Group meeting - February 14, 2020

Paper review & Research progress

Truong Son Hy *

*Department of Computer Science
The University of Chicago

Ryerson Physical Lab

GNNExplainer: Generating Explanations for Graph Neural Networks

GNNExplainer: Generating Explanations for Graph Neural Networks

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec
Stanford University, Department of Computer Science

Problem Setup

Understanding GNN's predictions is important and useful for several reasons:

a) Can increase trust in predictions and give insights into important graph structures.

b) Can improve model transparency and interpretability in decision-critical applications

c) Can identify and correct systematic patterns of mistakes made by GNNs before deploying them in the real world.

GNNExplainer, the first approach for explaining predictions made by GNNs:

- **Any machine learning task on graphs**. It is model-agnostic and provides explanations for node classification, link prediction, graph classification.
- **Any GNN model**. It is model-agnostic and can explain predictions of any GNN variant using the message-passing neural architecture.
- **Rich graph structural information and node features.** It incorporates relevant node features and relational information, the essence of graphs.

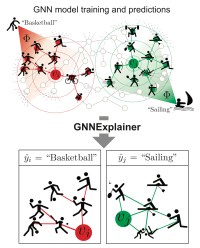
Given a trained GNN and its prediction(s), GNNExplainer returns an explanation in the form of a small subgraph of the input graph together with a small subset of node features that are most influential for the prediction(s).

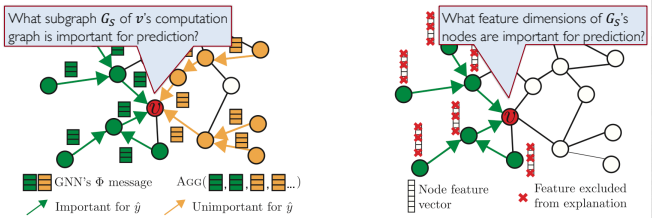
Key Idea and Explanation Mask

A GNN model Φ is trained on a social interaction graph to predict future sport activities.

Given a prediction $\hat{y}_i$ = "Basketball" for person v_i, GNNExplainer finds a small subgraph of the input graph together with a small subset of node features that are most influential for $\hat{y}_i$. Examining explanation for $\hat{y}_i$, we see that many friends in one part of vi's social circle enjoy ball games, and so the GNN predicts that v_i will like basketball.

Similarly, examining explanation for $\hat{y}_j$, we see that v_j's friends and friends of his friends enjoy water and beach sports, and so the GNN predicts $\hat{y}_j$ = "Sailing."





GNNExplainer: Design and Results

Objective: Maximize mutual information between the prediction and the explanation:

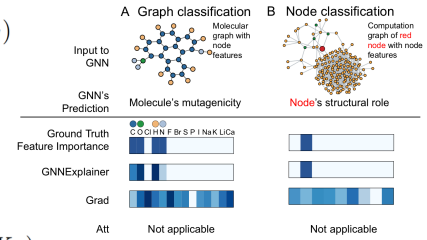
$$\max_{G_S} MI\,(Y, (G_S, X_S)) = H(Y) - H(Y|G = G_S, X = X_S)$$

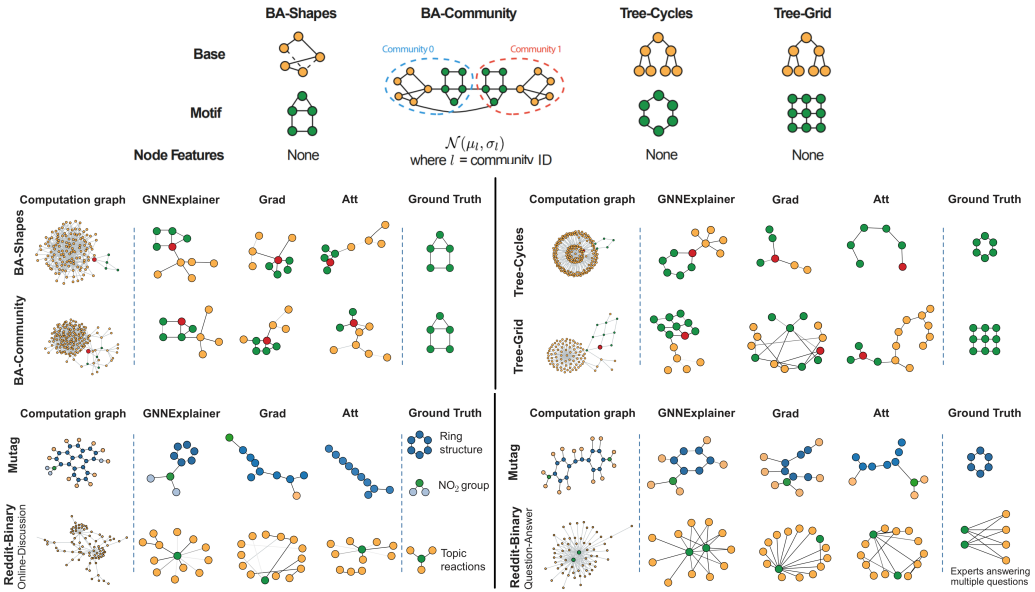
Conditional Entropy:

