
- Data and domain knowledge.
- Very brief sketch of protein structure determination:
  - including gamification and Monte Carlo methods: related ideas are used in many very different machine learning applications.
  - and a very little about AlphaFold.
What happens in catchup sessions?

- Lecture and demonstrated session scheduled as in normal session.
- Lecture material is non-examinable.
- Time for you to catch-up in demonstrated sessions or attempt some starred ticks.
- Demonstrators help as usual.
A biological application: the data

#MNQGKIWTVVNPAGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKAA

• top line records the amino acid sequence (one character per amino acid)
• bottom line shows the states:
  • i: inside the cell
  • M: within the cell membrane
  • o: outside the cell
Domain knowledge

Is this possible?

#MNQGIWTIVNPAIGIPALLGSVTIVAILVHAILSHTTWFPAYWQGKVKA

What about?

#MNQGIWTIVNPAIGIPALLGSVTIVAILVHAILSHTTWFPAYWQGKVKA

And this?

#MNQGIWTIVNPAIGIPALLGSVTIVAILVHAILSHTTWFPAYWQGKVKA
Domain knowledge

Is this possible?

#MNQGKIWTVVNPAGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA

What about?

#MNQGKIWTVVNPAGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKAA
Domain knowledge

Is this possible?

#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKA

What about?

#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKA

And this?

#MNQGKIWTVVNPAIGIPALLGSVTVIAILVHLAILSHTTWFPAYWQGGVKKA
Cell membranes and proteins

By LadyofHats Mariana Ruiz - Own work. https://commons.wikimedia.org/w/index.php?curid=6027169
Transmembrane protein: schematic diagram

1. a single transmembrane $\alpha$-helix (bitopic membrane protein)
2. a polytopic transmembrane $\alpha$-helical protein
3. a polytopic transmembrane $\beta$-sheet protein

Machine learning, abstractly

data preparation

task

data

evaluation

model/algorithm

feature extraction
Standard tasks

- Task
- Data preparation
- Model/algorithm
- Evaluation
- Feature extraction
- Provided
- Experiment
And actually . . .

real task

abstraction
task

data

PROVIDED
EXPERIMENT

data preparation

evaluation

feature extraction

model/algorithm
Tasks, data and domain knowledge

- Most ML researchers and textbooks ignore issues relating to data collection and task definition.
- Lots of examples of tasks that bear little resemblance to real applications.
- Real data is noisy and sometimes systematically biased.
  - Deep learning techniques are extremely good at exploiting data biases.
Tasks, data and domain knowledge

- Most ML researchers and textbooks ignore issues relating to data collection and task definition.
- Lots of examples of tasks that bear little resemblance to real applications.
- Real data is noisy and sometimes systematically biased.
  - Deep learning techniques are extremely good at exploiting data biases.
- Domain expertise is required to define the task and evaluation and to collect and check data.
Tasks, data and domain knowledge

- Most ML researchers and textbooks ignore issues relating to data collection and task definition.
- Lots of examples of tasks that bear little resemblance to real applications.
- Real data is noisy and sometimes systematically biased.
  - Deep learning techniques are extremely good at exploiting data biases.
- Domain expertise is required to define the task and evaluation and to collect and check data.
- ML expert plus domain expert: is ML approach modelling constraints? HMMs and membrane proteins?
Transmembrane protein example: (bovine) rhodopsin

- rhodopsin: one of the visual pigments
- accurate structure via x-ray crystallography: difficult and time-consuming, membrane location not determined
Protein structure

- Levels of structure:
  - Primary structure: sequence of amino acid residues.
  - Secondary structure: highly regular substructures, especially $\alpha$-helix, $\beta$-sheet.
  - Tertiary structure: full 3-D structure.

- In the cell: an amino acid sequence (as encoded by DNA) is produced and folds itself into a protein.

- Secondary and tertiary structure crucial for protein to operate correctly.

