580.694: HW #2

Max Collard

Clustering vertices

Write down a statistical decision theoretic framework for evaluating graph vertex binary clustering methods.

.

Sample space Each sample is a graph with labeled vertices; hence, our sample space is the space of ordered pairs of graphs with n vertices and binary labellings of n things:

$$\mathcal{S} = \mathcal{G}_n \times \{0,1\}^n$$

(Note that these true labellings are "latent", *i.e.*, not known to the decision rule.)

Model class For simplicity, we will only consider graphs distributed according to a sotchastic blockmodel; *i.e.*, our model class $\mathcal{M}$ will be the set of all

$$(G, L) \sim \text{SBM}_n^2(\boldsymbol{\rho}, P)$$

with 2 blocks, $\boldsymbol{\rho} \in \Delta_2$ the block membership probabilities, and $P \in [0,1]^{2\times 2}$ the interblock connection probability matrix.

Action space Our goal is to assign a binary label to each vertex; hence, our action space is

$$\mathcal{A} = \{0,1\}^n$$

Decision rule class In the most general sense, we could simply let our decision rule class be the space of all functions that take in graphs and produce labellings, *i.e.*,

$$\Phi = \mathcal{A}^{\mathcal{G}_n}$$

One could also make a simplification, and let Φ be the set of all decision rules produced by a particular algorithm (for example, 2-means), noting that all of the possible hyperparameter choices must be included in this space.

Loss One decent choice of loss functional, $L : \mathcal{A} \times \mathcal{A} \to [0, \infty)$, is the adjusted Rand index (ARI) between the true labelling A^* and guessed labelling $\hat{A}$:

$$L(A^*, \hat{A}) = \text{ARI}(A^*, \hat{A})$$

Risk Consider (G, A^*) jointly distributed according to the model class described above. Then, our end goal will be to select a decision rule $F \in \Phi$ that minimizes the risk

$$R[F] = E[L(A^*, F(G))]$$

where the expectation is taken over the entire joint distribution.

Hence, a "valid rule" $F : \mathcal{G}_n \to \mathcal{A}$ is a function that tells us, from an observed sample of an unlabeled graph G, a "guess" of the latent vertex labellings, $\hat{A} = F(G)$. We obtain the *optimal* rule F^* by taking

$$F^* = \arg\min_{F \in \Phi} R[F]$$

Our goal, in this decision-theoretic framework, would be to rigorously obtain this optimal F^*.