$$\min_{\mathcal{G}} \mathbb{E}_{G_S \sim \mathcal{G}} H(Y|G = G_S, X = X_S)$$

Tractable approximation via masking the graph structure and node features::

$$\min_{M} -\sum_{c=1}^{C} \mathbb{1}[y = c] \log P_\Phi(Y = y|G = A_S \odot \sigma(M), X = X_S)$$





GNNExplainer: Generating Explanations for Graph Neural Networks. *NeurIPS 2019.*

http://cs.stanford.edu/~name name = {rexy, jiaxuan, marinka, jure}

Graph Transformer Networks

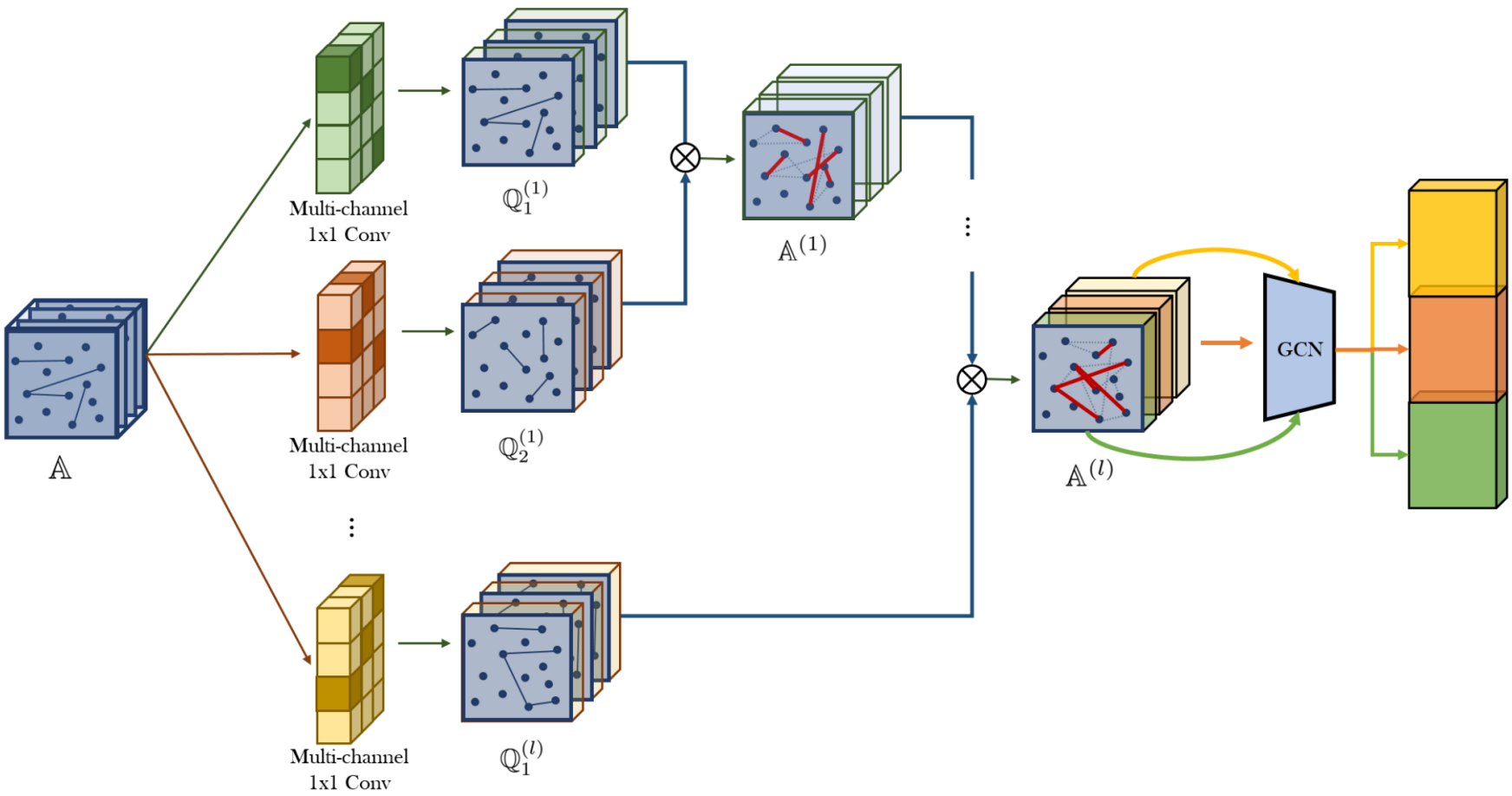


Figure 2: Graph Transformer Networks (GTNs) learn to generate a set of new meta-path adjacency matrices $\mathbb{A}^{(l)}$ using GT layers and perform graph convolution as in GCNs on the new graph structures. Multiple node representations from the same GCNs on multiple meta-path graphs are integrated by concatenation and improve the performance of node classification. $\mathbb{Q}_1^{(l)}$ and $\mathbb{Q}_2^{(l)} \in \mathbf{R}^{N \times N \times C}$ are intermediate adjacency tensors to compute meta-paths at the lth layer.

Questions for nCRP

1. How to initialize the topic-hierarchy tree given a set of documents?
2. How to adjust the tree given new documents? The algorithm is not clear to me.
3. Is each word in a document labeled to a particular topic? Or is it unsupervised in which the topics are **abstract**?
4. Maybe this work **Autoencoding Variational Inference For Topic Models** (ICLR 2017) is the right way to go? I did not have time to cover it yet.

Hierarchical Topic Models and the Nested CRP (1)

Problem

We address the problem of learning topic hierarchies from data. The model selection problem in this domain is daunting - which of the **large collection of possible trees** to use?

Hierarchical Topic Models and the Nested CRP (1)

Problem

We address the problem of learning topic hierarchies from data. The model selection problem in this domain is daunting - which of the **large collection of possible trees** to use?

