In short:

Multivariate categorical data –
- in data analysis, language processing, medical diagnosis...
- the number of possible vectors of observations $N$ grows exponentially with the number of discrete variables $D$ in the vector.
- the diversity of data points is poor compared to the exponentially many possible observations.

Example: Wisconsin breast cancer (Institute of Oncology, Ljubljana)

<table>
<thead>
<tr>
<th>#</th>
<th>Age</th>
<th>Education</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Negative</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Positive</td>
</tr>
</tbody>
</table>

| 683 patients, $2 \times 10^9$ possible configurations. |
- We develop a model for distribution estimation of multivariate categorical data:
  \[ P(y | \{y_1, \ldots, y_n\} | n = 1, \ldots, N) \]
- We use a continuous latent Gaussian space and learn a non-linear transformation between it and the multivariate categorical observation space.
- We derive inference for our model based on recent developments in sampling-based variational inference and stochastic optimisation.

Relation to other models

Existing approaches use –
- Discrete representations: based on frequencies of observations, but cannot handle sparse samples well (e.g. Dirichlet-Multinomial).
- Continuous representations: linearly transform a latent space before discretisation, but cannot capture multi-modality in the data (e.g. latent Gaussian model).

Method

1. Monte Carlo integration approximates the likelihood obtaining noisy gradients:
   \[ E_{\mathbf{z}, \mathbf{z}'} [\text{log Softmax}(\mathbf{y} \mid \mathbf{z}, \mathbf{z}') \mid \mathbf{x}] = \frac{1}{N} \sum_{n=1}^{N} \text{log Softmax}(\mathbf{y}_n \mid \mathbf{f}(\mathbf{z}_n), U_j(\mathbf{z}_n), x_n(\mathbf{z}_n))) \]
2. Learning-rate free stochastic optimisation is used to optimise the noisy objective.
3. Symbolic differentiation is used to get simple and modular code:

Experiments

Test perplexity predicting randomly missing values

<table>
<thead>
<tr>
<th></th>
<th>Uniform</th>
<th>Multinomial</th>
<th>Bi-Dir-Mult</th>
<th>LGM</th>
<th>CLGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin breast cancer</td>
<td>8.88</td>
<td>4.41</td>
<td>3.57</td>
<td>3.41</td>
<td>2.39</td>
</tr>
</tbody>
</table>

Data Visualisation

LGM latent space
CLGP latent space

Example alphadigits

Code available at http://github.com/yaringal/CLGP.