Sequential Maximum Margin Classifiers for Partially Labeled Data

Elizabeth Hou†, Alfred O. Hero III†

† University of Michigan, EECS Department

Motivation

- Data is not collected as a batch, but sequentially over time
  - Speech or streaming text classification
  - Satellite only transmits data daily
  - Agency’s quarterly reports
  - Often not possible/desirable to wait until complete before analyzing
  - Additionally, some of the labels may be missing
  - At any time point $t$, observed data as $D(t) = \{X(t), y(t)\}$
  - $X(t)$ is a matrix of $n(t)$ samples and $p$ feature variables
  - $l(t) < n(t)$ of the samples have label $y_i \in \{1, -1\}$

Sequential Laplacian MED

Parameters:
- $\theta$ are weights for decision boundary
- $\gamma_i$ are margin parameters
- $b$ is a bias term for decision boundary
- $\lambda$ is a regularizer

Objective:

$$\min_{\theta, \gamma_i, \lambda} \text{KL} \left( P(\cdot|\{D(t)\}_{t=1}^T) \big|| P_0(\cdot|\{D(t)\}_{t=1}^T) \right)$$

subject to

$$E_{\theta, \gamma_i} \left[ y_i(X_i^T \theta + b) - \gamma_i \right] \left( \{D(t)\}_{t=1}^T \right) \geq 0 \forall i$$

$$E_{\theta, \lambda} \left[ \theta^T L(t) \left( X_i \theta - \lambda \right) \left( \{D(t)\}_{t=1}^T \right) \right] \leq 0$$

Decision Rule Corollary [4]

Assume $\beta_i$ are fixed parameters. Then the decision rule reduces to

$$\hat{y}_i = \text{sgn}\left( X_i \mu_i + b \right)$$

which is a function of the previous mean and optimal parameters

$$\hat{\alpha}_i = \arg \max_{\alpha_i} Z(\alpha_i, \beta_i) =$$

$$-\frac{1}{2} \alpha_i^T Y \mu_i^T J X_i \alpha_i + \frac{1}{2} \alpha_i^T Y \alpha_i + \frac{1}{2} \beta_i^T 1 - Y_i J X_i \mu_{t-1} - \frac{1}{2} \sum_{i=1}^{l(t)} \log \left( 1 + \frac{\alpha_i}{C_i} \right)$$

s.t. $\alpha_i \geq 0$ and $\beta_i \geq 0$.

The above can be extended to non-linear decision boundaries with the kernel trick [4].

Experiments

Simulation

- Training: ~100 per time point ($n(t) \approx [97, 103]$)
- Test: 1000 test points, Accuracy = $(TP + TN)/1000$

Supervised: Data from 200 categorical distributions

- 100 are sparse (high prob. of 0s)
- 50 are relevant (lower prob. of 0s)
- 50 are used to distinguish between 2 classes

Semi-supervised: Data from interior of a 3-D sphere

- One class is mostly in center/interior
- Other class is on the shell
- Only 10% of the samples are labeled

Isolet Speech Database

- Follows the experimental framework used in [1]
- Training: (isolet1 - isolet4) broken into 24 time points of 5 speakers, only first speaker is labeled
- Test: 1,559 samples from isolet5

Acknowledgments

This work was funded partially by the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0000154 and partially by the University of Michigan ECE Departmental Fellowship.

References