Proposal

- We take a Bayesian approach, generating an appropriate prior via a distribution on partitions that we refer to as the nested **Chinese restaurant process** (CRP).
- We build a hierarchical topic model by combining this prior with a likelihood that is based on a hierarchical variant of **latent Dirichlet allocation**.

Hierarchical Topic Models and the Nested CRP (2)

Problem: Learning a topic hierarchy from data

Given a collection of **documents**, each of which contains a set of **words** we wish to discover common usage patterns or **topics** in the documents, and to organize these topics into a hierarchy.

Hierarchical Topic Models and the Nested CRP (2)

Problem: Learning a topic hierarchy from data

Given a collection of **documents**, each of which contains a set of **words** we wish to discover common usage patterns or **topics** in the documents, and to organize these topics into a hierarchy.

Approach

- Specify a **generative probabilistic model for hierarchical structure**.
- **Our hierarchies are random variables**; moreover, these random variables are specified procedurally, according to an algorithm that constructs the hierarchy as data are made available.

Hierarchical Topic Models and the Nested CRP (3)

Approach

- A distribution on partitions of integers known as the **Chinese restaurant process**.
- Extend the Chinese restaurant process to a hierarchy of partitions.
- Each node in the hierarchy is associated with a topic, where a topic is a distribution across words.
- A document is generated by choosing a path from the root to a leaf, repeatedly sampling topics along that path, and sampling the words from the selected topics.
- This approach differs from models of topic hierarchies which are built on the premise that the distributions associated with parents and children are similar $\rightarrow$ No such constraint.

