

THE ORIGINAL TASK

- Download PetersonBarney.tar from class website
- 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
- Email me a tar file with a snapshot of your work after 45 minutes. This should have a README describing what you did.
- Time permitting, use the test set to measure how well you did
- Not part of final grade, but best three get valuable prizes

VALUABLE PRIZES!

- Last time prizes went to Bálint Turi-Kováts, Botond Kiss, and József Pintér
- This time prizes go to Melinda Juhász, Márton Veres, and 3rd place split between Botond Kiss and József Pintér
- **Problem children** EM89OJ, GFD2LD, GNIA1S, NCF0AL, O6U1ON, T8MF47, U9GPC8 no homework (sometimes no 0th test either). Send private msg on slack or zoom now
- Bunch of late submissions, please don't, we take 2 points off!
- Leaderboard

LEADERBOARD (ALL TRAIN NOSTAR TEST)

- Gradient Boosted Trees: 0.905 Stumphauer, Kiss
- Decision Trees 0.843 Havas
- Logistic Regression 0.870 Juhász A
- SVM 0.865 Ócsai
- Random Forest 0.892 Juhász M
- feedforward NN N/A
- k-Nearest Neighbor 0.892 Simon
- Prototype 0.854 Juhász M

TO GO FURTHER WITH THIS

- 1 Typical errors (e.g. testing on first 10 utterances, rather than all utterances of first 4+4+2 speakers) still need to be fixed by some!
- 2 Reporting of figures is hidden/nonexistent for some
- 3 Joint github repo? Your homework should be named AHWnXY.ipynb
- 4 You will also need to upload your data (if it's small)
- 5 There will be a midterm (week 6 or 7) and small homeworks
- 6 But the key issue now is your project

PROJECT DISCUSSION

- 1 What is your problem area? Speech, bio, ...
- 2 What data you want to use?
- 3 What is your goal?
- 4 Are you already set on a particular ML method?
- 5 How much work you think it will be?