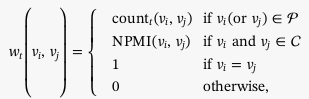
Reading summary for week of Sept 28th

This paper proposes a novel learning architecture for EHR graphs that takes unstructured data (ex. clinical notes in natural language) as well as structured data (diagnostic codes that make up most information in older GNNs for EHR tasks). This is motivated by the fact that different medical entities (patients, events, concepts) have connections that affect the features of the data. A key motivation for this work is the amount of incorrectly recorded codes present in most EHRs.

In short, HORDE embeds this heterogeneous graph into a low-dimensional latent space where similar concepts, diagnosis codes and visits are near one another. It does so by first building a graph out of the EHR data with multiple node classes (patient, event, concept), in which patient nodes vary over time to represent sequences of visits.

To weight edges between nodes, the algorithm accounts for co-occurrence of variables - for instance, concept edges are weighted based on conditional probability of one being mentioned within a sliding window in a patient' clinical notes given a mention of the other:



HORDE uses a graph convolutional network architecture to embed the nodes into a latent space. However, since nodes have different classes, they don't have feature vectors. The GCN takes the last output representation of each node as input and outputs a new embedding based on the nonlinear activation of a weighted sum of the last embeddings of all the node's neighbors. To account for the fact that some nodes vary over time, the full sequence of embeddings of those nodes are passed through an LSTM to produce the final embedding.

To calculate loss for each node, HORDE uses a random walk to compare embedding and graph proximity between each reachable node. Minimizing the loss function maximizes the similarity of reachable nodes.