

ADVANCED MACHINE LEARNING

LECTURE 1

András Kornai

BMETE91MM20

2026/9/9

ORGANIZATION

- Hybrid meetings, zoom click here for zoom link
- Slack is getting expensive, do people want Teams or Discord?
- Course webpage <https://kornai.homeip.net/2026/AML> Also reachable as [kornai.com](#) → [Teaching](#) → [2026](#) → [AML](#)
- Attendance sheet
- Add github repo?

PREREQUISITES

- Introduction to data science 1 (BMETE95AM36) is highly recommended
- Material presented in that course and **not** discussed here includes least squares, linear regression, . . .
- Also required: ability to deal with real-world (or at least Kaggle) data, using python (scikit-learn), R, pandas, jupyter, etc.
- Notions refreshed: precision, recall, F-measure, and other figures of merit; gradient method; maximum likelihood; Estimate-Maximize
- Access to frontier LLM (Anthropic, OpenAI, Google)

GRADING

- Prelim test (today) gives 0% of the grade, but if you don't do well, this course is not for you!
- Individual project gives 50% of the grade
- Collective projects (2-4 people) are possible, discuss with instructor at 2nd class. May give 25% of the grade
- Final exam (oral) gives remainder of the grade (25-50%). There will be 2 random questions of the syllabus available at <https://kornai.homeip.net/2026/AML/Resources>

MATERIAL COVERED (NOT IN THIS ORDER)

- 1 Main tasks: classification, regression, generation
- 2 Main problem domains: speech- and character recognition (ASR, OCR), (biometric) identification, pattern classification, ranking/recommendation, info extraction, info retrieval, natural language processing (NLP)
- 3 Basics of descriptive statistics, linear algebra, optimization, information theory. Data reduction, principal component analysis, linear discriminant analysis, feature engineering
- 4 Survey of major machine learners: linear classifiers, maximum entropy, hidden Markov (HMM), nearest neighbor, max margin, genetic/evolutionary, boost, decision tree, Bayesian, neural net (NN)
- 5 Central NLP tasks: sequence tagging (POS, NER), chunking, parsing, anaphora resolution, disambiguation, language identification, role labeling, semantic similarity, paraphrase, dictionary building, machine translation

MATERIAL COVERED (CONT'D)

- 6 Learning finite automata, transducers, Eilenberg machines
- 7 Algorithmic information theory, Kolmogorov complexity, minimum description length
- 8 Unsupervised learning, structure learning

READINGS

Recommended: Duda, Hart, Stork: Pattern Classification (we won't cover the entire book); Ng: Machine Learning Yearning (very light reading)

Required: research papers. Always made available at the Resources page

WE BEGIN WITH A CLASSIC DATA SET

```
@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}}
```

HUGELY INFLUENTIAL PAPER

Google Scholar lists 5,235 citations (was 5,042 two years ago), data reconstructed in Watrous, 1991

0TH TEST

In the next 45 minutes

- Download PetersonBarney.tar and clawxiv.pdf from <https://kornai.homeip.net/2026/AML/Resources>
- Build a test set by taking all data from the first four male, the first four female, and the first two child speakers
- Build a train set from all remaining data
- Clean both sets by removing uncertain data (lines with * in the 4th column)
- Build a classifier: predict col 3 using only data from cols 5-8
- Write a clawxiv bundle that includes a human-readable paper (latex, bibtex, and pdf) describing what you did. Email me the bundle after 45 minutes
- Time permitting, use the test set to measure how well you did
- Not part of final grade, but best three get valuable prizes