1 Probability

Some solutions to hw2.

A log linear model

Question 7:

Goal:

Give a model

$$p(y|x) : \text{ on "my understanding level } y \text{ after class given condition } x. \text{ "}$$

This is supposed to help me decide whether I should attend a lecture.

- $\mathcal{X}$ be my "condition" before lecture. Each $x \in \mathcal{X}$ consists enough data, where for $i = 1, \ldots, N$, I can define a collection of functions

$$\{q_i : \mathcal{X} \to \mathbb{R}\}_{i=1}^{N}$$

- $\mathcal{Y} := \{0, 1\}$, 0 means I basically understood nothing, and 1 means I got something out.

Working example

To give an explicit example: let $N = 5$. $\mathcal{X} := \mathbb{R}^5$, for $i = 1, \ldots, 5$ $q_i : \mathbb{R}^5 \to \mathbb{R}$ be simply projection on to the ith component, $x = (x_1, \ldots, x_5) \mapsto x_i$.
$x = (x_1, \ldots, x_5) \in \mathcal{X}$ encodes the following data:

- x_1: mood in scale 1-10.
- x_2: sleepiness in scale 1-10.
- x_3: zoom-ness (that is 1 if lecture was on zoom, 0 otherwise)
- x_4: my background knowledge for the coming lecture.
- x_5 day since first class.

The training set can be collected from attending each lecture.

Some feature choices

We have the following features: where

$$f(-,-) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$$

All features are binary. Let me give five most important features $f_1, \ldots, f_5$ continuing the above explicit example. First we can have features that are dependent on my mood.

$$f_{1a}(x,y) = \begin{cases} 1 & \text{if } q_1(x) \geq 5 \wedge y = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{1b}(x,y) = \begin{cases} 1 & \text{if } q_1(x) \leq 5 \wedge y = 0 \\ 0 & \text{otherwise} \end{cases}$$

But one is really the "negation" of the other. So we can simply just have one such feature, let $f_1 := f_{1a}$. Next we can feature on how sleepy I am coming to the lecture.

$$f_2(x,y) = \begin{cases} 1 & \text{if } q_2(x) \geq 5 \wedge y = 1 \\ 0 & \text{otherwise} \end{cases}$$

A feature on whether the lecture was on zoom:

$$f_3(x,y) = \begin{cases} 1 & \text{if } q_3(x) = 0 \wedge y = 0 \\ 0 & \text{otherwise} \end{cases}$$

A feature on my background knowledge:

$$f_4(x,y) = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let my feature vector be weight $w \in \mathbb{R}^k$. As of my experience, I almost get nothing out whenever its a zoom lecture. So I will give a high weight w_3. There should also be some weight to f_4 as I still get something out.
Here is another feature that I might consider. This feature is not *binary*. This i my cumulative knowledge increase as getting older:

$$f_5(x,y) = \begin{cases} q_5(x) & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let me remark on how one can easily enlarge the number of features to hundreds.

- my scale can be finer, going from 1-100.

- I can vary my constraints on features.

$$f(x,y) = \begin{cases} 1 & \text{if } \bigwedge_{i=1}^{N} \{q_i(x) = a_i\} \wedge y = 1 \\ 0 & \text{otherwise} \end{cases}$$

  where $\{a_i\}_{i=1}^{N} \in \mathbb{R}$.

- Increase the value of N. In example, I have $N = 5$.

Training data

Clearly, these are based on my experience. Perhaps for better model, incorporating other student's data might help.

1.1 Word similarity

Question 8. here I used the `words-50.txt`. The most similar words to :
`seattle`

```
    dallas atlanta wichita tacoma lauderdale florida spokane chino
       dulles
```

`dog`

```
    badger cat hound puppy dachshund sighthound poodle rat keeshond
```

`jpg`

```
    svg szczepanek buteo pix gif image galleria regnum fiav
```

`the`

```
    its of which entire within from a part second
```

`google`

```
    word not in vocabulary
```

Below I copied for `words-20.txt` and `words-100.txt` for `the`.

```
    in within its between entire over part uninterrupted marked
```

```
    its in which entire a itself this second from
```

For larger values of d , the result seem closer to what we think of similarity.
Now let us make some additions:

```
python3 findsim.py words-50.txt king --minus man --plus woman
```

```
    queen throne carloman son melisende disgrace sibylla daughter
       betrothed
```

```
python3 findsim.py words-50.txt hitler --minus germany --plus italy
```

```
    cesare petacci innitzer banality honoria accomplices benito
       conspirators aetius
```

Some tests

Example 1: abstract realtion *worked ok.*

```
python3 findsim.py words-50.txt love --minus heart --plus brain
```

```
    emotion feelings ashamed senses thoughts unrequited imagining
       intellect hypnotist
```

Example 2: job-relations *not really worked.*

```
    python3 findsim.py words-200.txt teacher --minus school --plus
       doctor
```

```
    davros contemplating daleks yueh firefly mcgann pangloss marple
       scifi
```

Example 3: concrete relations which do not work

```
python3 findsim.py words-50.txt meow --minus cat --plus bark
```

On `words-50.txt`

```
    tablecloth pomegranates bark quenched dried nisibis sinai fayyum
       groves
```

and on `words-100.txt`

```
    willow trees pine alder sap munching poplar hardwoods quercus
```

and on `words-200.txt`

```
    excelsa pine trees leaf olive poplar leaves bushes ebony
```

Example 4: people-relation *Worked.*

```
python3 findsim.py words-100.txt teacher --minus student --plus boss
```

```
    henchmen stooge karras realises kingpin riddler thief dugan
       vasquez
```

Discussion:

- I was expecting things where relation is more "concrete" to work better. This wasn't the case for `meow-cat+bark`.
- In some cases, as example 2,3 Increasing from 50- 200 did not help.

