GNN caste prediction

Goal

The goal for this exercise is to predict the caste for a given household having household characteristics and graph network adjacency matrices. This is an intermediate step for the structural model simulating information diffusion based on Banerjee et al (2013) paper.

Setup

For the "Diffusion of Microfinance" (2013) paper, Banerjee et al surveyed households in Karnataka rural villages in India. The first wave of the survey happened in 2006, and the data were published in 2011. Microfinance organization BSS entered 43 villages, but only for 29 of them the data on caste were collected on the household level. The second wave survey was conducted in 2012, and the data were published alongside the 2023 paper, "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets," by Banerjee et al. The second wave collected missing caste data for the remaining 14 villages, but some households left the area. As a result, the latter 14 villages have from 4% to 23% of original households missing caste information. Notably, households tend to cluster by caste.

Data

43 villages with household-level data, including household characteristics such as roof type, household size, age of each household member, number of females and males, number of rooms and beds, electricity and latrine ownership. Adjacency matrices for each village that allow for calculating degree centrality, local clustering coefficient, between centrality, etc. Five possible castes in the data: OBC, General, Scheduled Caste, Scheduled Tribe, and Muslim.

As a result, 9598 households with 258 households missing caste information.

Model and result

A two-layer GraphSAGE model with mean aggregation, ReLU activations, dropout regularization, and a log-softmax output for node classification. Trained for 25k epochs with $3 \cdot 10^5$ learning rate with Adam optimizer and negative log-likelihood loss.

The model shows slight overfitting with 83.5% accuracy for the training set and 74.5-76% accuracy for the validation set. The results are below:

Final Model Evaluation on Validation Set

	precision	recall	f1-score	support
OBC	0.74	0.84	0.79	688
General	0.66	0.58	0.62	168
Scheduled Caste	0.76	0.74	0.75	366
Scheduled Tribe	0.50	0.11	0.18	75
Muslims	0.97	0.98	0.98	62