Chapter 7

Clustering

7.1 Decompositions and multicuts

This section is concerned with learning and inferring decompositions (clusterings) of a graph. We introduce some terminology of Horňáková et al. (2017):

Definition 16 Let $G = (A, E)$ be any graph. A subgraph $G' = (A', E')$ of G is called a *component* of G iff G' is non-empty, node-induced[1] and connected[2]. A partition Π of the node set A is called a *decomposition* of G iff, for every $U \in \Pi$, the subgraph $(U, E \cap \binom{U}{2})$ of G induced by U is connected (and thus a component of G).

For any graph G, we denote by D_G the set of all decompositions of G. Useful in the study of decompositions are the multicuts of a graph:

Definition 17 For any graph $G = (A, E)$, a subset $M \subseteq E$ of edges is called a *multicut* of G iff, for every cycle $C \subseteq E$ of G, we have $|C \cap M| \neq 1$.

For any graph G, we denote by M_G the set of all multicuts of G. For any decomposition of a graph G, the set of those edges that straddle distinct components is a multicut of G. This multicut is said to be induced by the decomposition. In fact, the map from decompositions to induced multicuts is a bijection from D_G to M_G (Horňáková et al., 2017, Lemma 2). This bijection allows us to state the problem of learning and inferring decompositions as one of learning and inferring multicuts.

The characteristic function $y\colon E \to \{0, 1\}$ of a multicut $y^{-1}(1)$ decides, for every edge $\{a, a'\} = e \in E$, whether the incident nodes belong to the same component ($y_e = 0$) or distinct components ($y_e = 1$). By the definition of a multicut, these decisions are not necessarily independent. More specifically:

Lemma 12 *For any graph $G = (V, E)$ and any $y\colon E \to \{0, 1\}$, the set $y^{-1}(1)$ of those edges that are mapped to 1 is a multicut of G iff the following inequalities are satisfied:*

$$\forall C \in \text{cycles}(G)\ \forall e \in C\colon \quad y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'} \tag{7.1}$$

Exercise 7 *a) Prove Lemma 12.*

b) Show that it is sufficient in (7.1) to consider only chordless cycles.

[1] I.e. $E' = E \cap \binom{A'}{2}$

[2] A component is not necessarily maximal w.r.t. the subgraph relation.

Now that we have a finite set E, decisions $y\colon E \to \{0,1\}$ and constraints (7.1), we can state the problem of learning and inferring multicuts as a learning and inference problem (4.1) with

$$S = E \tag{7.2}$$

$$\mathcal{Y} = \left\{ y\colon S \to \{0,1\} \;\middle|\; \forall C \in \text{cycles}(G)\ \forall e \in C\colon y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'} \right\} \tag{7.3}$$

7.2 Linear functions

7.2.1 Data

Throughout Section 7.2, we consider some graph $G = (A, E)$ and constrained data $(S, X, x, \mathcal{Y})$ with $S = E$, as in (7.2), $\mathcal{Y}$ defined as in (7.3), and $X = \mathbb{R}^V$ with some finite, non-empty set V. As a special case, we consider labeled data, i.e., $\mathcal{Y} = \{y\}$ with y satisfying the constraints (7.1).

7.2.2 Familiy of functions

Throughout Section 7.2, we consider linear functions. More specifically, we consider $\Theta = \mathbb{R}^V$ and $f : \Theta \to \mathbb{R}^X$ such that

$$\forall \theta \in \Theta\ \forall \hat{x} \in \mathbb{R}^V : \quad f_\theta(\hat{x}) = \langle \theta, \hat{x} \rangle \ . \tag{7.4}$$

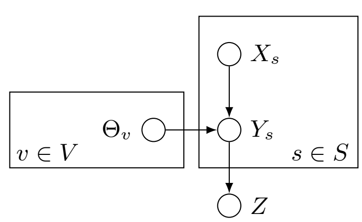
7.2.3 Probabilistic model

Random variables

- For any $\{a, a'\} \in S$, let $X_{\{a,a'\}}$ be a random variable whose realization is a vector $x_{\{a,a'\}} \in \mathbb{R}^V$, called the *attribute vector* of the pair $\{a, a'\}$.
- For any $\{a, a'\} \in S$, let $Y_{\{a,a'\}}$ be a random variable whose realization is a binary number $y_{\{a,a'\}} \in \{0,1\}$, called the *decision* of assigning a and a' to distinct components
- For any $v \in V$, let Θ_v be a random variable whose realization is a real number $\theta_v \in \mathbb{R}$, called a *parameter*
- Let Z be a random variable whose realization is a subset $z \subseteq \{0,1\}^S$. We are interested in $z = \mathcal{Y}$, a characterization of all multicuts (and hence, decompositions) of G

Conditional independence assumptions

We assume a probability distribution that factorizes according to the Bayesian net depicted below.