Hierarchical Topic Models and the Nested CRP (4)

The Chinese restaurant process (CRP) is a distribution on partitions of integers obtained by imagining a process by which M customers sit down in a Chinese restaurant with an infinite number of tables. The m-th subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} \text{p(occupied table } i \mid \text{previous customers)} &= \frac{m_i}{\gamma+m-1} \\ \text{p(next unoccupied table} \mid \text{previous customers)} &= \frac{\gamma}{\gamma+m-1} \end{aligned}$$

where m_i is the number of previous customers at table i, and γ is a parameter. After M customers sit down, the seating plan gives a partition of M items.

Hierarchical Topic Models and the Nested CRP (4)

The Chinese restaurant process (CRP) is a distribution on partitions of integers obtained by imagining a process by which M customers sit down in a Chinese restaurant with an infinite number of tables. The m-th subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} \text{p(occupied table } i \mid \text{previous customers)} &= \frac{m_i}{\gamma+m-1} \\ \text{p(next unoccupied table} \mid \text{previous customers)} &= \frac{\gamma}{\gamma+m-1} \end{aligned}$$

where m_i is the number of previous customers at table i, and γ is a parameter. After M customers sit down, the seating plan gives a partition of M items.

Unrelated

This makes me remember **Hilbert's paradox of the Grand Hotel**!

Hierarchical Topic Models and the Nested CRP (5)

A nested Chinese restaurant process:

- There are an infinite number of infinite-table Chinese restaurants in a city.
- One restaurant is determined to be the **root restaurant** and on each of its infinite tables is a card with the name of another restaurant. And so on.
- On the first evening, a customer enters the root Chinese restaurant and selects a table. On the second evening, he goes to the restaurant identified on the first night's table and chooses another table.
- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree.

Hierarchical Topic Models and the Nested CRP (6)

Consider a data set composed of a **corpus** of documents. Each document is a collection of **words**, where a **word** is an item in a **vocabulary**.

Hierarchical Topic Models and the Nested CRP (6)

Consider a data set composed of a **corpus** of documents. Each document is a collection of **words**, where a **word** is an item in a **vocabulary**.

Assumption

Our basic assumption is that the words in a document are generated according to a mixture model where the mixing proportions are random and document-specific.

Hierarchical Topic Models and the Nested CRP (7)

Document-specific mixture distribution:

$$p(w|\theta) = \sum_{i=1}^{K} \theta_i p(w|z = i, \beta_i)$$

where:

- θ: document-specific mixing proportions.
- Assume K different topics in the corpus.
- z ranges over K possible values.
- β is a parameter.
- $p(w|z, \beta)$ is a distribution over words.

Hierarchical Topic Models and the Nested CRP (8)

Two-level generative probabilistic process for generating adocument:

1. Choose a K-vector θ of topic proportions from a distribution $p(\theta|\alpha)$, where α is a corpus-level parameter.
2. Repeatedly sample words from the mixture distribution $p(w|\theta)$ for the chosen value of θ.

When the distribution $p(\theta|\alpha)$ is chosen to be a Dirichlet distribution, we obtain the **latent Dirichlet allocation** model (LDA).

Hierarchical Topic Models and the Nested CRP (9)

Extend the process to a hierarchy of topics:

1. Suppose we are given an L-level tree and each node is associated with a topic.
2. Choose a path from the root of the tree to a leaf.
3. Draw a vector of topic proportions θ from an L-dimensional Dirichlet.
4. Generate the words in the document from a mixture of the topics along the path from root to leaf, with mixing proportions θ.

Hierarchical Topic Models and the Nested CRP (10)

Use the nested CRP to relax the assumption of a fixed tree structure. The nested CRP can be used to place a prior on possible trees $\rightarrow$ **Hierarchical LDA** (hLDA):

1. Let c_1 be the root topic.
2. For each level $\ell \in \{2, .., L\}$:
   - Draw a topic from $c_{\ell-1} \rightarrow c_\ell$.
3. Draw an L-dimensional topic proportion vector θ from Dir(α).
4. For each word $n \in \{1, .., N\}$:
   - Draw $z \in \{1, .., L\}$ from Multinomial(θ).
   - Draw w_n from the topic w_z.