- Some diseases thought to be caused by problems in protein folding.
Alpha helix
Bovine rhodopsin

- found in the rods in the retina of the eye
- a bundle of seven helices crossing the membrane (membrane surfaces marked by horizontal lines)
- supports a molecule of retinal, which changes structure when exposed to light, also changing the protein structure, initiating the visual pathway
7-bladed propeller fold (found naturally)

http://beautifulproteins.blogspot.co.uk/
Peptide self-assembly mimic scaffold (an engineered protein)

http://beautifulproteins.blogspot.co.uk/
Protein folding

- Anfinsen’s hypothesis: the structure a protein forms in nature is the **global minimum** of the free energy and is determined by the amino acid sequence.

- Levinthal’s paradox: protein folding takes milliseconds — not enough time to explore the space and find the global minimum. Therefore kinetic function must be important.
Protein structure determination and prediction

- Primary structure may be determined directly or from DNA sequencing: relatively easy.
- Secondary and tertiary structure can be determined by x-ray crystallography and other direct methods, but difficult, expensive, time-consuming.
- Given amino acid sequence, can we predict the structure? i.e., determine how the protein will fold.
- Secondary structure prediction is relatively tractable: various prediction methods, including HMMs.
- Tertiary structure prediction is very difficult.
Protein tertiary structure prediction

- Modelling protein structure fully is hugely computationally expensive. Ideally, should model all the water molecules too . . .
- Several approaches, including:
  1. Molecular Dynamics (MD): modelling chemistry. folding@home: use home computers to run simulations.
  2. Foldit: get lots of humans to work on the problem (an example of **gamification**). [https://fold.it/portal/](https://fold.it/portal/)
  3. Use **Monte Carlo methods** (repeated random sampling) to explore possibilities.
  4. Use additional information either a) previously determined structures or b) **evolutionary coupling** (e.g., DeepMind’s AlphaFold)
2: Foldit: combined human-computer intelligence
Objective: find lowest energy state of protein.

Idea: start with secondary structure, try (pseudo)random move, see if result is lower energy and repeat.

Problem: local minima — locally good move may not be part of best solution.

So: also sometimes accept a move that increases energy.

Specific approach Metropolis-Hastings: a type of Markov Chain Monte Carlo method (e.g., Rosetta).
Monte Carlo methods in general

- Using random sampling to solve intractable numerical problems.
- Buffon’s needle problem used for estimating $\pi$ (‘experiment’ by Lazzarini 1901).
Monte Carlo methods

- Physicists developed modern Monte Carlo methods at Los Alamos: programmed into ENIAC by von Neumann.
- Bayesian statistical inference not until 1993 (Gordon et al): essential for many modern machine learning approaches.
- Gibbs sampling is a special case of Metropolis-Hastings.
- Much more about this in later courses.
- Practical introduction by Geyer in
  
  http://www.mcmchandbook.net/HandbookTableofContents.html
4: Using additional information in protein folding

1. use previously determined structures of similar proteins.
2. **evolutionary couplings**: databases of proteins in an evolutionary relationship, mutations tend to be correlated if amino acids are physically close in folded protein:
   - generate likely contacts (nowadays using deep learning), feed info into folding program;
   - Deep Mind’s AlphaFold: produce full probability distribution of distances, statistical potential function which is directly minimized by **gradient descent**.

https://deepmind.com/blog/alphafold/
(updated version in Bioinformatics)
Conclusions

Protein structure prediction is still unsolved. Possible approaches involve several techniques used elsewhere in machine learning:

- gamification: an example of human-computer collaboration
- Monte Carlo methods
- using additional information sources (domain knowledge)
Conclusions

Protein structure prediction is still unsolved. Possible approaches involve several techniques used elsewhere in machine learning:

- gamification: an example of human-computer collaboration
- Monte Carlo methods
- using additional information sources (domain knowledge)

General discussion in deep learning: constraints/priors vs tabula rasa approaches (also question as to what counts as tabula rasa . . .)