

A Gaussian Latent Variable Model for Large Margin Classification of Labeled and Unlabeled Data

Do-kyum Kim, Matthew Der, and Lawrence K. Saul
UC San Diego



1. Overview

Motivation

- Semi-supervised learning: learn classifiers from n labeled and m unlabeled examples when $n \ll m$.
- How can we utilize unlabeled examples to improve classification?

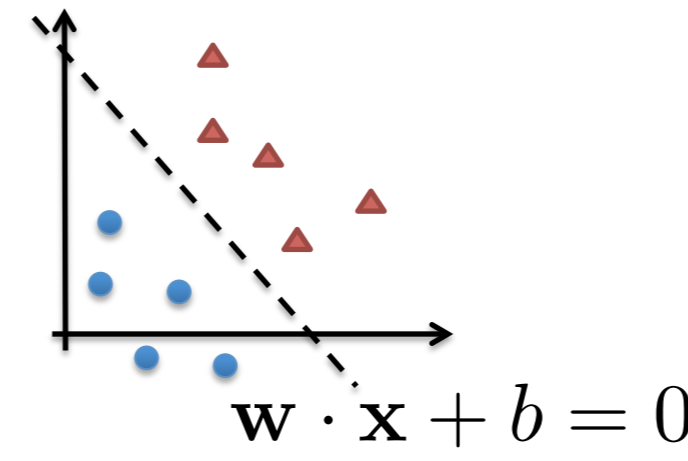
Our contribution

- Investigated a Gaussian latent variable model of linear large margin classifiers.
- Developed simple and highly scalable EM algorithm for learning.
- Utilized a Lyapunov central limit theorem to constrain unlabeled data to have a similar ratio of positive to negative examples as labeled data.

2. Background

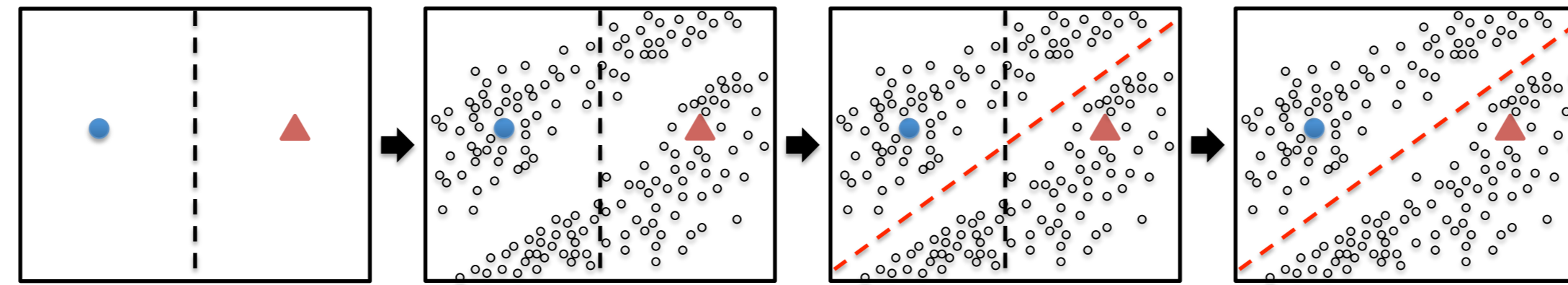
Linear classifier

- Data $\mathbf{x} \in \mathbb{R}^d$
- Linear score $z = \mathbf{w} \cdot \mathbf{x} + b$
- Label $y = \text{sign}(z)$



