Modeling Binary Data

Main Topic of our paper

Bernoulli-Logistic Latent Gaussian Models (bLGMs)

bLGMs - Classification Models

Bayesian Logistic Regression and Gaussian Process Classification


Figures reproduced using GPML toolbox
bLGMs - Latent Factor Models

Parameter Learning is Intractable

Logistic Likelihood is not conjugate to the Gaussian prior.

\[
\int (1 + \exp(z))^{-1} \times dZ
\]

We propose piecewise bounds to obtain tractable lower bounds to marginal likelihood.
Learning in bLGMs
Bernoulli-Logistic Latent Gaussian Models

\[ p(z) = \mathcal{N}(\mu, \Sigma) \]
\[ p(y_d = 1) = \sigma(w_d^T z) \]
\[ \sigma(x) = (1 + \exp(x))^{-1} \]

Parameter Set
\[ \Theta = \{\mu, \Sigma, W\} \]
Learning Parameters of bLGMs

\[ \mathcal{L}(\theta | y_1, y_2, \ldots, y_N) = \sum_{n=1}^{N} \log \int \prod_{d=1}^{D} p(y_{dn} | z, \theta) \mathcal{N}(z | \mu, \Sigma) dz \]

\[ \log \int \prod_{d=1}^{D} dz \times \]
Variational Lower Bound (Jensen’s)

\[ \mathcal{L}(\theta | y) = \log \int \prod_{d=1}^{D} p(y_d | z, \theta) \mathcal{N}(z | \mu, \Sigma) dz \]

\[ = \log \int \frac{\prod_{d=1}^{D} p(y_d | z, \theta) \mathcal{N}(z | \mu, \Sigma)}{\mathcal{N}(z | m, V)} \mathcal{N}(z | m, V) dz \]

\[ \geq \max_{m, V} \sum_{d=1}^{D} \int [\log p(y_d | z, \theta)] \mathcal{N}(z | m, V) dz - KL [\mathcal{N}(m, V) || \mathcal{N}(\mu, \Sigma)] \]

\[ = \max_{m, V} \sum_{d=1}^{D} \int [-\log(1 + e^{x_d})] \mathcal{N}(\tilde{m}_d, \tilde{v}_d) dx_d \quad + \quad \text{some other tractable terms in } m \text{ and } V \]
Quadratic Bounds

• Bohning’s bound (Bohning, 1992)
Quadratic Bounds

- Bohning’s bound (Bohning, 1992)
- Jaakkola’s bound (Jaakkola and Jordan, 1996)
- Both bounds have unbounded error.
Problems with Quadratic Bounds

1-D example with $\mu = 2$, $\sigma = 2$

$$p(y = 1|\mu, \sigma^2) = \int (1 + \exp(z))^{-1} \mathcal{N}(z|\mu, \sigma^2) dz$$

Generate data, fix $\mu = 2$, and compare marginal likelihood and lower bound wrt $\sigma$

As this is a 1-D problem, we can compute lower bounds without Jensen’s inequality. So plots that follow have errors only due to error in bounds.
Problems with Quadratic Bounds

Bohning

Jaakkola

Piecewise

$t_0 = -\infty$ $t_1$ $t_2$ $t_3 = \infty$
Piecewise Bounds
Finding Piecewise Bounds

- Find Cut points, and parameters of each pieces by minimizing maximum error.

- Linear pieces (Hsiung, Kim and Boyd, 2008)

- Quadratic Pieces (Nelder-Mead method)

- Fixed Piecewise Bounds!

- Increase accuracy by increasing number of pieces.
Linear Vs Quadratic

Maximum Error vs Number of Pieces
Results
Binary Factor Analysis (bFA)

- UCI voting dataset with D=15, N=435.
- Train-test split 80-20%
- Compare cross-entropy error on missing value prediction on test data.
bFA – Error vs Time

Imputation Error vs Time on Voting Data

Error

Bohning
Jaakkola
Piecewise Linear with 3 pieces
Piecewise Quad with 3 pieces
Piecewise Quad with 10 pieces

Time in Seconds
bFA – Error Across Splits

bFA on Voting data

Error with Piecewise Quadratic

Error with Bohning and Jaakkola

Bohning
Jaakkola
Gaussian Process Classification

- We repeat the experiments described in Kuss and Rasmussen, 2006
- We set $\mu = 0$ and squared exponential Kernel
- $\Sigma_{ij} = \sigma \exp\left[\frac{(x_i - x_j)^2}{s}\right]$
- Estimate $\sigma$ and $s$.
- We run experiments on Ionosphere ($D = 200$)
- Compare Cross-entropy Prediction Error for test data.
GP – Marginal Likelihood

Bohning-lowerBound

Jaakkola-lowerBound

PiecewiseQuad-lowerBound

EP-Approx

ICML 2011.
Mohammad Emtiyaz Khan
GP – Prediction Error
EP vs Variational

• We see that the variational approach underestimates the marginal likelihood in some regions of parameter space.

• However, both methods give comparable results for prediction error.

• In general, the variational EM algorithm for parameter learning is guaranteed to converge when appropriate numerical methods are used,

• Nickisch and Rasmussen (2008) describe the variational approach as more principled than EP.
Conclusions
Conclusions

• Fixed piecewise bounds can give a significant improvement in estimation and prediction accuracy relative to variational quadratic bounds.

• We can drive the error in the logistic-log-partition bound to zero by letting the number of pieces increase.

• This increase in accuracy comes with a corresponding increase in computation time.