2 Interpretability

- Interpretability cheat sheet.
- A summary for explainable AI in industry

Perhaps it is *impossible* to achieve a completely interpretable model. We may thus ask for a model which is *trustable.* In the following way:

- One *trusts* a model when deployed.
- One *trusts* the prediction to take action depending on it.

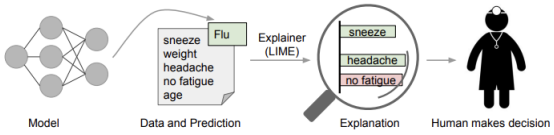

Figure 1: [?ribeiro2016should]

There are various forms of *explainations*. How should we organize it? Often it would be hard to formalize this, but we give a basic taxonomy.

- Local/global explanation. *Local* explanation, this is particular to an instance. *Global* explaination, reveals the full prediction logic, as is often in decision trees.
- Self-explaining/post-hoc. *Post-hoc* explanation usually involves a second step.

2.0.1 Why do we need interpretability

If you are convinced of its importance, then please *safely* ignore this section. In this section we explain the need of explainability from a *execution* point of view.

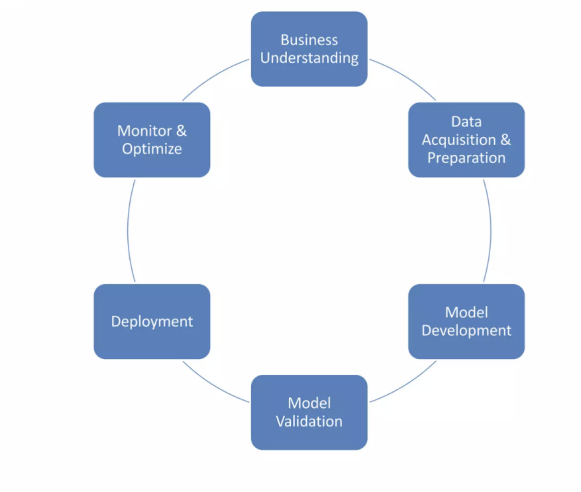

Figure 2: The AI Life Cycle

We refer to [?danilevsky2021explainability] for further details of role of explainability.

1. Monitor, and optimize. We have to put regular check to make sure that the system has continued effectiveness.

Example, [?Szegedy2013IntriguingPO]. The various stake holders.

Field	Examples
Buisness	Making loans.
Regulation	In US, there are several laws against discriminatio such as age, disability status, rac
Model developers	Which training exampels are most influential ? Debug and improve
Legal professionals	

2.0.2 Explainability and accuracy

Examples of XAI system.

2.1 Overview of techniques

The papers here do not

1. Post-hoc explainable AI: LIME, DeepLIFT.
    - Individual predictions: input features. The attribution problem, this can used to analyze robustness.

- Saliency maps. There has been an explosion of methods since 2018. The paper: Sanity checks for Saliency maps.

2. Interpretability.

2.2 Saliency map/Feature attribution

Definition 2.1. Feature importance: derives importance scores of different features used to output the final prediction.

Example, in the classical form of encoder-decoder for machine translation, [?Bahdanau2014NeuralMT]. In this paper, one has a *weight combination* of the latent features. Weights become the *importance scores* of these latent features. This an example of a local self-explaining approach, see also [?thorne2019generating].

2.2.1 LIME

Local interpretable model-agnostic explainations (LIME), is a model agnostic method, [?ribeiro2016should].

3 Mechanistic Interpretabiltiy

What is the area of *mechanistic interpretability*? Interpretability can mean rather different things in different context. One way to see what it is, is that

Discussion

Let us understand the following three questions:

- Who does the interpreting?
- Why is this needed?
- What type of reasoning?

Non-example: a *doctor* using ML for *diagonsis*, requiring reasoning for *specific* cases.

Non-example: Methods which increases predictive capability and improves calibrations of ML models.

Unclear Definition 3.1. *Mechanistic* refers to the understanding of the

- algorithms

Arguably, MI is a subfield of (*post-hoc*) *inner interpretability* or *white-box interpretability*. The first collection of works begins with the idea of *circuits*, **??**. To demonstrate the field, it is perhaps easier to look through examples.

