The Supervised IBP: Neighbourhood Preserving Infinite Latent Feature Models

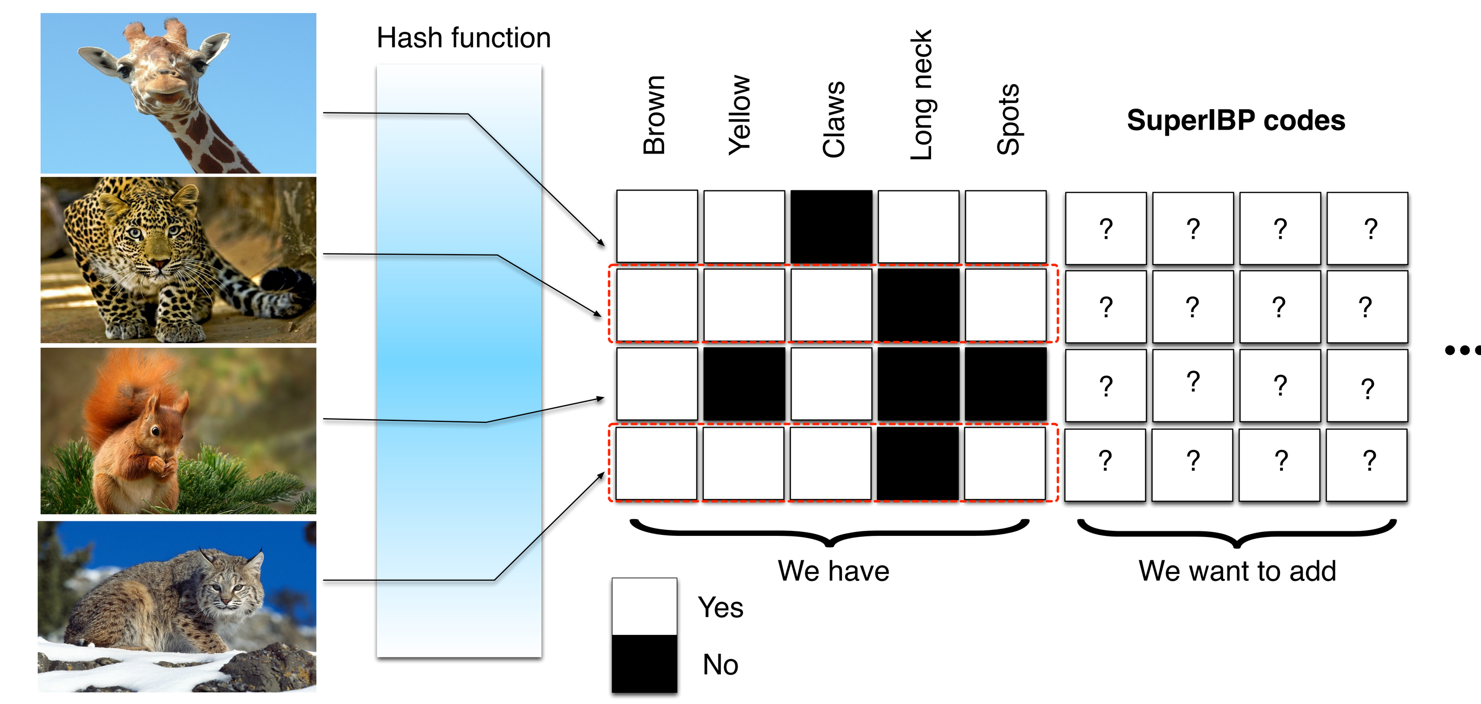
Novi Quadrianto[1] Viktoriia Sharmanska[2] David A. Knowles[3] Zoubin Ghahramani[1]
1: Cambridge Machine Learning Group, University of Cambridge 2: IST Austria (Institute of Science and Technology Austria) 3: Computer Science Department, Stanford University

Abstract

WHAT: a probabilistic model to infer binary latent variables that preserve neighbourhood structure of the data

- **WHY**: to perform a nearest neighbour search for the purpose of retrieval
- **WHEN**: in dynamic and streaming nature of the Internet data
- **HOW**: the Indian Buffet Process prior coupled with a preference relation
- **WHERE**: dynamic extension of hash codes

Motivating Example: Dynamic Hash Codes Extension



Neighbourhood Relation

- Supervision is in the form of triplets ($sample$ i, its $neighbour$ j, its non-$neighbour$ l).
- Triplets encode neighbourhood preference relation $j \succ_i l$: sample i prefers its neighbour j to its non-neighbour l.

$$\Pr(j \succ_i l | \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_l, \mathbf{w}) = \frac{1}{C} \sum_k w_k \mathbb{I}[z_i^k = z_j^k](1 - \mathbb{I}[z_i^k = z_l^k])$$

