1. Motivation

Goal: binary classification on large datasets. Linear classifiers are fast but large datasets are usually too complex for good linear separation.

Example: separate from.

2. Notations

\( x \), \( y \) denote vectors and matrices, \( x \) is a column vector.

3. Margins for Polytopes

Worst-case margin

The Convex Polytope Machine (CPM) formulation arises from maximizing the total margin:

\[
\sum_{i=1}^{n} \min \left( 1, 2y_i (\mathbf{w}^T \mathbf{x}_i + b) \right)
\]

This is equivalent to:

\[
\sum_{i=1}^{n} \min \left( 1, \mathbf{w}_i \mathbf{x}_i + b \right)
\]

The objective is smoothed by minimizing the sum of squared inverses, costs are split between positive and negative instances:

4. Convex Polytope Machine

The Convex Polytope Machine (CPM) formulation arises from maximizing the total margin:

\[
\sum_{i=1}^{n} \min \left( 1, 2y_i (\mathbf{w}^T \mathbf{x}_i + b) \right)
\]

This is equivalent to:

\[
\sum_{i=1}^{n} \min \left( 1, \mathbf{w}_i \mathbf{x}_i + b \right)
\]

The objective is smoothed by minimizing the sum of squared inverses, costs are split between positive and negative instances:

5. SGD-Based Learning

Parameters

\[ T \] number of iterations
\[ k \] number of sub-classifiers
\[ \gamma \] regularization factor

Empirical observation: instances of the same class sometimes cluster well

\[ \sum_{i=1}^{n} \mathbf{w}_i \mathbf{x}_i + b \]

Sub-classifier assignment choice

\( \mathbf{w}_i \mathbf{x}_i + b \)

6. Assignment Heuristic

Natural choice for \( x \) is to take the sub-classifier with highest score. This can lead to intractable suboptimal solutions where all sub-classifiers are assigned positive instances, while the remaining are assigned negative instances.

Basic heuristic idea:

- Count number of assigned instances per sub-classifier
- Count number of assigned instances per sub-classifier
- Choose minimum sub-classifier otherwise, average such that counts are uniform.

We use entropy as a proxy for "sufficiency spread." The final assignment task parameter is specifying the triggering uniformity threshold.

7. Evaluation

8. Influence of \( k \)

The influence of \( k \) on the generalization performance is an open problem. Empirical evidence suggests an optimal \( k \) depends on the dataset.

9. Conclusion

CPM provides a rich, large-margin, non-linear decision model while still enjoying the computational efficiency of a simple linear separator.

On a single core machine, CPM can learn CPM data from state-of-the-art parallelized SVM implementations while producing a comparable or higher accuracy model.

High-performance implementation freely available at https://github.com/alexkantchelian/CPM

10. References and Acknowledgments

This research was supported by NSF CAREER Expeditions grant (CNS-0954046), ONR grant (N00014-11-1-0723), NSF grant (IIS-0916405), and academic research grants from Yahoo, Cloudera, and EMC. The authors would like to acknowledge the anonymous reviewers for their feedback.

Alex Kantchev
Michael C. Tschantz

Peter L. Bartlett
Ling Huang

Anthony D. Joseph
J. D. Tygar