• Unlike many other frameworks, we have a very fine grained control over the speed-accuracy trade-off through controlling the number of pieces in the bound.
Thank You
Piecewise-Bounds: Optimization Problem

\[
\begin{align*}
\min_{t,a} \quad & \max_{r \in \{1, \ldots, R\}} \quad \max_{t_{r-1} \leq x < t_r} \quad a_r x^2 + b_r x + c_r - \text{lse}(x) \\
\text{s.t.} \quad & \quad a_r x^2 + b_r x + c_r - \text{lse}(x) \geq 0 \quad \forall r \in \{1, \ldots, R\}, \forall x \in [t_{r-1}, t_r] \\
\end{align*}
\]

\[
\min_{t,a} \quad \max_{r \in \{1, \ldots, R\}} \left( \max_{t_{r-1} \leq x < t_r} a_r x^2 + b_r x - \text{lse}(x) \right) - \left( \min_{t_{r-1} \leq x < t_r} a_r x^2 + b_r x - \text{lse}(x) \right)
\]
\[ E_{q_n}(\mathbf{z}|\gamma_n)[\log p(\mathbf{y}_n|\mathbf{z}, \theta)] \]
\[ \geq \sum_{d=1}^{D} \left( y_{dn} \mathbf{W}_d^T \mathbf{m}_n - E_{q_n}(\mathbf{z}|\gamma_n)[B_{\alpha}(\mathbf{W}_d^T \mathbf{z})] \right) \]
\[ = \sum_{d=1}^{D} \left( y_{dn} \mathbf{W}_d^T \mathbf{m}_n - E_{q_n}(\eta|\tilde{\gamma}_{dn})[B_{\alpha}(\eta)] \right) \]
\[ \tilde{\gamma}_{dn} = \{ \tilde{m}_{dn}, \tilde{v}_{dn} \}, \quad \tilde{m}_{dn} = \mathbf{W}_d^T \mathbf{m}_n, \quad \tilde{v}_{dn} = \mathbf{W}_d^T \mathbf{V}_n \mathbf{W}_d \]
\[ E_{q_n}(\eta_{dn}|\tilde{\gamma}_{dn})[B_{\alpha}(\eta)] = \sum_{r=1}^{R} f_{r}(\tilde{m}_{dn}, \tilde{v}_{dn}, \alpha) \]
\[ = \sum_{r=1}^{R} \int_{t_{r-1}}^{t_r} (a_r \eta^2 + b_r \eta + c_r) \mathcal{N}(\eta|\tilde{m}_{dn}, \tilde{v}_{dn})d\eta \]
Algorithm 1 bLGM Generalized EM Algorithm

**E-Step:**

\[
\frac{\partial \mathcal{L}_{Q,J}}{\partial \mathbf{m}_{kn}} \leftarrow \sum_{d=1}^{D} y_{dn} \mathbf{W}_{dk} - \sum_{l=1}^{K} (\Sigma^{-1})_{lk} (\mathbf{m}_{ln} - \mu_l) - \sum_{r=1}^{R} \sum_{d=1}^{D} \mathbf{W}_{rk} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \alpha)}{\partial \tilde{m}_{dn}}
\]

\[
\frac{\partial \mathcal{L}_{Q,J}}{\partial \mathbf{V}_{kl}} \leftarrow \frac{1}{2} (\Sigma^{-1})_{kl} - \frac{1}{2} (\mathbf{V}_n^{-1})_{kl} - \sum_{r=1}^{R} \sum_{d=1}^{D} \mathbf{W}_{rk} \mathbf{W}_{dl} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \alpha)}{\partial \tilde{v}_{dn}}
\]

**M-Step:**

\[
\mu \leftarrow \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n
\]

\[
\Sigma \leftarrow \frac{1}{N} \sum_{n=1}^{N} (\mathbf{V}_n + (\mathbf{m}_n - \mu)(\mathbf{m}_n - \mu)^T)
\]

\[
\frac{\partial \mathcal{L}_{Q,J}}{\partial \mathbf{W}_{dk}} \leftarrow \sum_{n=1}^{N} \left[ \mathbf{m}_{kn} \left( y_{dn} - \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \alpha)}{\partial \tilde{m}_{dn}} \right) \right]
\]

\[
- \left( 2 \sum_{l=1}^{K} \mathbf{V}_{lkn} \mathbf{W}_{dk} \right) \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \alpha)}{\partial \tilde{v}_{dn}}
\]
(b) 5D bLGGM Covariance

(c) 5D bLGGM KL Divergence
Latent Gaussian Graphical Model

LED dataset, 24 variables, N=2000

\[
\begin{align*}
\mu & \quad \Sigma \\
\rightarrow & \\
\rightarrow & \quad \rightarrow \\
Z_{1n} & \quad Z_{2n} \quad \cdots \quad Z_{Dn} \\
\downarrow & \\
y_{1n} & \quad y_{3n} \quad \cdots \quad y_{Dn} \\
\end{align*}
\]

\(n=1:N\)
Sparse Version

$sBLGGM$ on LED data

Error vs. Regularization parameter $\lambda$

Error with $Q20$

Error with $B$ and $J$
Binary Latent Gaussian Models

We are interested in maximum likelihood estimate of parameters

\[ p(z) = \mathcal{N}(\mu, \Sigma) \]
\[ p(y_d = 1) = \sigma(w^T_d z) \]
\[ \sigma(x) = (1 + \exp(x))^{-1} \]

We are interested in maximum likelihood estimate of parameters

\[ \Theta = \{ \mu, \Sigma, W \} \]