Peer-Grading in a Course on Algorithms and Data Structures

Machine Learning Algorithms do not Improve over Simple Baselines

Learning @ Scale
26.04.2016

Mehdi S. M. Sajjadi
Morteza Alamgir
Ulrike von Luxburg

MPI IS Tübingen
Uni Hamburg
Uni Tübingen

Peer-Grading Setting

- How to aggregate grades?
- Are peer grades as accurate as TA grades?
- Challenges
  - Inexperience
  - Bias
  - Limited number of grades
  - Cheating

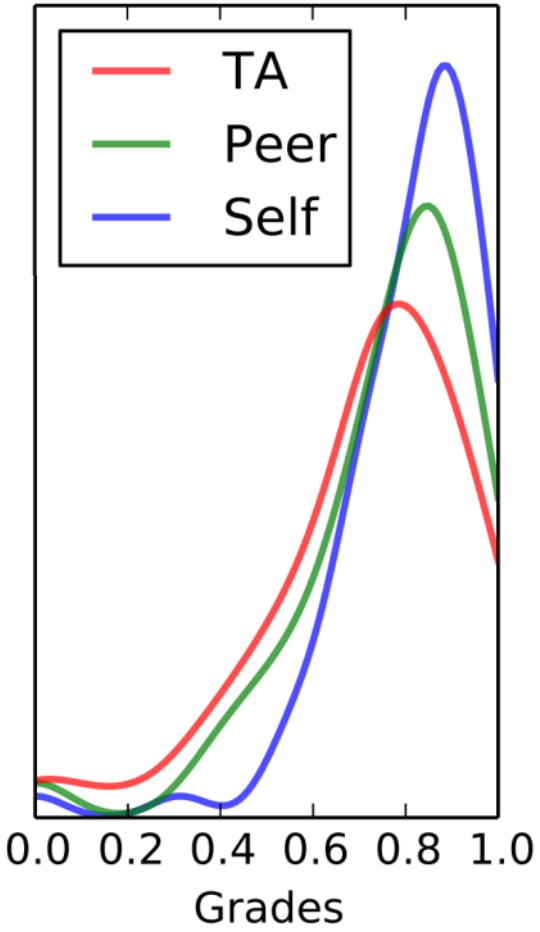
Our University Class

- Algorithms and Data Structures (easy-demanding tasks)
- 1 semester, 220 students, ~14.000 grades
- Groups of 3 students for solving exercises
- Detailed scoring rubric
- Grading done by everyone on their own
- All submissions graded in 3 different ways:
  - Self-assessment, peer-grading, TA grading

A glance at one exercise…

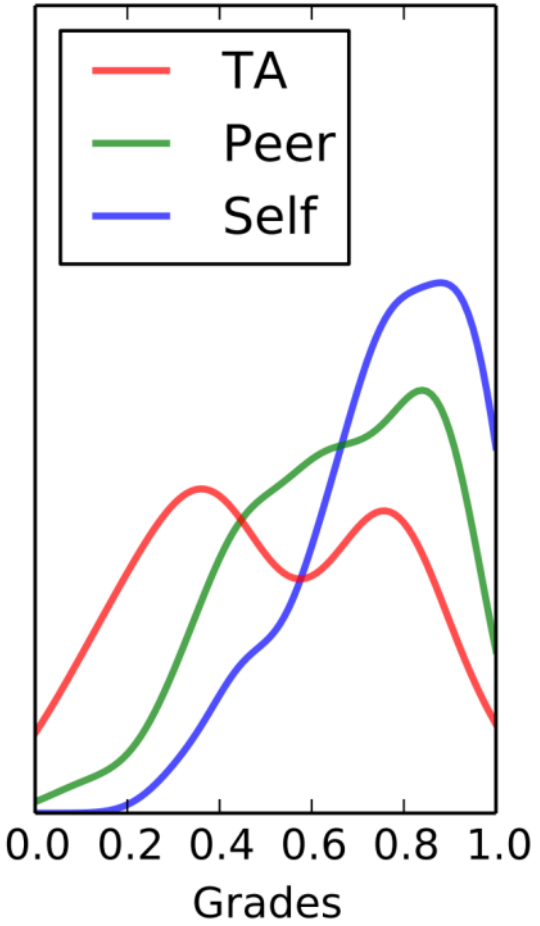
- Easy multiple choice with proofs, 0 or 1 point each

