Advanced Machine Learning

András Kornai

AML 2024/09/11

Is there anyone who knows how to

- code (in Python)?
- compute with vectors and matrices (linear algebra)?
- use probabilities and statistics?
- work in a team?

Course website https://nessie.ilab.sztaki.hu/~kornai/2024/AML

Topics that will be discussed (not necessarily in this order)

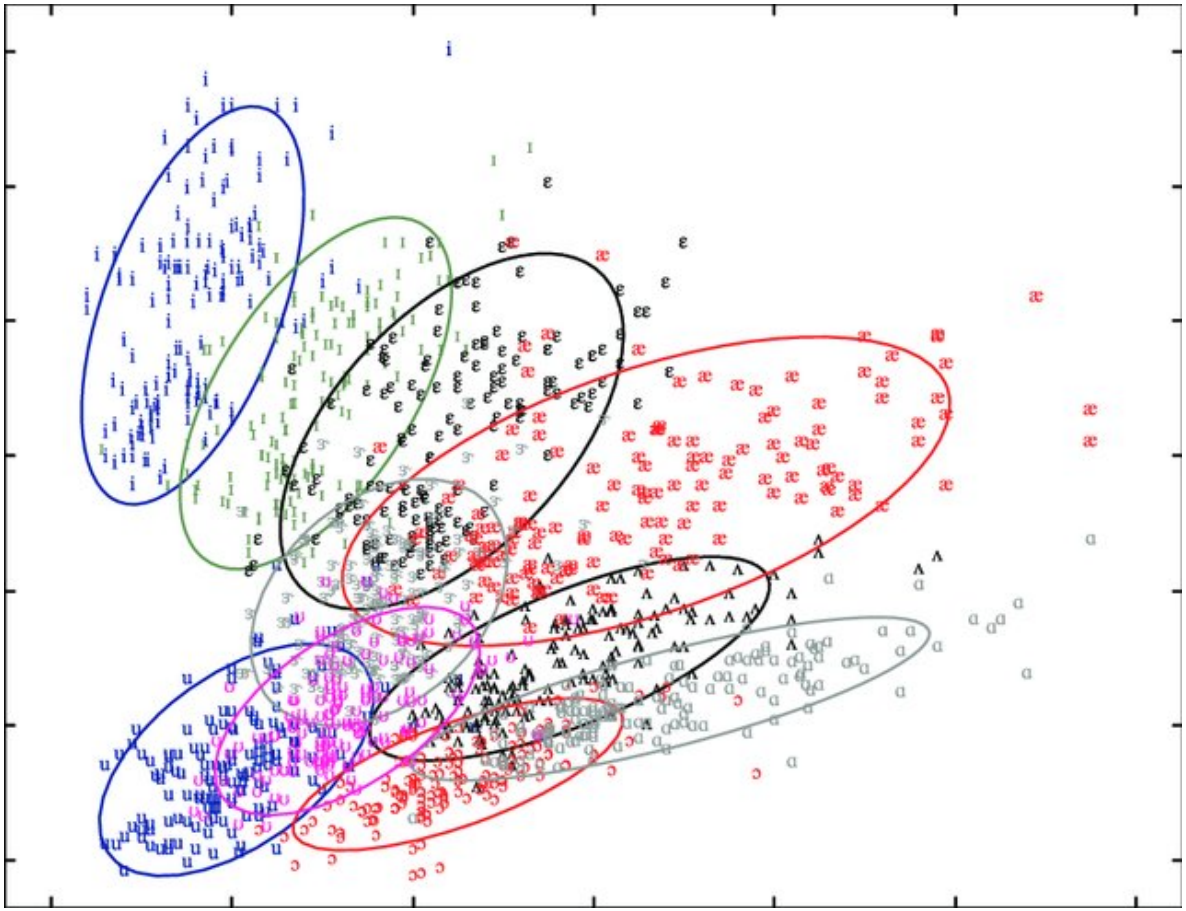
1. The main tasks: classification, regression, generation
2. The main areas: speech and character recognition, info extraction, info retrievalm ranking/recommendation, biometric ID, NLP tasks
3. Descriptive statistics, linear algebra, optimization, information theory, data reduction There will be a test next week
4. The main ML algorithms: linear classifiers, maximumm entropy, hidden Markov, nearest neighbor, max margin, genetic/evolutionary, boosting, decision trees, Bayes, NNs
5. NLP tasks: sequence labeling (POS, NER), chunking, parsing, anaphor resolution, disambiguation, language ID, role labeling, paraphrase, dictionary building, machine translation. We will read Ignat et al., 2024