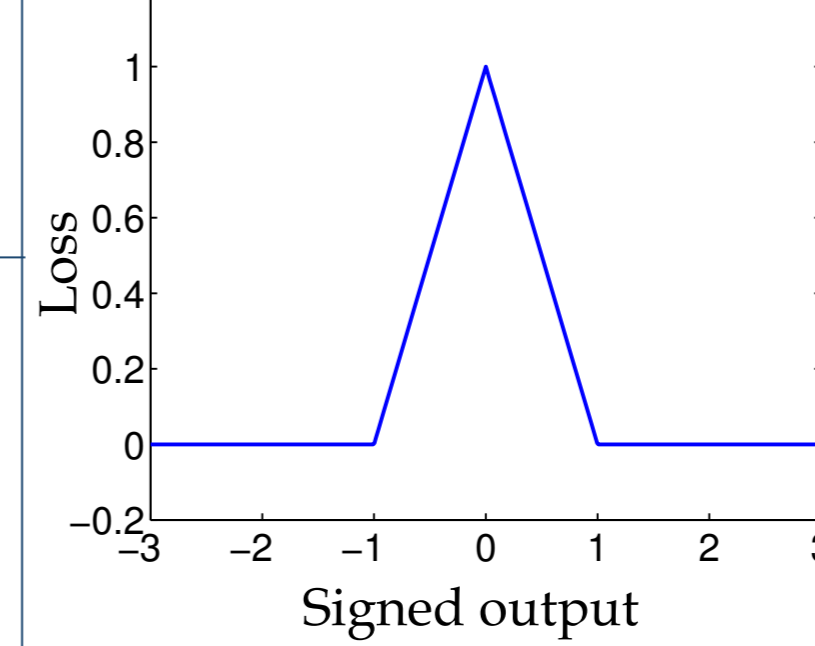
Cluster assumption

- Points in the same cluster are likely to share the same label.
- Find decision boundaries that cross through low density regions of unlabeled examples.



Semi-supervised SVMs (S³VMS)

- Extend SVMs to handle partially labeled data based on the cluster assumption.



Cost function for unlabeled data in S³VMS [Chapelle and Zien, AISTATS 2005].
This non-convex loss is much more difficult to optimize than those for SVMs.

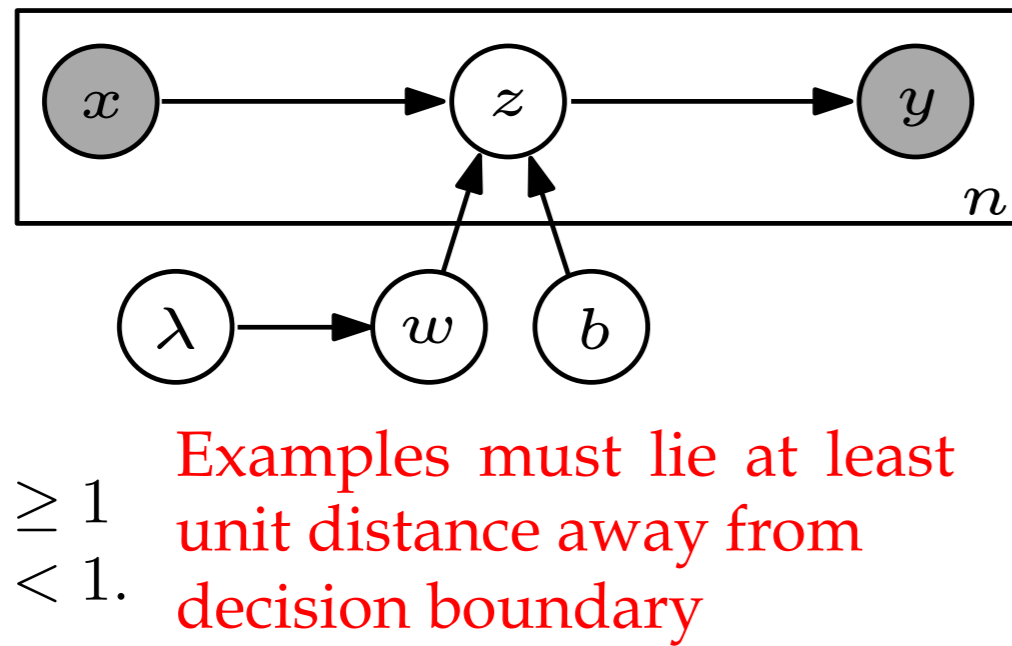
Why latent variable models?

- Key intuition: model missing labels as latent variables; use EM algorithm for learning.
- Successes in clustering (mixture of Gaussians), sequencing (hidden Markov models), and topic modeling (LDA).

