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

Sören Sonnenburg and Vojtech Franc
Fraunhofer FIRST.IDA, Berlin

joint work with
Vojtech Franc, Elad Yom-Tov and Michele Sebag
Outline

1. Large Scale Learning
2. Motivation
3. Challenge
Large Scale Learning

Definition

Large Scale Problems

What makes a Problem Large Scale?

- Large number of data points
- Extremely high dimensionality
- High effort algorithms $O(N^3)$
- Large memory requirements

⇒ Anything that reaches current computers limits: computational, memory, transfer costs

Applications

- Bioinformatics (Splice Sites, Gene Boundaries, ...)
- IT-Security (Network traffic)
- Text-Classification (Spam vs. Non-Spam)
- Image Recognition
**Our Motivation**

**Current SVM solvers**
- Joachims 2005, $\text{SVM}^{\text{perf}}$ is *much* faster than $\text{SVM}^{\text{light}}$
- Own experiments: $\text{SVM}^{\text{light}}$ is *much* faster than $\text{SVM}^{\text{perf}}$
- Shalev-Shwartz et.al. 2007, Pegasos is much faster than $\text{SVM}^{\text{light,perf}}$
- Own experiments: Pegasos is much slower than $\text{SVM}^{\text{light,perf}}$
- Teo et.al. 2007, $\text{SVM}^{\text{perf}}$ is a special case of BMRM
- Own experiments: BMRM is much faster than $\text{SVM}^{\text{perf}}$
- new $\text{SVM}^{\text{perf2.1}}$ similar in speed to BMRM
- Bottou 2007, SGD done right outperforms competitors

**There is no reliable way to tell which method is faster!**
Evaluation was done using different criteria!

- Different Parameters $C, \varepsilon, \lambda, \ldots$
- Meaning of parameters different
- Evaluation based on test error, objective value, \ldots
- Programming Errors, Inefficient Code
- Other accidental mistakes.
Proposal for a Large Scale Learning Challenge

- **Main Goal**
  - Evaluation under exact same fair conditions to answer: Which learning method is most accurate given limited resources?
  - Evaluation based on training time, test error (or objective value, etc. specific to method)

- **Additional Goals**
  - Which method gives the overall best classification performance?
  - Which classifier is the most training time efficient while achieving a good test error?
  - Approximation vs. Exact Algorithms?
  - What should one tune? Data representation? Feature selection? Core algorithm?
Large Scale Learning

Motivation

Proposal

Competition

- **Two tracks:**
  - Wild Competition
  - Method Specific:
    - Linear SVM
    - RBF SVM

- **Setup:**
  - Method are trained on diverse labeled datasets (size \(10^2, 10^3, 10^4, 10^5, 10^6, 10^7, \cdots\)); unlabeled validation set and test set

- **Evaluation**
  - Record training time, validation and test output for 10 intermediate points
  - Timing “calibrated” using program measuring floating point, integer, memory speed; At the end re-evaluation on a single machine.
  - Live feedback for validation set
  - Feedback for test set after end of competition
  - Competitors are required to submit a detailed explanation of the used methods.
Datasets

Different properties: sparse/dense, high/low dimensional. Different splits: training, testing and validation parts.

- Real World Datasets:
  - 55M examples - human splice dataset (strings of length 201)
  - 500K examples - web-spam data (16M dims)
  - 3M examples - face detection 1K dimensions
  - 5M examples - OCR 1K dimensions

- Artificial Datasets:
  - Generated from known distribution ⇒ results can be compared with the optimal classifier.
  - Datasets with different properties will be generated:
    - Separable versus Non-separable data.
    - Data with low and high intrinsic dimensionality.
    - Data with different scale of features.
Time Line

- February - Start of Competition
- Beginning of June - End of Open Competition
- We perform re-evaluation on a single CPU Linux machine with 32G of memory
- 9 July 2008 - Evaluation in an ICML’2008 workshop

Proceedings in LNCS Springer for best performing methods
Setup and Evaluation Criteria

Setup Evaluation Criteria

- Time vs. Test Error or Objective Value
- Dataset Size vs. Time ($O(n^s)$)
- Dataset Size vs. Test Error or Objective Value

We will compute **Performance Figures** and **Scalar Measures**.

**Categories**

- Overall winner for given dataset based on test error.
- Overall winner based on average rank computed on scalar measures over datasets.
- ...

**Evaluation: Time vs. Test Error**

Scalar Measures: Test Error, Time for fixed error, Area under curve
Dataset Size vs. Time

Scalar Measure - Slope in Log-Log Plot $O(n^s)$
Scalar Measures: Dataset size for fixed error, Area under curve
Goals for SVMs

- What is the relation between objective value vs. test error?
- What is the relation between stopping conditions and test error?
- Which algorithm is good on what kind of data set ((un)balanced, high or low dimensional, range of C, etc.)
Adjusted Evaluation for SVMs

Setup and Evaluation Criteria for SVMs

- Linear SVM with sparse data representation
- RBF Kernel SVM with dense data representation
- Run SVM for given fixed values of C and kernel width
- Record objective value while training
- Additional stopping criterion: target objective value
- Figures: Time vs. C, Time vs. Objective, Time vs. Test Error and Objective
- Scalars: Total time for model selection (all Cs and kernel widths), Time to reach target objective
Take part in the Challenge!

Pascal Large Scale Learning Challenge

<table>
<thead>
<tr>
<th>Submitter</th>
<th>Title</th>
<th>Dataset</th>
<th>Track</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is</td>
<td>speedy</td>
<td>Webspam</td>
<td>Wild Competition</td>
<td>25.01.2008 22:34 CET</td>
</tr>
<tr>
<td>Is</td>
<td>bummer</td>
<td>Webspam</td>
<td>Wild Competition</td>
<td>25.01.2008 22:08 CET</td>
</tr>
<tr>
<td>Is</td>
<td>brutal</td>
<td>Webspam</td>
<td>Wild Competition</td>
<td>25.01.2008 19:16 CET</td>
</tr>
</tbody>
</table>

Training Time vs. Test Error

Dataset Size vs. Training Time

Dataset Size vs. Test Error
Start of Competition in February / End: June

- **Two Tracks:**
  - Wild Competition
  - SVM

- **10 Datasets**
  - 4 real world datasets (up to 55Mio examples)
  - 6 artificial datasets

- **Evaluation**
  - Figures: Time vs. Error
  - Dataset Size vs. Time
  - Dataset Size vs. Error

Take part in the Challenge!