- The label preference likelihood function is given by

$$\Pr(\mathbf{T}|\mathbf{Z}, \mathbf{w}) = \prod_{(i,j,l) \in \mathcal{T}} \Pr(j \succ_i l | \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_l, \mathbf{w}).$$

The Supervised IBP: $Z \to X$ Linear Gaussian Model

- **Data related likelihood:**
  the data point $\mathbf{x}_n \in \mathbb{R}^M$ is generated as follows:

$$\mathbf{x}_n | \mathbf{z}_n, \mathbf{V}, \sigma_x \sim \mathcal{N}(\mathbf{V}\mathbf{z}_n, \sigma_x^2 \mathbf{I}), \quad \mathbf{V} \in \mathbb{R}^{M \times K}.$$

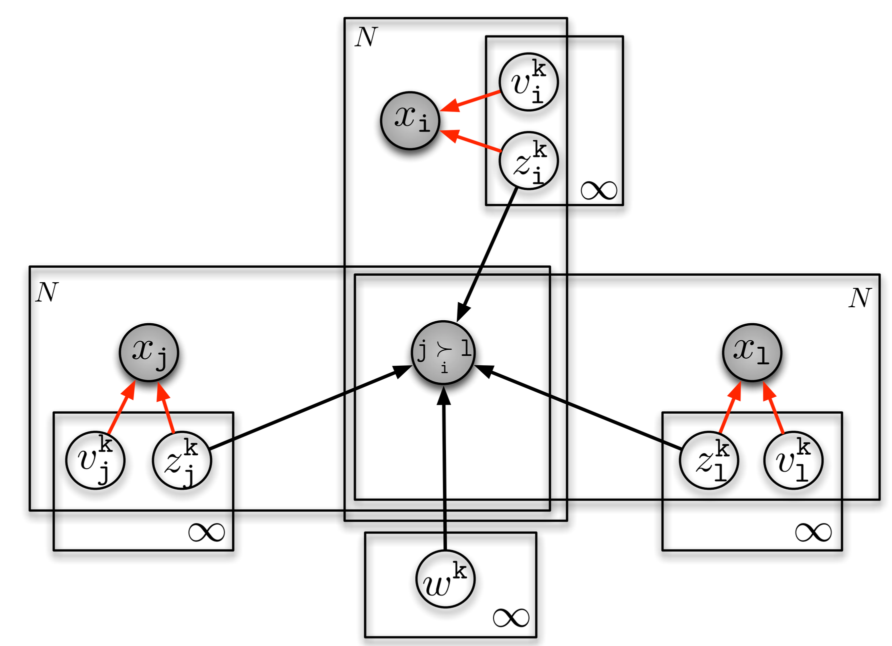
- **Inference**: alternating between M-H sampling on Z, and slice sampling on w.
- **Prediction**: for a $test$ point $\mathbf{x}_* \in \mathbb{R}^M$:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{x}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Z}\mathbf{Z}^\top + \sigma_x^2/\sigma_v^2 \mathbf{I} & \mathbf{Z}\mathbf{z}_*{}^\top \\ \mathbf{z}_*\mathbf{Z}^\top & \mathbf{z}_*\mathbf{z}_*{}^\top + \sigma_x^2/\sigma_v^2 \mathbf{I} \end{bmatrix}\right)$$

and conditional $\Pr(\mathbf{x}_* | \mathbf{z}_*, \mathbf{Z}, \mathbf{X}) \sim \mathcal{N}(\mu_*, \Sigma_*)$ with

$$\mu_* = \mathbf{z}_*(\mathbf{Z}^\top\mathbf{Z} + \sigma_x^2/\sigma_v^2 \mathbf{I})^{-1}\mathbf{Z}^\top\mathbf{X}$$
$$\Sigma_* = \mathbf{z}_*\mathbf{z}_*{}^\top - \mathbf{z}_*(\mathbf{Z}^\top\mathbf{Z} + \sigma_x^2/\sigma_v^2 \mathbf{I})^{-1}\mathbf{Z}^\top\mathbf{Z}\mathbf{z}_*{}^\top.$$

We use the predictive mean of $\Pr(\mathbf{x}_* | \mathbf{z}_*, \mathbf{Z}, \mathbf{X})$ and approximate $\mathbf{z}_*$ by solving a linear system of equations, resulting in a continuous estimate $\hat{\mathbf{z}}_*$ of the binary vector $\mathbf{z}_*$.