3.1 Circuits in vision models

A nice analogy in [?olah2020zoom] is the central philosophy of neural networks are the following three hypothesis:

Hypothesis

1. There are fundamental units of neural networks, called *features*. [a] We will later discuss what *features* mean.
2.
3. Universality.

[a] One interpretation is being the *directions*.

3.1.1 Case study of curve dectectors

In analogy of how language model were designed for next word prediction turns out to have other capabilities, it is unclear why CNNs trained to classify image show have curve detectors falling out of gradient descent.

Question

- Is every neuron meaningful? See polyseismic neuron.

It is often hard to establish what exactly is meant by *neurons being curve detectors*. Here are a few criteria:

Hypothesis

A detector of feature X should satisfy the following properties:

- Causal.
- General. The detector should tolerate a wide range of features X and largely invariant to other attributes.
- Pure. There have no meaningful secondary function.
- Family. They come in varieties spanning all "similar" X.

A lot of techniques are *emprical*.

3.1.2 References

- L14. In lecture they discuss, how one could visualize the first layer effectively. Another method is last year.

The article importance of interpretable bases, conjectures of *prefered basis*.

Phenomena

Neural networks contain polysemantic neurons.

But why do they form? Why argument is from [?elhage2022toy], referring as *superposition*.

Hypothesis

Features are directions.

Hypothesis

Features are sparse.[a] This appear in the context of biology [?olshausen1997sparse].

[a] Let us phrase this mathematically. Given k vectors in $\{x_i\}_{i=1}^k \subset \mathbb{R}^n$. Let $\mathcal{S} := \{M : |M_{i\cdot}|_2 \leq 1\} \subset \mathbb{R}^{m\times n}$ be the "space of dictinoaries". In *sparse coding* one way to find an "over" complete set of vectors. then the goal would be to solve the optimization problem

$$\text{argmin}_{W\in\mathcal{S}} \sum_{i=1}^{k} |x_i - Wr_i|^2 + \lambda S(r_i)$$

Hypothesis

Priveledged basis.

- There is no reason to expect individual coordinates in stream to have any particular meaning.

3.1.3 Prerequisites

There are many noncanonical terminology in these sequences of works where we refer to Neel Nanda's glossary, if not explained.

3.2 Case study of superposition

Intuition: there is the *Johnson-Lindenstrauss lemma*

Theorem 3.2. Given $0 < \varepsilon < 1$.

To *test* the hypothesis, we

- create a NN which outputs a representation

$$Wx$$

where $W : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, with $n \gg m$. [1] The matrix W already tells us a lot. We can thus visualize two things

  - The extent of which feature i of x, dentoed x_i, is represented is $|W_i|$, where $W_i := Wx_i$. [2]One value associated in the paper is *dimensionality*

$$D_i := \frac{|W_i|^2}{\sum_{j=1}^{n} \hat{W}_i \cdot W_j^2}$$

    The denominator can be thought of as *how many features share the same representation.*

  - How *much* superposition occurs is measured by

$$\sum_{\{j\,:\,j \neq i\}} \hat{W}_i^t \cdot W_j, \quad \hat{W_i}^t := W_i^t / |W_i^t|$$

- Being a neural embedding implies that the weights of W are optimized under a choice of *loss function*, which is to measure how much it is lost from its original space. But to compare this to its original space we consider[3]

$$f(x, W) := \text{ReLU}(W^T W x + b)$$

or

$$f^L(x, W) := W^T W x$$

We define loss as

$$L_x := \sum_{i=1}^{n} I_i (x_i - f(x, W)_i)$$

Example: consider the case $n = 4, m = 2$. Suppose each four basis (column) vectors of $\mathbb{R}^4$, $e_1 = (1, 0, 0, 0), \ldots, e_4 = (0, 0, 0, 1)$ well be embedded. If

$$W := \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix}$$

Then $We_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$. On the other hand, when we "write" out our embedding,

$$W^t \begin{pmatrix} x \\ y \end{pmatrix}$$

this is done by taking the dot product of which each $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$

[1] I think the use of word *embedding* is misleading, especially in the context of geometry.

[2] For example, we can imagine a sentence where each word is embedded as the standard basis vectors $e_1, \ldots, e_n$ of $\mathbb{R}^n$.

[3] ReLUs introduce prevliged basis.

3.3 Short history of works

As this is a relatively new field, it is not hard to discuss the time line of works, a great survey is given [?rauker2023toward], of which we follow its taxonomy: we divide current interpretability techniques by which part of the DNN's graph they explain.

- Ablation methods.
- *Modularity*: this is an important concept in the field of *systems biology*. [4]

3.4 Philosophical thoughts

It is useful to step back two aspects :

- *epistemology* - how we come to accept a knowledge (and justify the acceptance).
- *empiricism*: it is worth reviewing the "orthodox narrative" as held by early modern British philosophers like Lock, Berkeley and Hume.

Embeddings encode *linguistic regularities*, [?levy-goldberg-2014-linguistic]. [5]

[4] Biological data is often represented as networks, e.g. genes are represented as nodes, and edges will mean some interaction.

[5] Examples of regularities include "man:woman" , "king:queen"; "france:french", "Mexico:Spanish".