3. Model for labeled data

Graphical model

- Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Latent variable $z_i \sim \mathcal{N}(\mathbf{w} \cdot \mathbf{x}_i + b, 1)$
- Label $y_i = \begin{cases} \text{sign}(z_i) & \text{if } |z_i| \geq 1 \\ 0 & \text{if } |z_i| < 1 \end{cases}$
- Regularization $\mathbf{w} \sim \mathcal{N}(0, \lambda^{-1} \mathbf{I}_d)$

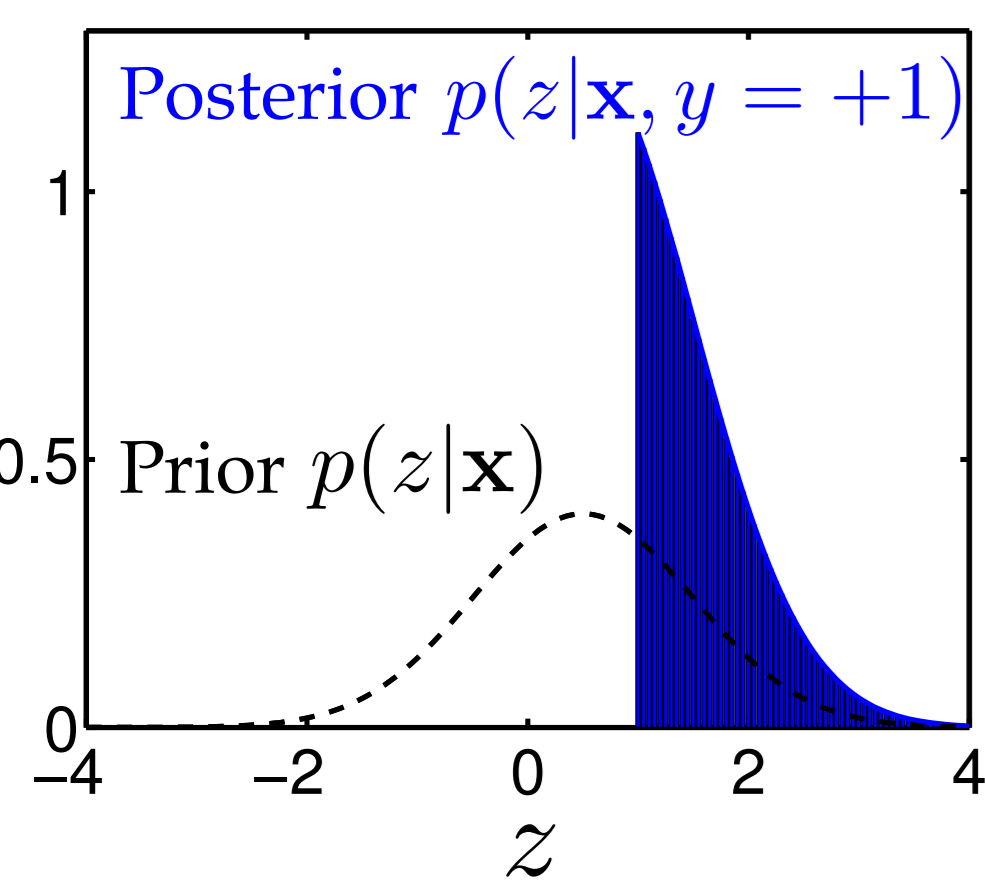


Examples must lie at least unit distance away from decision boundary

Learning

- Learn by maximizing likelihood: $\mathcal{L}_{\text{labeled}}(\mathbf{w}, b) = \sum_{i=1}^n \log P(y_i | \mathbf{x}_i, \mathbf{w}, b) - \frac{\lambda}{2} \|\mathbf{w}\|^2$
- Use Expectation-Maximization (EM) algorithm
 - E-step: compute posterior mean $\bar{z}_i = E[z_i | \mathbf{x}_i, y_i, \mathbf{w}, b]$
 - M-step: solve least squares $\min_{\mathbf{w}, b} \sum_{i=1}^n (\bar{z}_i - \mathbf{w} \cdot \mathbf{x}_i - b)^2 + \lambda \|\mathbf{w}\|^2$
- Converge to global maximum of $\mathcal{L}_{\text{labeled}}$