Posterior probability: $\Pr(\mathbf{Z}, \mathbf{w} | \mathbf{X}, \mathbf{T}) \propto$

$$\Pr(\mathbf{T}|\mathbf{Z}, \mathbf{w}) \int \Pr(\mathbf{X}|\mathbf{Z}, \mathbf{V}, \sigma_x)\Pr(\mathbf{V}|\sigma_v) d\mathbf{V} \Pr(\mathbf{w}|\gamma_w, \theta_w)\Pr(\mathbf{Z}|\alpha)$$

The Supervised IBP: $X \to Z$ Linear Probit Model

- **Standard stick-breaking IBP**:

$$z_n^k | b_k \sim \text{Bernoulli}(b_k); \; b_k = v_k b_{k-1}; \; v_j \sim \text{Beta}(\alpha, 1), \; b_0 = 1.$$

Bernoulli(b_k) random variable can be represented as:

$$z_n^k | b_k = \mathbb{I}[u_n^k < \Phi^{-1}_{\mu,\sigma^2}(b_k)], \quad u_n^k \sim \mathcal{N}(\mu, \sigma^2).$$

- **Data dependent IBP**
  by linearly parameterising the cut off variable u_n^k:

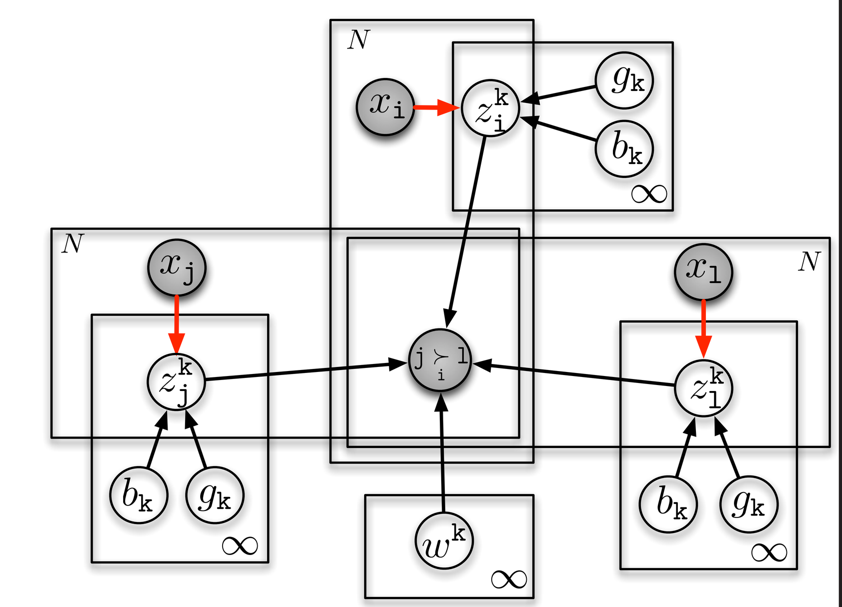
$$u_n^k \mid \mathbf{x}_n, \mathbf{g}_k \;\sim\; \mathcal{N}(u_n^k | -\mathbf{x}_n^\top \mathbf{g}_k, 1),$$

$\mathbf{g}_k \in \mathbb{R}^M$ is a vector of regression coefficients for each feature k.
Feature presence probability:

$$z_n^k | \mathbf{x}_n, \mathbf{g}_k, b_k \sim \text{Bernoulli}(\Phi_{0,1}(\mathbf{x}_n^\top \mathbf{g}_k + \Phi^{-1}_{\mu,\sigma^2}(b_k))).$$

- **Inference**: We adapt a slice sampling procedure with stick breaking representation of Teh et al., 2007, and use elliptical slice sampling (Murray et al., 2010) for updating $\mathbf{g}_k$.
- **Prediction**: For a new test point $\mathbf{x}_* \in \mathbb{R}^M$:

$$\Pr(\mathbf{z}_*^k = 1 | \mathbf{G}, b_k, \mathbf{x}_*) = \Phi_{0,1}(\mathbf{x}_*^\top \mathbf{g}_k + \Phi^{-1}_{0,1}(b_k)).$$