Hierarchical Topic Models and the Nested CRP (11)

Suppose we are given a corpus of M documents $\boldsymbol{w}_1, .., \boldsymbol{w}_M$:

- $w_{m,n}$ is the n-th word of the m-th document.
- $c_{m,\ell}$ is the restaurant/node corresponding to the ℓ-th topic (at level ℓ-th of the tree) in document m.
- $z_{m,n}$ is the assignment of the n-th word in the m-th document to one of the L available topics.

Hierarchical Topic Models and the Nested CRP (12)

Conceptually, we divide the Gibbs sampler into two parts. First, given the current state of the CRP, we sample the $z_{m,n}$ variables of the underlying LDA model following the algorithm developed in [12], which we do not reproduce here. Second, given the values of the LDA hidden variables, we sample the $c_{m,\ell}$ variables which are associated with the CRP prior. The conditional distribution for $\mathbf{c}_m$, the L topics associated with document m, is:

$$p(\mathbf{c}_m \mid \mathbf{w}, \mathbf{c}_{-m}, \mathbf{z}) \propto p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) p(\mathbf{c}_m \mid \mathbf{c}_{-m}),$$

where $\mathbf{w}_{-m}$ and $\mathbf{c}_{-m}$ denote the $\mathbf{w}$ and $\mathbf{c}$ variables for all documents other than m. This expression is an instance of Bayes' rule with $p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z})$ as the likelihood of the data given a particular choice of $\mathbf{c}_m$ and $p(\mathbf{c}_m \mid \mathbf{c}_{-m})$ as the prior on $\mathbf{c}_m$ implied by the nested CRP. The likelihood is obtained by integrating over the parameters β, which gives:

$$p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) = \prod_{\ell=1}^{L} \left(\frac{\Gamma(n^{(\cdot)}_{c_{m,\ell},-m} + W\eta)}{\prod_w \Gamma(n^{(w)}_{c_{m,\ell},-m} + \eta)} \frac{\prod_w \Gamma(n^{(w)}_{c_{m,\ell},-m} + n^{(w)}_{c_{m,\ell},m} + \eta)}{\Gamma(n^{(\cdot)}_{c_{m,\ell},-m} + n^{(\cdot)}_{c_{m,\ell},m} + W\eta)} \right),$$

where $n^{(w)}_{c_{m,\ell},-m}$ is the number of instances of word w that have been assigned to the topic indexed by $c_{m,\ell}$, not including those in the current document, W is the total vocabulary size, and $\Gamma(\cdot)$ denotes the standard gamma function. When $\mathbf{c}$ contains a previously unvisited restaurant, $n^{(w)}_{c_{m,\ell},-m}$ is zero.

Variational Inference for the Nested Chinese Restaurant Process (1)

CRP:

$$p(c_d = k|c_{1:(d-1)}) \propto \begin{cases} m_k & \text{if } k \text{ is previous occupied} \\ \gamma & \text{if } k \text{ is a new table,} \end{cases}$$

nCRP:

1. Draw a path $\boldsymbol{c}_n|\boldsymbol{c}_{1:(n-1)} \sim \text{nCRP}(\gamma, \boldsymbol{c}_{1:(n-1)})$, which contains L nodes from the tree.
2. Draw a latent variable $\boldsymbol{x}_n \sim p(\boldsymbol{x}_n|\lambda)$.
3. Draw an observation $\boldsymbol{t}_n \sim p(\boldsymbol{t}_n|\boldsymbol{W}_{\boldsymbol{c}_n}, \boldsymbol{x}_n, \tau)$.

The parameters λ and τ are associated with the latent variables $\boldsymbol{x}$ and data generating distribution, respectively. Note that $\boldsymbol{W}_{\boldsymbol{c}_n}$ contains the $\boldsymbol{w}_i$s selected by the path $\boldsymbol{c}_n$. Specific applications of the nCRP mixture depend on the particular forms of $p(\boldsymbol{w})$, $p(\boldsymbol{x})$ and $p(\boldsymbol{t}|W_{\boldsymbol{c}}, \boldsymbol{x})$.