Let's get started

```
@Article{   Peterson:1952,
author = {Gordon E. Peterson and Harold L. Barney},
title = {Control methods used in the study of vowels},
journal = {Journal of the Acoustical Society of America
volume = {24},
pages = {175--184},
year = {1952}}
```

Very influential paper, over 5k citations in Google Scholar, data recovered by Watrous, 1991

This is how the data looks



https://nessie.ilab.sztaki.hu/~kornai/2024/AML/PetersonBarney.tar

The data

1	1	1	IY	160.	240.	2280.	2850.
1	1	1	IY	186.	280.	2400.	2790.
1	1	2	IH	203.	390.	2030.	2640.
1	1	2	IH	192.	310.	1980.	2550.
1	1	3	EH	161.	490.	1870.	2420.
1	1	3	*EH	155.	570.	1700.	2600.
1	1	4	*AE	140.	560.	1820.	2660.
1	1	4	AE	180.	630.	1700.	2550.
1	1	5	AH	144.	590.	1250.	2620.
1	1	5	AH	148.	620.	1300.	2530.

The metadata

33 male (1); 28 female (2); 15 child (3) 10 vowels 2x (1520 recordings)

Data in 8 columns: 1 gender; 2 speaker; 3 phoneme; 4 phoneme in ascii; F0; F1; F2; F3

1	IY	[i]
2	IH	[I]
3	EH	[e]
4	AE	[ae]
5	AH	[^]
6	AA	[a]
7	AO	[o]
8	UH	[U]
9	UW	[u]
10	ER	[3]

Asterisk in ARPABET phoneme field means utterance failed of unanimous identification in listening test (26 listeners)

How can we deal with it?

- We have vectors and labels (truth)
- There will also be cases where it's not vectors, but sequences of vectors, but this is steady-state
- What is the first task?
- You need to cut this to train, validation, and test sets 80/10/10
- What is the second task?
- You need to build a simple baseline classifier
- What's the figure of merit?
- This is a classification task, we just look at accuracy
- If class sizes are very different, we can average per-class accuracies

The baseline

- We know the center of gravity for each phoneme
- This is a model as it stands
- At test time we look at which model is nearest
- This is already tricky: shall we include the starred data in train/test?
- Do we need all the data? Maybe it will be better without F0
- Is it worth normalizing the data?
- Is it worth adding all training data points (including far outliers)?
- Is it worth complicating the model?

GMM

- Gaussian Mixture Model
- Each class is considered a probability distribution
- Which is modeled by a mixture of n-dim gaussians
- One Gaussian is given by its mean (n parameters) and its covariance matrix ($n(n-1)/2$ parameters)
- Important special case: assume only variances but zero covariances (diagonal covariance matrix, only n parameters)
- You can work with a mixture of r Gaussians
- If there is enough data!

Homework

- Submit, by email, to `levai.math@gmail.com`
- With Subj: AML
- A jupyter notebook that contains
- Your baseline model of P&B, including the evaluation
- A model that you trained on the same data, ideally better than the baseline
- The body of your message should describe if you used any help (including ChatGPT CoPilot etc)
- Attach the file MYNEPT.NN.ipynb where MYNEPT is your neptun code, NN is the week when the homework is due (so this one is 02)
- No later than Saturday 6pm!

Ignat, Oana et al. (2024). *Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models*. arXiv: 2305.12544 [cs.CL]. URL: https://arxiv.org/abs/2305.12544.

Watrous, Raymond L. (1991). "Current status of Peterson-Barney vowel formant data". In: *Journal of the Acoustical Society of America* 89.5, pp. 2458–2459.