Prior vs. posterior

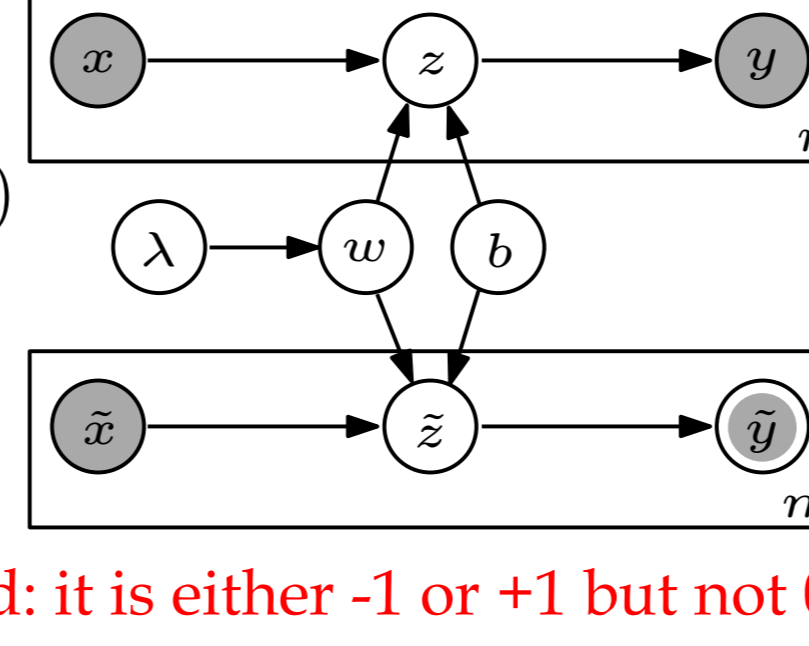


Prior and posterior for a positive example when prior mean $E[z|x] = 0.5$

4. Model for partially labeled data

Incorporating unlabeled data

- Unlabeled data $\{\tilde{\mathbf{x}}_j\}_{j=1}^m$
- Latent variable $\tilde{z}_j \sim \mathcal{N}(\mathbf{w} \cdot \tilde{\mathbf{x}}_j + b, 1)$
- Label $\tilde{y}_j = \begin{cases} \text{sign}(\tilde{z}_j) & \text{if } |\tilde{z}_j| \geq 1 \\ 0 & \text{if } |\tilde{z}_j| < 1 \end{cases}$
- Constraint $\tilde{y}_j \neq 0$, i.e. $|\tilde{z}_j| \geq 1$

