Large Scale Learning - Challenge
(Learning with Millions of Examples and Dimensions)

Sören Sonnenburg, Vojtech Franc,
Elad Yom-Tov and Michele Sebag

Fraunhofer FIRST.IDA, Berlin

July 8, 2008
Lessons learned I

**Evaluation Criteria: Curves**

- Training Time vs. Test Error
- Dataset Size vs. Training Time
- Dataset Size vs. Test Error

Optimize points in Training Time vs. Test Error curve independently?

Pre-specify running time must be at least T?

How to treat methods that use different dataset sizes?
Lessons learned IIa

**Evaluation criteria: Scores**
(properties derived from the curves, introduced to declare a winner)

1. minimum aoPRC
2. auTime vs. PRC
3. auSize vs. PRC
4. Time aoPRC 5%
5. Size aoPRC 5%
6. Effort
Lessons learned IIb

**Evaluation criteria: Scores**

- Dataset-score is average rank based on these 6 values
- Overall-score is average rank over all datasets (ranking last on datasets where the method did not participate)

1. aoPRC
2. auTime vs. PRC
3. auSize vs. PRC
4. Time aoPRC 5%
5. Size aoPRC 5%
6. Effort

- of interest? real-valued?
- fair for different time?
- too much focus on small scale?
- won if next opponent is aoPRC 5% away
- won if next opponent is aoPRC 5% away
- is it cheating to have effort < 1?

Ranking not monotone, better use Elo ranking or ... ???
Ways to cheat

- Multiple scores are harmful. Simplified criterion?
- Timing/Objective numbers from participants cannot be trusted.
- Calibration of little use.
- Submit single point.

Other Lessons

- Re-evaluation should have fixed format.
- Writing a web interface from scratch takes > 1 week.
- *Public* LS-Real-world datasets are hard to find.
Future

Improvements for next LS-Challenge

- How to incorporate model selection time and or prior knowledge?
- Limit number of free models/parameters?
- Is there interest in having another LS-Challenge?
- All data in memory?
- One group of tasks of same type vs. diverse datasets?
- Fixed representation?
- Measure time at all? Wall-clock-time?
- Anyone interested in organizing?
- Datasets?
- Improvements to web-interface?
- Parallel?
- More?