- Mean absolute deviation from TA grade
  - 0.08 Peer
  - 0.12 Self
  - 0.08 Peer + Self



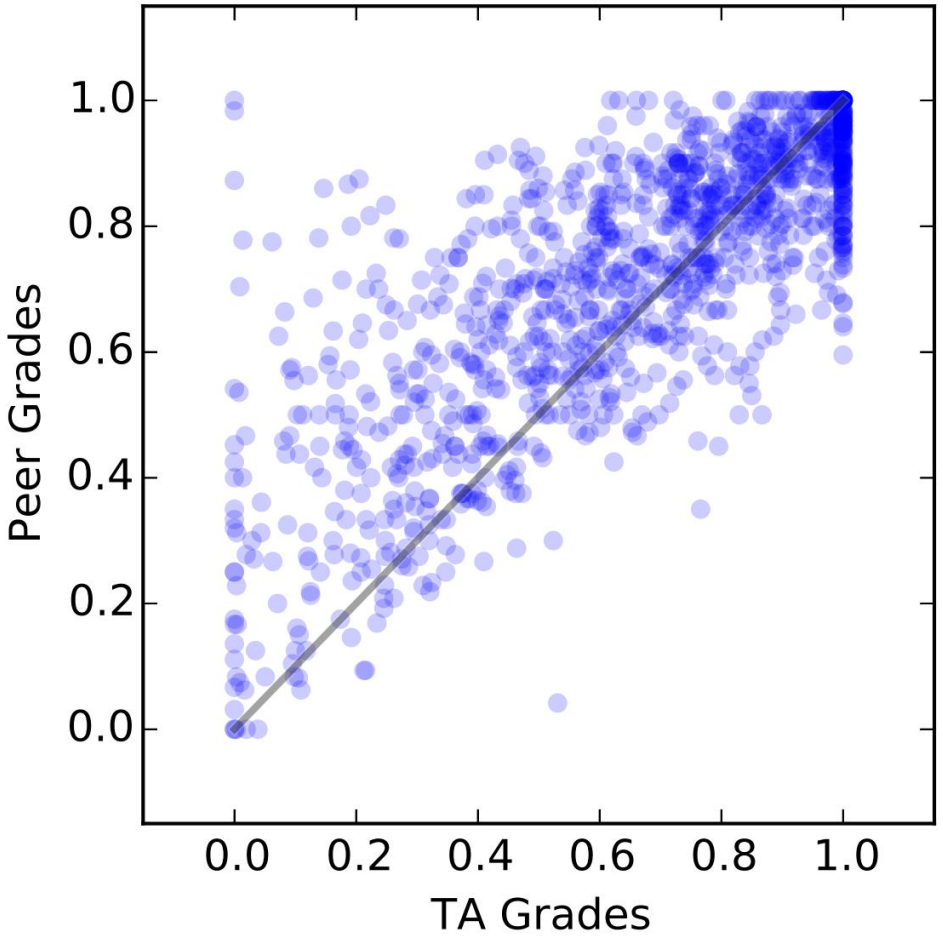
…and another exercise.

- Design algorithm, prove runtime
- Algorithms with bad runtime
- Students did not realize mistake
- Mean absolute deviation from TA grade
  - 0.18 Peer
  - 0.28 Self
  - 0.21 Peer + Self



Overall Grade Comparison

- Overall bias
  - 0.06 Peer
  - 0.12 Self
- Large variance
- **Good base for improvements?**