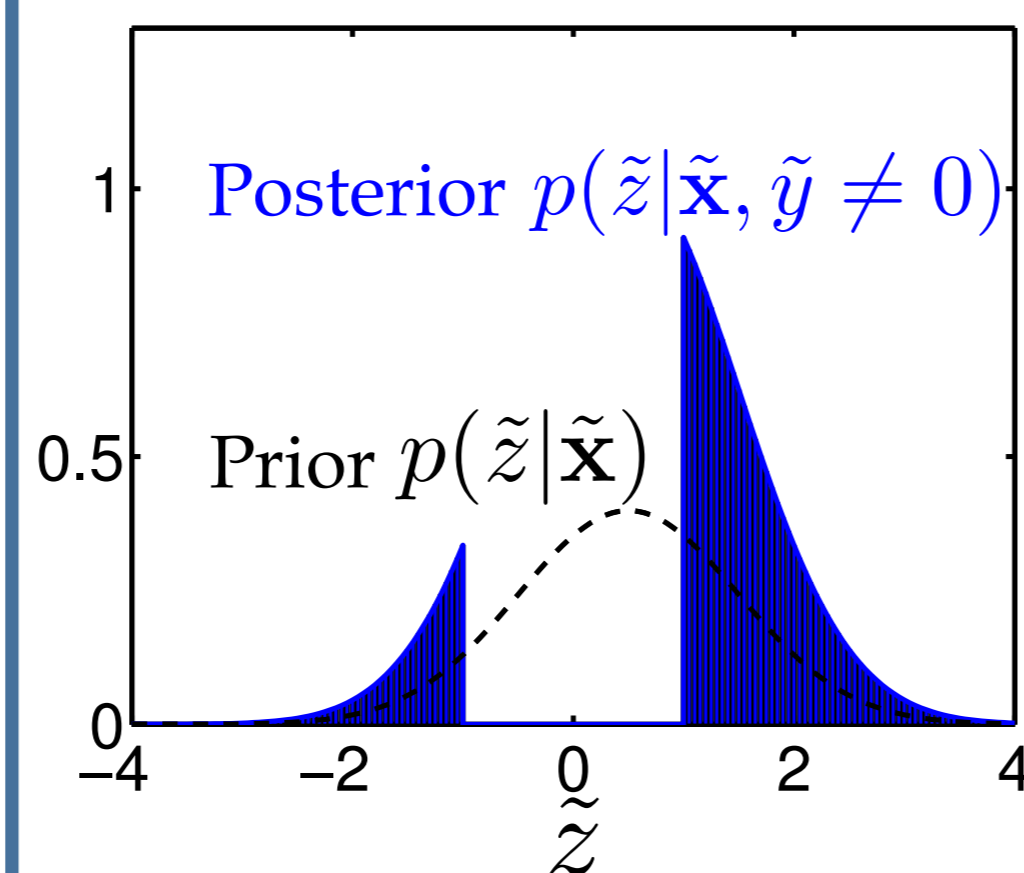


\tilde{y} is partially shaded: it is either -1 or +1 but not 0.

Learning with unlabeled data

- Learn by maximizing likelihood: $\mathcal{L}_{\text{ss}}(\mathbf{w}, b) = \mathcal{L}_{\text{labeled}}(\mathbf{w}, b) + \sum_{j=1}^m \log P(\tilde{y}_j \neq 0 | \tilde{\mathbf{x}}_j, \mathbf{w}, b)$
- Use EM algorithm
 - E-step: compute posterior mean
 - For labeled example $\bar{z}_i = E[z_i | \mathbf{x}_i, y_i, \mathbf{w}, b]$
 - For unlabeled example $\hat{z}_j = E[\tilde{z}_j | \tilde{\mathbf{x}}_j, \tilde{y}_j \neq 0, \mathbf{w}, b]$
 - M-step: solve least squares $\min_{\mathbf{w}, b} \sum_{i=1}^n (\bar{z}_i - \mathbf{w} \cdot \mathbf{x}_i - b)^2 + \sum_{j=1}^m (\hat{z}_j - \mathbf{w} \cdot \tilde{\mathbf{x}}_j - b)^2 + \lambda \|\mathbf{w}\|^2$
- Converge to local maximum of \mathcal{L}_{ss} (non-convex)