Variational Inference for the Nested Chinese Restaurant Process (2)

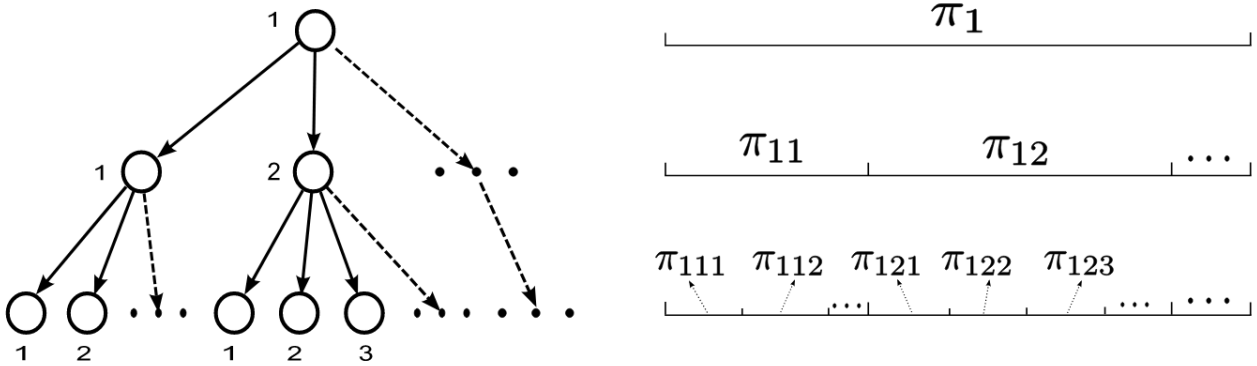


Figure 1: **Left.** A possible tree structure in a 3-level nCRP. **Right.** The tree-based stick-breaking construction of a 3-level nCRP.

Variational Inference for the Nested Chinese Restaurant Process (3)

CRP mixtures can be equivalently formulated using the Dirichlet process (DP) as a distribution over the distribution of each data point's random parameter [21, 4]. An advantage of expressing the CRP mixture with a DP is that the draw from the DP can be explicitly represented using the stick-breaking construction [22]. The DP bundles the scaling parameter γ and base distribution G_0. A draw from a DP(γ, G_0) is described as

$$v_i \sim \text{Beta}(1, \gamma), \;\; \pi_i = v_i \prod_{j=1}^{i-1}(1 - v_j), \;\; \boldsymbol{w}_i \sim G_0, \;\; i \in \{1, 2, \cdots\}, \;\; G = \sum_{i=1}^{\infty} \pi_i \delta_{\boldsymbol{w}_i},$$

where $\boldsymbol{\pi}$ are the stick lengths, and $\sum_{i=1}^{\infty} \pi_i = 1$ almost surely. This representation also illuminates the discreteness of a distribution drawn from a DP.

Variational Inference for the Nested Chinese Restaurant Process (4)

The tree-based stick-breaking construction lets us calculate the conditional probability of a path given $\boldsymbol{V}$. Let the path $\boldsymbol{c} = [1, c_2, \cdots, c_L]$,

$$p(\boldsymbol{c}|\boldsymbol{V}) = \prod_{\ell=1}^{L} \pi_{1,c_2,\cdots,c_\ell} = \prod_{\ell=1}^{L} v_{1,c_2,\cdots,c_\ell} \prod_{j=1}^{c_\ell - 1} (1 - v_{1,c_2,\cdots,j}). \tag{2}$$

By integrating out $\boldsymbol{V}$ in Equation 2, we recover the nCRP. Given Equation 2, the joint probability of a data set under the nCRP mixture is

$$p(\boldsymbol{t}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{c}_{1:N}, \boldsymbol{V}, \boldsymbol{W}) = p(\boldsymbol{V})p(\boldsymbol{W}) \prod_{n=1}^{N} p(\boldsymbol{c}_n|\boldsymbol{V})p(\boldsymbol{x}_n)p(\boldsymbol{t}_n|\boldsymbol{W}_{\boldsymbol{c}_n}, \boldsymbol{x}_n). \tag{3}$$

This representation is the basis for variational inference.

Autoencoding Variational Inference For Topic Models

Maybe this work is the right way to go? I did not have time to cover this paper yet.

Q & A

Thank you very much for your attention!