Factorization

These conditional independence assumptions imply the following factorizations:

- Firstly:

$$P(X, Y, Z, \Theta) = P(Z \mid Y) \prod_{s \in S} P(Y_s \mid X_s, \Theta) \prod_{s \in S} P(X_s) \prod_{v \in V} P(\Theta_v) \tag{7.5}$$

- Secondly:

$$\begin{aligned}
P(\Theta \mid X, Y, Z) &= \frac{P(X,Y,Z,\Theta)}{P(X,Y,Z)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid X, Y)\, P(X,Y)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid Y)\, P(X,Y)} \\
&= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X,Y)} \\
&\propto P(Y \mid X, \Theta)\, P(\Theta) \\
&= \prod_{s \in S} P(Y_s \mid X_s, \Theta) \prod_{v \in V} P(\Theta_v) \qquad (7.6)
\end{aligned}$$

- Thirdly,

$$\begin{aligned}
P(Y \mid X, Z, \theta) &= \frac{P(X,Y,Z,\Theta)}{P(X,Z,\Theta)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X,Z,\Theta)} \\
&\propto P(Z \mid Y)\, P(Y \mid X, \Theta) \\
&= P(Z \mid Y) \prod_{s \in S} P(Y_s \mid X_s, \Theta) \qquad (7.7)
\end{aligned}$$

Forms

Here, we consider:

- The *logistic distribution*

$$\forall s \in S: \qquad p_{Y_s \mid X_s, \Theta}(1) = \frac{1}{1 + 2^{-f_\theta(x_s)}} \qquad (7.8)$$

- A $\sigma \in \mathbb{R}^+$ and the *normal distribution*:

$$\forall v \in V: \qquad p_{\Theta_v}(\theta_v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_v^2 / 2\sigma^2} \qquad (7.9)$$

- A uniform distribution on a subset:

$$\forall z \subseteq \{0,1\}^S: \quad p_{Z \mid Y}(z) \propto \begin{cases} 1 & \text{if } y \in z \\ 0 & \text{otherwise} \end{cases} \qquad (7.10)$$

  Note that $p_{Z|Y}(\mathcal{Y})$ is non-zero iff $y^{-1}(1)$ is a multicut and hence defines a decomposition of G.

7.2.4 Learning problem

Corollary 1 *Estimating maximally probable parameters θ, given attributes x and labels y, i.e.,*

$$\underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta \mid X, Y}(\theta, x, y)$$

is identical to the supervised learning problem w.r.t. L, R and λ such that

$$\forall r \in \mathbb{R}\ \forall \hat{y} \in \{0,1\}: \quad L(r, \hat{y}) = -\hat{y} r + \log\left(1 + 2^r\right) \qquad (7.11)$$

$$\forall \theta \in \Theta: \qquad R(\theta) = \|\theta\|_2^2 \qquad (7.12)$$

$$\lambda = \frac{\log e}{2\sigma^2} \qquad (7.13)$$

7.2.5 Inference problem

Corollary 2 *For any constrained data as defined above and any $\theta \in \mathbb{R}^V$, the inference problem has the form of* CORRELATION-CLUSTERING, *i.e.*

$$\min_{y\colon S \to \{0,1\}} \quad \sum_{\{a,a'\} \in S} (-\langle \theta, x_{\{a,a'\}} \rangle)\, y_{\{a,a'\}} \tag{7.14}$$

$$\text{subject to} \quad \forall C \in \text{cycles}(G)\ \forall e \in C\colon \quad y_e \leq \sum_{e' \in C \setminus \{e\}} y_{e'} \ . \tag{7.15}$$

CORRELATION-CLUSTERING has been studied intensively, notably by Chopra and Rao (1993), Bansal et al. (2004) and Demaine et al. (2006).

Lemma 13 (Bansal et al. (2004)) CORRELATION-CLUSTERING *is* NP*-hard.*

Bansal et al. (2004) establish NP-hardness of CORRELATION-CLUSTERING by a reduction of k-TERMINAL-CUT whose NP-hardness is an important result of Dahlhaus et al. (1994).