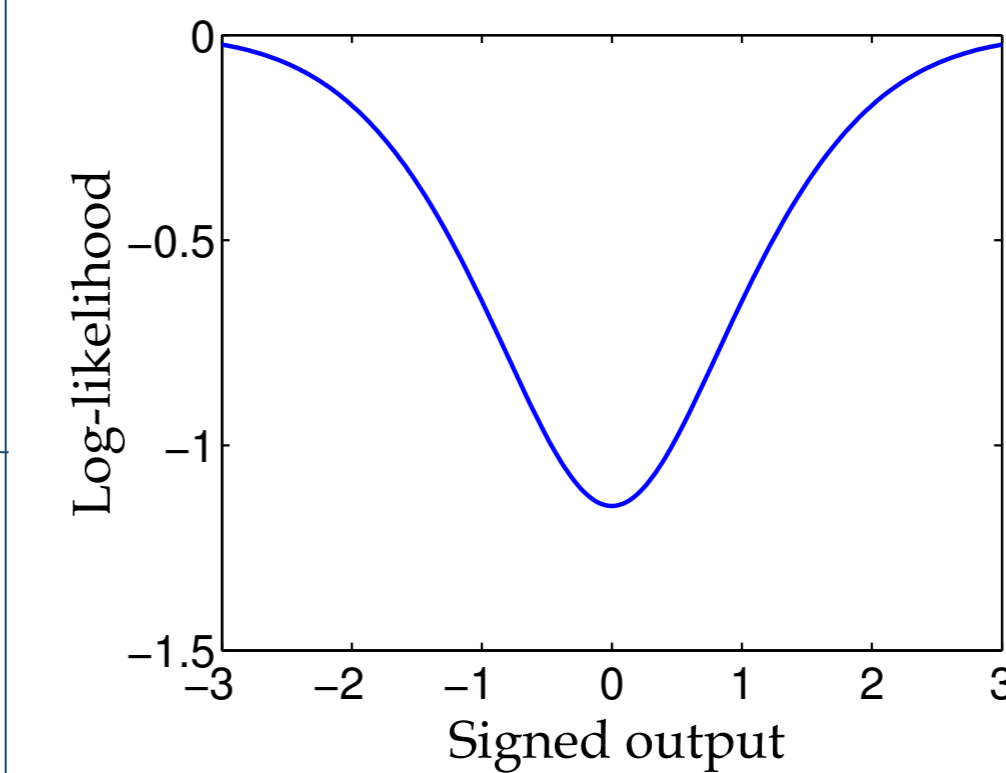
Prior vs. posterior



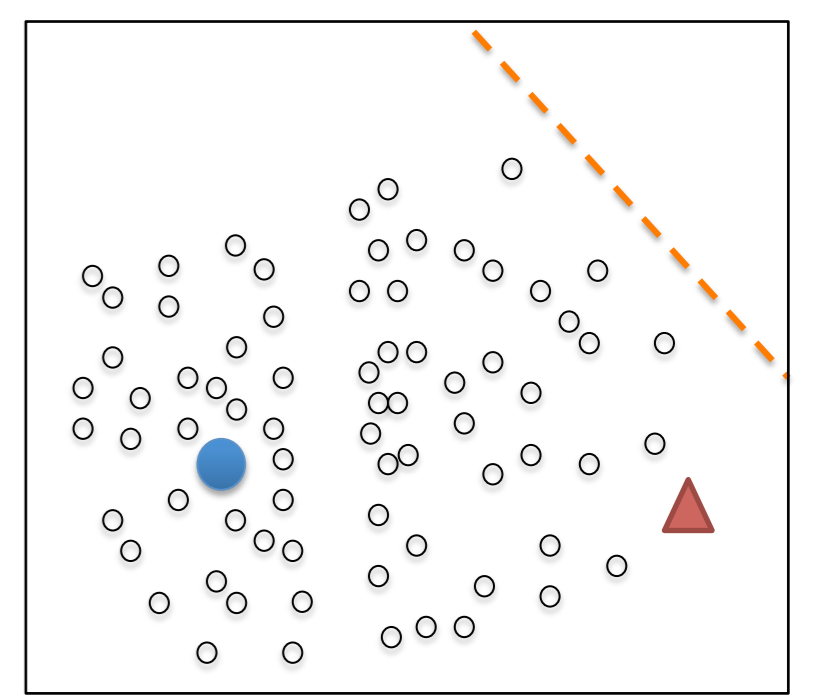
Prior and posterior for an unlabeled example when prior mean $E[\tilde{z} | \tilde{\mathbf{x}}] = 0.5$

Degenerate solution

- Large-margin constraints push decision boundary out of unlabeled examples; this often leads to degenerate solutions.



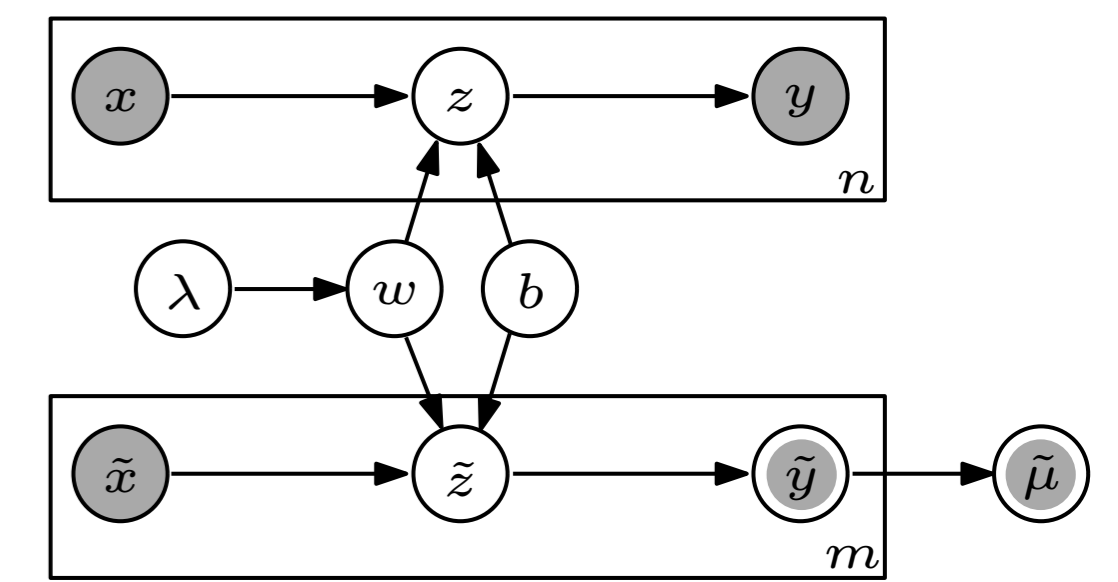
Log-likelihood for unlabeled examples



All examples are assigned to the same class.

