Latent Gaussian Processes for Distribution Estimation of Multivariate Categorical Data


Yarin Gal    Yutian Chen    Zoubin Ghahramani

University of Cambridge

18 March 2016

Presented by: Kyle Ulrich
Multivariate categorical data appear in many applications
- e.g., medical diagnosis with various categorical test results for each patient

*Big labelled data* allows for measuring similarity between vectors of categorical variables

What about *small unlabelled data*?
- Distribution estimation
- Non-linear transformations (Gaussian processes?) capture multi-modality in distribution estimation
- Continuous latent representations address sparsity

Authors propose a model called the *categorical latent Gaussian process* (CLGP)
Proposed Model in Relation to the Literature

Gaussian Processes

A Gaussian process is a distribution over functions,

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$  \hspace{1cm} (1)

where $$x \in \mathbb{R}^Q$$ is an input variable. The mean function $$m(x)$$ and covariance kernel $$k(x, x')$$ are defined by\(^1\)

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x, x') = \text{cov}(f(x), f(x'))$$

Any collection of function evaluations are distributed according to

$$[f(x_1), f(x_2), \ldots, f(x_N)]^T \sim \mathcal{N}(0, K(X, X))$$  \hspace{1cm} (2)

where elements of the Gram matrix are $$(K(X, X))_{ij} = k(x_i, x_j)$$

\(^1m(x)$$ is typically set to 0 without loss of generality.
Gaussian Process Latent Variable Models

- $D$-dimensional data $\{y_n\}_{n=1}^N$, with $Y = [y_1, \ldots, y_N]^T$
- Each $y_n$ is located at $x_n \in \mathbb{R}^Q$, with $X = [x_1, \ldots, x_N]^T$
- $X$ is unobserved, to be inferred

Consider independent GPs across features:

$$y_d \sim \mathcal{N}(0, K_d(X, X) + \beta^{-1} I_N)$$  \hspace{1cm} (3)

The GP-LVM places a prior on the location of each $y_n$ with

$$x_n \sim \mathcal{N}(0, \sigma_x^2 I_Q)$$  \hspace{1cm} (4)

$X$ is inferred jointly with $\beta$ and $\theta_d$ (covariance kernel parameters)
Data:

- $D$-dimensional categorical data $\{y_n\}_{n=1}^N; \ Y = [y_1, \ldots, y_N]^T$
- Each variable $y_{nd}$ can take integer values $0, \ldots, K_d$
- All categorical variables have same cardinality $K_d \equiv K$

Authors seek to infer a latent variable $x_n \in \mathbb{R}^Q$ for these types of observations
The generative model for the **categorical latent Gaussian process:**

\[
\begin{align*}
  x_{nq} & \sim \mathcal{N}(0, \sigma_x^2) \quad (5) \\
  F_{dk}(\mathbf{x}) & \sim \mathcal{GP}(0, k_d(\mathbf{x}, \mathbf{x}')) \quad (6) \\
  f_{ndk} & = F_{dk}(\mathbf{x}_n) \quad (7) \\
  y_{nd} & \sim \text{Softmax}(f_{nd}) \quad (8)
\end{align*}
\]

where the Softmax distribution is given by

\[
\text{Softmax}(y = k; \mathbf{f}) = \text{Cat}\left( \frac{\exp(f_k)}{1 + \sum_{k'=1}^{K} \exp(f_{k'})} \right) \quad (9)
\]

for \( k = 0, \ldots, K \)
Inducing Point Approximation

Define $f_{dk} = (f_{1dk}, \ldots, f_{Ndk})$ and let $u_{dk} = (u_{1dk}, \ldots, u_{Mdk})$ be a set of $M$ ‘inducing points’ at the locations $z_1, \ldots, z_M$.

The model is adjusted such that $u_{mdk} = F_{dk}(z_m)$ and the function evaluations at $X$ are now independently *induced* through $u$:

$$p(f_{dk}|X, u_{dk}) = \prod_{n=1}^{N} p(f_{ndk}|x_n, u_{dk})$$  \hspace{1cm} (10)
The effective generative model for the CLGP now becomes:

\[ x_{nq} \sim \mathcal{N}(0, \sigma^2_x) \]
\[ \mathcal{F}_{dk}(x) \sim \mathcal{GP}(0, k_d(x, x')) \]
\[ u_{mdk} = \mathcal{F}_{dk}(z_m) \quad (11) \]
\[ f_{ndk} \sim \mathcal{N}(a^T_{nd}u_{dk}, b_{nd}) \quad (12) \]
\[ y_{nd} \sim \text{Softmax}(f_{nd}) \]

where

\[ a_{nd} = K_d(Z, Z)^{-1}K_d(Z, x_n) \quad (13) \]
\[ b_{nd} = K_d(x_n, x_n) - K_d(x_n, Z)a_{nd} \quad (14) \]

Known as the fully independent training conditional (FITC) approximation, GP complexity is reduced to \( \mathcal{O}(NM^2) \)
Consider a variational approximation to the posterior distribution

\[ q(X, F, U) = q(X)q(U)p(F|X, U) \]  \hspace{1cm} (15) 

The resulting variational lower bound on the log evidence is

\[
\log p(Y) \geq \mathcal{L}
= -\text{KL}(q(X)||p(X)) - \text{KL}(q(U)||p(U))
\]

\[
+ \sum_{n=1}^{N} \sum_{d=1}^{D} \int q(x_n)q(U_d)p(f_{nd}|x_n, U_d)\ 
\cdot \log p(y_{nd}|f_{nd})dx_ndf_{nd}dU_d \]  \hspace{1cm} (16) 