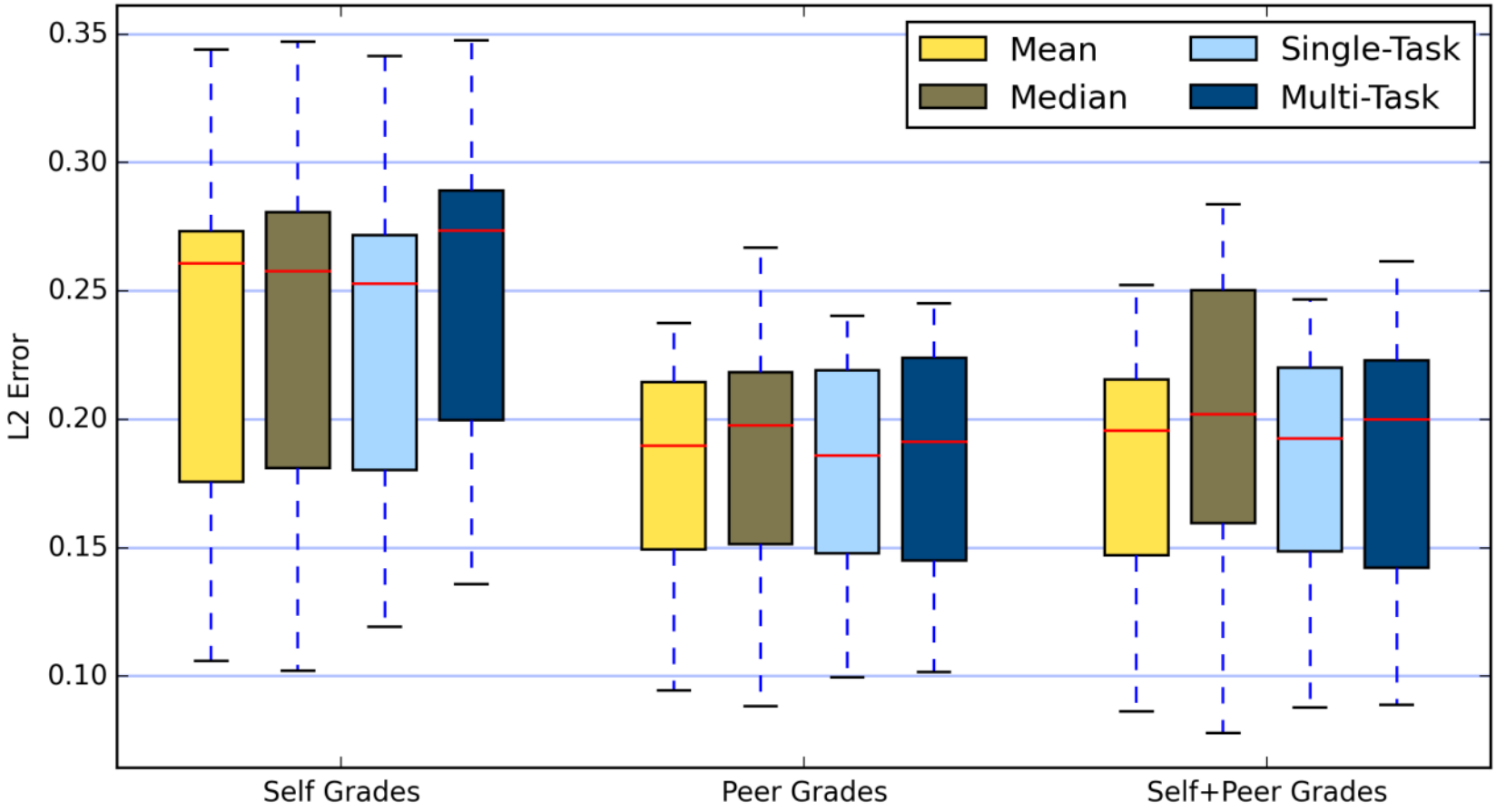


Probabilistic Model

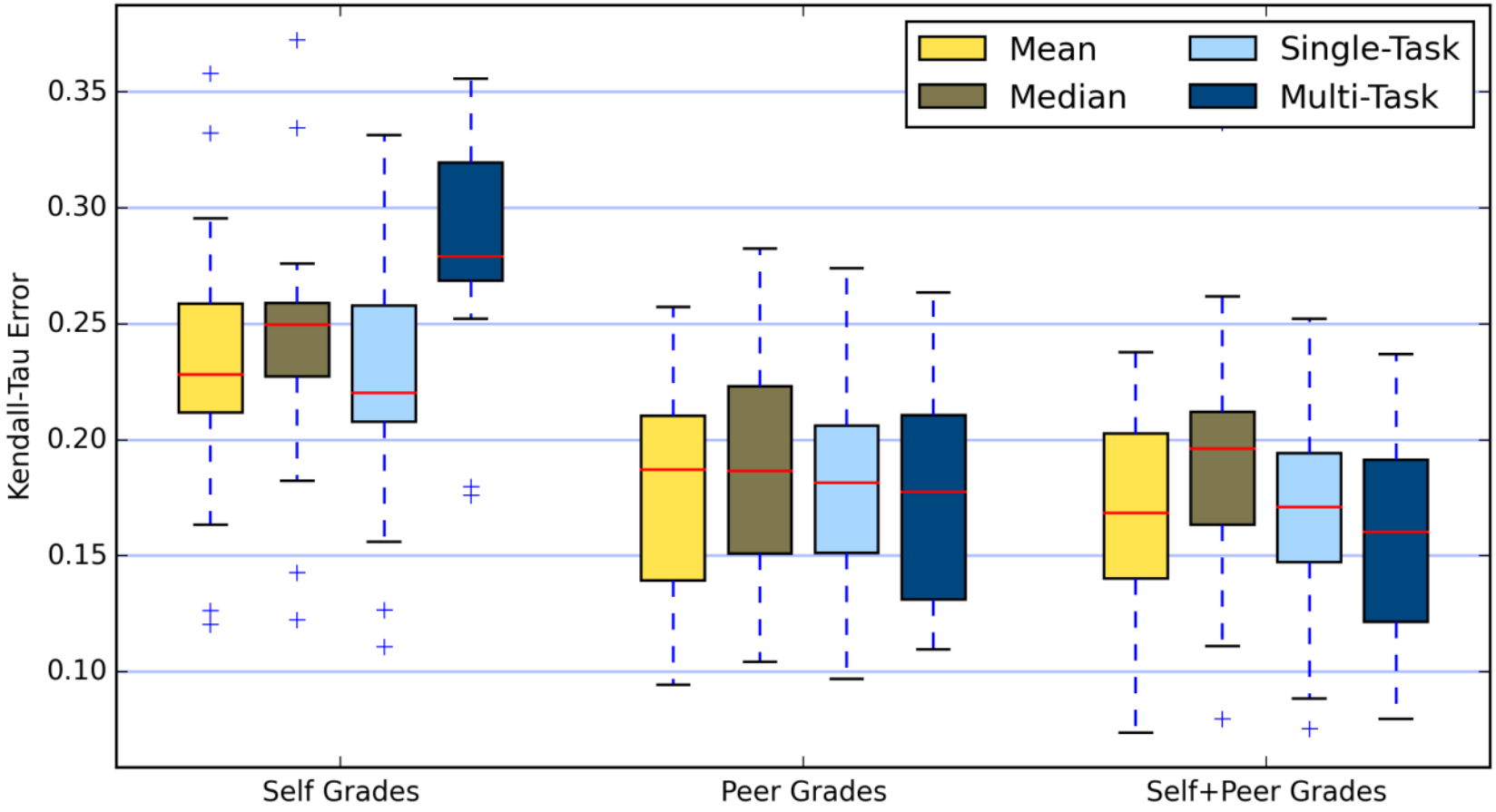
- $z_a^g \sim N(s_a + b_g, 1/r_g)$ Reported Scores
- $s_a \sim N(\mu, \sigma^2)$ True Scores
- $b_g \sim N(0, \eta^2)$ Bias
- $r_g \sim \Gamma(\alpha, \beta)$ Reliability
- EM algorithm for parameter estimation
- **Works well on artificial data**
  - Other algorithms in the literature yield similar results

C. Piech et al. Tuned Models of Peer Assessment in MOOCs. EDM, 2013.

Results on our dataset



What about the ranking?



Why no improvements?

Possible Reasons I

- **Amount of data?** (6 peer grades per assignment)
  - Same results with 15 peer grades
  - Bias and reliability estimation over all assignments
- **Wrong priors?**
  - They barely change the results
- **Unmotivated students** / useless reviews?
  - Very few, and models should benefit from them anyway
- **Different types of assignments?**
  - Grouping them by grading difficults did not change results

Possible Reasons II

- (Different) **TAs as baseline?**
  - The 6 TAs graded similarly
  - Some noisy in ground truth does not hurt models
- **Bias vs. Reliability!**
  - Most errros due to low reliabilities, not bias
  - This kind of error is hard to correct (artificial models)
- **Other sources of errors!**
  - Errors often a result of lacking knowledge
  - Hard to correct for this

Conclusion

- Models fail to improve over mean estimator
- Main reasons
  - Sources of errors are different from the assumption
  - Not much bias
  - Reliability difficult to estimate and correct
- Are complicated models acceptable for students?
- Is peer grading a viable option for university courses?
- The dataset is publicly available on our websites