Class-balancing constraints

- Class balance in labeled and unlabeled examples: $\mu = \frac{1}{n} \sum_{i=1}^n y_i$, $\tilde{\mu} = \frac{1}{m} \sum_{j=1}^m \tilde{y}_j$
- Constrain $|\tilde{\mu} - \mu| < \epsilon$
- Compared to equality constraints in S³VMS, our range constraints give the model more flexibility to fit the data.



$\tilde{\mu}$ is partially specified

Learning with class-balancing

- Learn by maximizing likelihood: $\mathcal{L}_{\text{ss}}^{\text{bal}}(\mathbf{w}, b) = \mathcal{L}_{\text{ss}}(\mathbf{w}, b) + \log P(|\tilde{\mu} - \mu| < \epsilon | \{\tilde{\mathbf{x}}_j, \tilde{y}_j \neq 0\}_{j=1}^m, \mathbf{w}, b)$
- $\tilde{\mu}$ is sum of independent, but non-identical random variables.
- Lyapunov central limit theorem provides excellent approximation.

5. Experiments

Setup

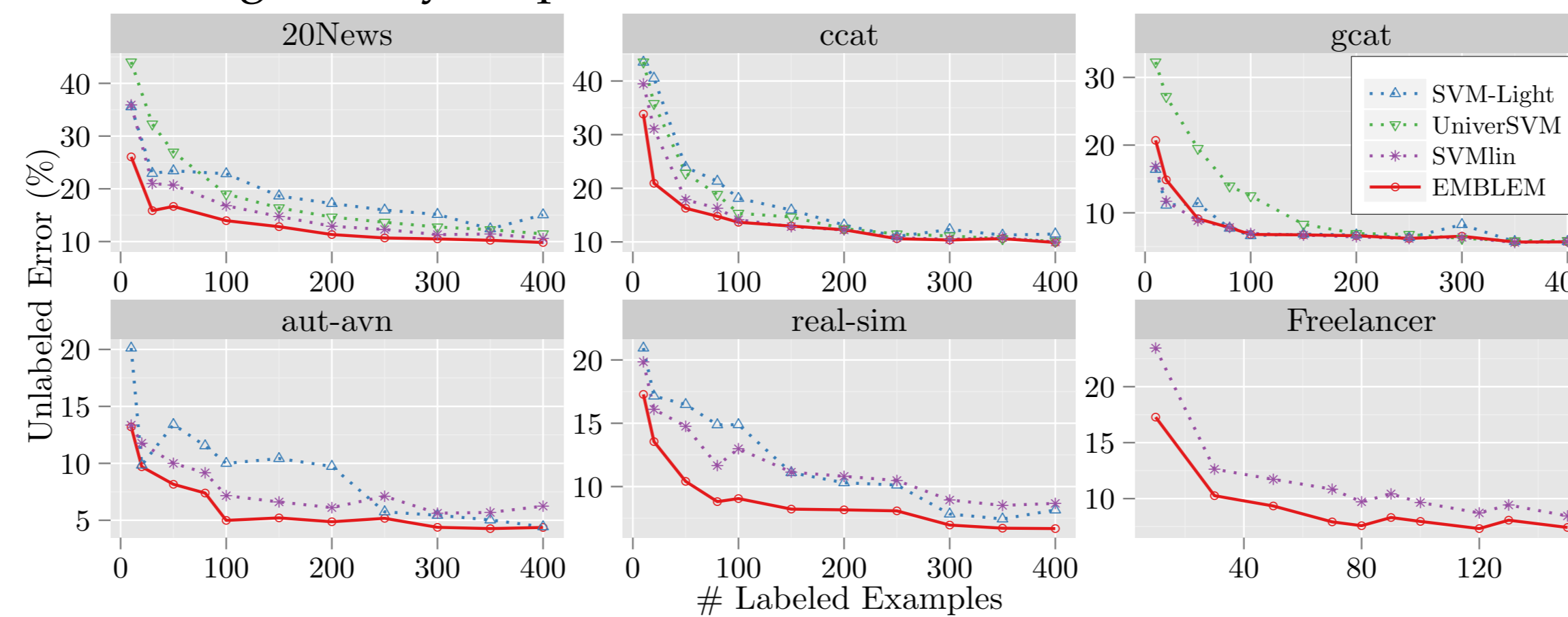
- Use transductive setting
 - Train on whole data set (partially labeled).
 - Test on unlabeled part.
- Perform 6 tasks in document classification from:

Corpus	Source	# Doc.	# Terms
20-News	UseNet articles	19K	61K
RCV1	News articles	23K	47K
SRAA	UseNet articles	72K	21K
Freelancer	Crowdsourcing postings	355K	28K

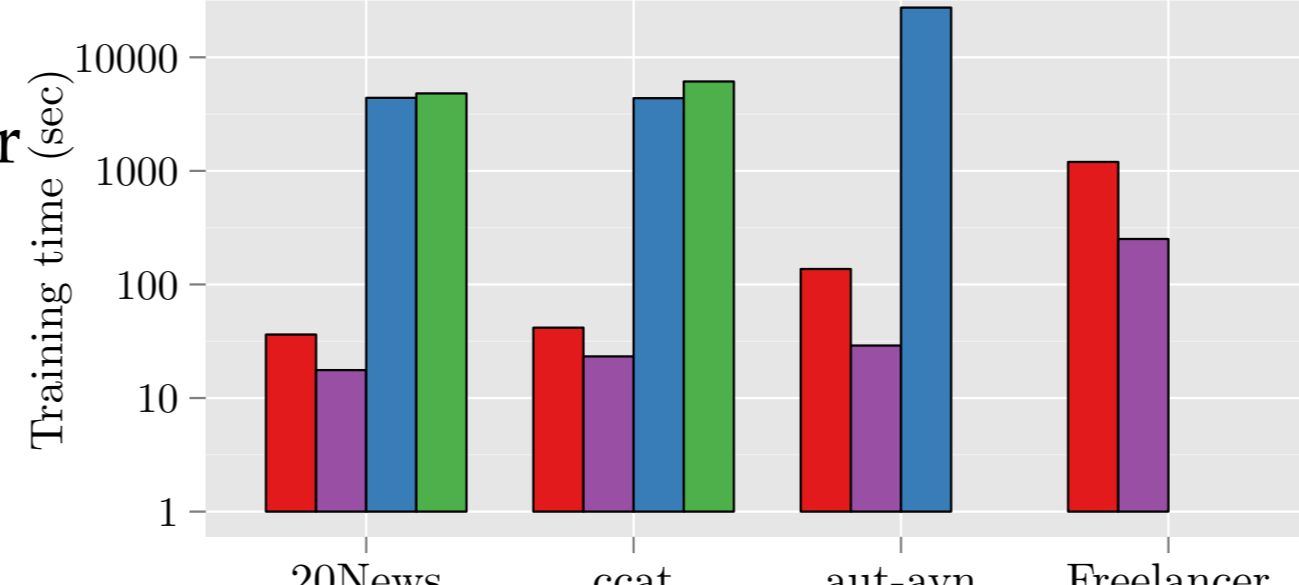
- Compare to three implementations of S³VMS:
 - SVM-Light [Joachims, ICML 1999].
 - UniverSVM [Collobert et al., JMLR 2006].
 - SVMLin [Sindhwani and Keerthi, SIGIR 2006].
- Refer to our algorithm as EMBLEM: EM algorithm for Binary Large Margin classification

Results

- EMBLEM generally outperforms S³VMS



- EMBLEM is orders-of-magnitude faster than SVM-Light and UniverSVM, but somewhat slower than SVMLin (EMBLEM is implemented in MATLAB, SVMLin is in C).



6. Advantages

- Handles unlabeled examples transparently – it differs only in the formula used to compute posterior means.
- Scales well – we can leverage highly optimized solvers for least-squares.
- Outperforms S³VMS in terms of accuracy.
- Easy to parallelize – E-step works on each example, and there are parallelized solvers for least squares.

Check out our code at:

<https://github.com/dokyum/EMBLEM>

References

- Chapelle and Zien. Semi-supervised classification by low density separation. AISTATS 2005.
- Collobert, Sinz, Weston, and Bottou. Large scale transductive svms. JMLR 2006.
- Joachims. Transductive inference for text classification using support vector machines. ICML 1999.
- Sindhwani and Keerthi. Large scale semi-supervised linear svms. SIGIR 2006.