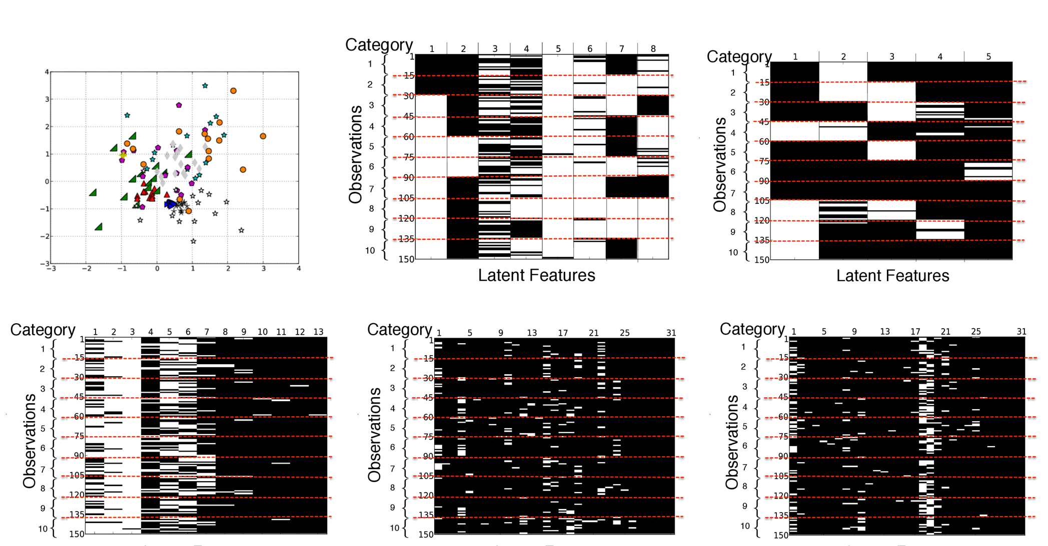


Posterior: $\Pr(\mathbf{Z}, \mathbf{b}, \mathbf{w}, \mathbf{G} | \mathbf{X}, \mathbf{T}) \propto$

$$\Pr(\mathbf{T}|\mathbf{Z}, \mathbf{w})\Pr(\mathbf{Z}|\mathbf{X}, \mathbf{G}, \mathbf{b})\Pr(\mathbf{w}|\gamma_w, \theta_w)\Pr(\mathbf{G}|\sigma_g)\Pr(\mathbf{b}|\alpha).$$

Experimental Results

Synthetic Experiments



Extending Hash Codes

- Dataset: Animals with Attributes with 30,475 images and 50 classes.
  Features: Colour histograms with a codebook of size 128.
  Performance metric: k-NN accuracy.
- Hash codes: Each class is assigned an Osherson's attribute binary string of length D. We learn D logistic regression functions, one for each bit position. When a new data point arrives, we evaluate each of the D $test$ logistic regressors to generate a D-bit hash code.
- Setup: 27,032 images from 45 classes define $initial$ $image$ $corpus$ for learning the hash codes. From the remaining five classes, we randomly sample 300 images with uniform class proportions to form a $refinement$ set for training, and the test set using 50/50 split. Likewise for a refinement set with 10 new classes.

	Hash	IBP [1]	dd-IBP [3] Input	dd-IBP [3] Output	Super IBP LGM (ours)	Super IBP LPM (ours)	Bayesian Averaging	Reference
5 animal categories								
1 NN	26.3 ± 2.2	30.3 ± 2.0	27.9 ± 6.3	29.7 ± 3.5	33.8 ± 1.6	$\mathbf{42.8 \pm 2.4}$	-0.2%	$\mathbf{40.9 \pm 4.7}$
3 NN		29.5 ± 2.9	29.6 ± 3.6	31.0 ± 1.4	34.6 ± 2.3	$\mathbf{41.9 \pm 3.4}$	+5.3%	$\mathbf{40.2 \pm 3.4}$
15 NN		31.5 ± 2.6	27.8 ± 2.8	28.1 ± 3.2	35.5 ± 1.0	$\mathbf{44.5 \pm 2.1}$	+4.6%	39.3 ± 3.7
30 NN		29.5 ± 3.2	24.3 ± 3.0	23.6 ± 3.4	33.8 ± 0.7	$\mathbf{45.9 \pm 4.1}$	+1.5%	36.1 ± 2.8
10 animal categories								
1 NN	12.7 ± 2.5	17.1 ± 3.1	12.9 ± 2.6	15.9 ± 2.3	17.3 ± 1.2	$\mathbf{25.0 \pm 2.9}$	+3.2%	$\mathbf{25.0 \pm 2.2}$
3 NN		17.9 ± 2.8	13.1 ± 2.4	15.3 ± 2.3	18.2 ± 1.2	$\mathbf{25.1 \pm 3.0}$	+8.2%	$\mathbf{26.0 \pm 1.7}$
15 NN		16.4 ± 2.5	14.7 ± 2.3	15.1 ± 1.8	18.0 ± 1.5	$\mathbf{26.6 \pm 2.7}$	+5.2%	$\mathbf{27.8 \pm 2.8}$
30 NN		17.7 ± 3.4	14.5 ± 1.9	14.0 ± 1.9	18.3 ± 1.4	$\mathbf{27.5 \pm 2.4}$	-1.0%	$\mathbf{25.8 \pm 1.4}$

[1] T. Griffiths and Z. Ghahramani, Infinite latent feature models and the IBP, NIPS, 2005
[2] Y. W. Teh, D. Görür, and Z. Ghahramani, Stick-breaking construction for the IBP, AISTATS, 2007
[3] S. J. Gershman, P. I. Frazier and D. M. Blei, Distance dependent infinite latent feature models, arXiv, 2012
[4] I. Murray, R. P. Adams, and D. J. C. MacKay, Elliptical slice sampling, AISTATS, 2010