Inference Issues

Problems:

- Softmax likelihood is not conjugate to Gaussian prior
- Integrating the latent variables with a Softmax distribution is intractable

Solutions

- Use an alternative likelihood to the Softmax
- Further bound the ELBO
- Exploit recent advances in sampling-based variational inference
  - The authors choose this approach

---

2 e.g., Khan et al., “A stick-breaking likelihood for categorical data analysis with latent Gaussian models,” 2012
3 e.g., Blei & Lafferty, “Correlated topic models,” 2006
4 e.g., Blei et al., “Variational Bayesian inference with stochastic search,” 2012
Monte Carlo Gradient Estimation

Consider further the mean field approximation:

\[
q(U) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}(u_{dk} | \mu_{dk}, \Sigma_d)
\]

\[
q(X) = \prod_{n=1}^{N} \prod_{i=1}^{Q} \mathcal{N}(x_{ni} | m_{ni}, s_{ni}^2)
\]  \hspace{1cm} (17)

A deterministic axillary parametrization trick is useful\(^5\):

\[
x_{ni} = m_{ni} + s_{ni} \epsilon_{ni}^{(x)}, \quad \epsilon_{ni}^{(x)} \sim \mathcal{N}(0, 1)
\]  \hspace{1cm} (18)

\[
u_{dk} = \mu_{dk} + L_d \epsilon_{dk}^{(u)}, \quad \epsilon_{dk}^{(u)} \sim \mathcal{N}(0, I_M)
\]  \hspace{1cm} (19)

\[
f_{ndk} = a_{nd}^T u_{dk} + \sqrt{b_{nd}} \epsilon_{ndk}^{(f)}, \quad \epsilon_{ndk}^{(f)} \sim \mathcal{N}(0, 1)
\]  \hspace{1cm} (20)

\(^5\)See Kingma & Welling, “Auto-Encoding Variational Bayes,” 2014
Monte Carlo Gradient Estimation

\[ \mathcal{L} = -KL(q(X)||p(X)) - KL(q(U)||p(U)) \]
\[ + \sum_{n=1}^{N} \sum_{d=1}^{D} \mathbb{E}_{\epsilon(x), \epsilon(u), \epsilon(f)} \left[ \log \text{Softmax} \left( y_{nd} | f_{nd} \left( \epsilon_{nd}, U_d(\epsilon_u), x_n(\epsilon_x) \right) \right) \right] \]  \hspace{1cm} (21)

The Softmax likelihood term may be estimated with Monte Carlo integration:

\[ \mathcal{L}^{nd}_s \approx \frac{1}{T} \sum_{i=1}^{T} \log \text{Softmax} \left( y_{nd} | f_{nd} \left( \epsilon_{nd,i}, U_d(\epsilon_u), x_n(\epsilon_x) \right) \right) \]  \hspace{1cm} (22)

The derivatives of \( \mathcal{L} \) are now straightforward, and noisy gradient descent methods (e.g., RMSPROP) may be utilized.
A practical issue with stochastic approximate inference is the variance of the gradient approximation may be large\(^6\).

Authors suggest empirically this variance decreases with iteration.

Figure: ELBO and standard deviation per iteration for Alphadigits data.

\(^6\)Control variates may help by considering existing lower bounds, Paisley, Blei, & Jordan, “Variational Bayesian inference with stochastic search,” 2012.
Relational Learning – Exclusive Or

Train model with binary XOR observations

\[(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0)\]

Evaluate probability of model imputing new missing labels

\[(0, 0, ?), (0, 1, ?), (1, 0, ?), (1, 1, ?)\]

Figure shows \(p(y_d = 1 | x)\) as a function of \(x\) for \(d = 0, 1, 2\)
**Wisconsin breast cancer dataset**: 683 data points, 9 categorical variables (1–10), and 1 binary variable

On testing set (75% — 25%), randomly remove one of 10 categorical variables to impute

Use test-set perplexity to evaluate

Report three different splits of dataset

<table>
<thead>
<tr>
<th>Split</th>
<th>Baseline</th>
<th>Multinomial</th>
<th>Uni-Dir-Mult</th>
<th>Bi-Dir-Mult</th>
<th>LGM</th>
<th>CLGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.68</td>
<td>4.41</td>
<td>4.41</td>
<td>3.41</td>
<td>3.57 ± 0.208</td>
<td>2.86 ± 0.119</td>
</tr>
<tr>
<td>2</td>
<td>8.68</td>
<td>∞</td>
<td>4.42</td>
<td>3.49</td>
<td>3.47 ± 0.252</td>
<td>3.36 ± 0.186</td>
</tr>
<tr>
<td>3</td>
<td>8.85</td>
<td>4.64</td>
<td>4.64</td>
<td>3.67</td>
<td>12.13 ± 9.705</td>
<td>3.34 ± 0.096</td>
</tr>
</tbody>
</table>
Sparse Small Data – Handwritten Binary Alphadigits

10 × 8 binary images of 10 handwritten digits and 26 capital letters

Each class has 39 images (randomly split 30 training, 9 test)

On testing set, randomly remove 20% of all pixels

Use 2 latent dimensions for visual comparison