Markov blanket discovery in positive-unlabelled and semi-supervised data: Supplementary Material

Konstantinos Sechidis and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.sechidis,gavin.brown}@manchester.ac.uk

Proofs of Section 4

Before proving the theorems of this section, we will prove the following useful Lemma.

Lemma 1.
In the positive unlabelled scenario, under the selected completely at random assumption, for any subset of features $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$ *it holds:*

$$p(x|y^+, \mathbf{z}) = p(x|s_P^+, \mathbf{z}) \quad \forall \ \ \mathbf{z} \in \boldsymbol{\mathcal{Z}}.$$

Proof. In order to prove this Lemma we will start from the *rhs* of the desired equation:

$$p(x|s_P^+, \mathbf{z}) \overset{\text{when } S=s_P^+ \text{ then } Y=y^+}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=} p(x|s_P^+, y^+, \mathbf{z})$$

This hold because in PU setting if an instance is labelled then it must be positive. Then by using the Bayes theorem and the chain rule we get:

$$p(x|s_P^+, y^+, \mathbf{z}) \overset{\text{Bayes theorem}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=} \frac{p(x, s_P^+|y^+, \mathbf{z})}{p(s_P^+|y^+, \mathbf{z})} \overset{\text{Chain rule}}{=\!=\!=\!=\!=\!=\!=\!=\!=} \frac{p(s_P^+|x, y^+, \mathbf{z})p(x|y^+, \mathbf{z})}{p(s_P^+|y^+, \mathbf{z})}$$

Because of the selected completely at random assumption:

$$p(s_P^+|y^+, x, \mathbf{z}) = p(s_P^+|y^+, \mathbf{z}) \tag{1}$$

As a result the last expression becomes:

$$\frac{p(s_P^+|x, y^+, \mathbf{z})p(x|y^+, \mathbf{z})}{p(s_P^+|y^+, \mathbf{z})} \overset{\text{eq. (1)}}{=\!=\!=\!=\!=} p(x|y^+, \mathbf{z}) \tag{2}$$

This finishes the proof, since we derived the *lhs* of the desired equation.

An interesting point to clarify is that equation (1) holds for any subset of features. To show that, without loss of generality, let us assume that the entire set of features $\mathbf{x}$ consists of the variables $x, \mathbf{z}$ and $\mathbf{w}$, where x is a single variable and $\mathbf{z}$, $\mathbf{w}$ sets of variables. The $x, \mathbf{z}$ and $\mathbf{w}$ can be created by any feature combination

as long their intersection is the empty set and their union is the entire feature space. Now we can re-write the selected completely at random assumption [2, eq. (3)]

$$\begin{aligned} p(s_P^+|y^+,\mathbf{x}) = p(s_P^+|y^+) \Leftrightarrow \\ p(s_P^+|y^+,x,\mathbf{z},\mathbf{w}) = p(s_P^+|y^+) \Leftrightarrow \\ p(s_P^+,x,\mathbf{z},\mathbf{w}|y^+) = p(s_P^+|y^+)p(x,\mathbf{z},\mathbf{w}|y^+). \end{aligned} \tag{3}$$

Now marginalising out the variable $\mathbf{w}$ we get:

$$\begin{aligned} \sum_{\mathbf{w}\in\mathcal{W}} p(s_P^+,x,\mathbf{z},\mathbf{w}|y^+) = p(s_P^+|y^+) \sum_{\mathbf{w}\in\mathcal{W}} p(x,\mathbf{z},\mathbf{w}|y^+) \Leftrightarrow \\ p(s_P^+,x,\mathbf{z}|y^+) = p(s_P^+|y^+)p(x,\mathbf{z}|y^+) \Leftrightarrow \end{aligned} \tag{4}$$

$$p(s_P^+|y^+,x,\mathbf{z}) = p(s_P^+|y^+) \tag{5}$$

Furthermore in equation (4) by marginalising out the variable x we get:

$$\begin{aligned} \sum_{x\in\mathcal{X}} p(s_P^+,x,\mathbf{z}|y^+) = p(s_P^+|y^+) \sum_{x\in\mathcal{X}} p(x,\mathbf{z}|y^+) \Leftrightarrow \\ p(s_P^+,\mathbf{z}|y^+) = p(s_P^+|y^+)p(\mathbf{z}|y^+) \Leftrightarrow \\ p(s_P^+|y^+,\mathbf{z}) = p(s_P^+|y^+) \end{aligned} \tag{6}$$

Thus from equations (5) and (6) we can derive equation (1). □

1 Proof of Theorem 1

Theorem 1 (Testing conditional independence in PU data). *In the positive unlabelled scenario, under the selected completely at random assumption, a variable X is independent of the class label Y given a subset of features $\mathbf{Z}$ if and only if X is independent of S_P given $\mathbf{Z}$, so it holds:*

$$X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp S_P|\mathbf{Z}.$$

Proof. To prove $X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp S_P|\mathbf{Z}$ we need to prove that

$$p(x,s_P|\mathbf{z}) = p(x|\mathbf{z})p(s_P|\mathbf{z}) \Leftrightarrow p(x,y|\mathbf{z}) = p(x|\mathbf{z})p(y|\mathbf{z}) \forall x \in \mathcal{X}, y \in \mathcal{Y}, s_P \in \mathcal{S}_\mathcal{P} \text{ and } \mathbf{z} \in \boldsymbol{\mathcal{Z}}$$

Since the random variables S_P and Y are binary it is sufficient to prove this for the two classes. For the first class we have:

$$\begin{aligned} p(x,s_P^+|\mathbf{z}) = p(x|\mathbf{z})p(s_P^+|\mathbf{z}) \Leftrightarrow p(x|s_P^+,\mathbf{z}) = p(x|\mathbf{z}) \overset{\text{Lemma 1}}{\Leftrightarrow} \\ p(x|s_P^+,\mathbf{z}) = p(x|\mathbf{z}) \Leftrightarrow p(x,y^+|\mathbf{z}) = p(x|\mathbf{z})p(y^+|\mathbf{z}) \end{aligned}$$

Using the above result for the first class, we will prove it for the second:

$$\begin{aligned} p(x,s_P^-|\mathbf{z}) = p(x|\mathbf{z})p(s_P^-|\mathbf{z}) \Leftrightarrow p(x|\mathbf{z}) - p(x,s_P^+|\mathbf{z}) = p(x|\mathbf{z})(1-p(s_P^+|\mathbf{z})) \Leftrightarrow \\ p(x,s_P^+|\mathbf{z}) = p(x|\mathbf{z})p(s_P^+|\mathbf{z}) \Leftrightarrow p(x,y^+|\mathbf{z}) = p(x|\mathbf{z})p(y^+|\mathbf{z}) \Leftrightarrow \\ p(x|\mathbf{z}) - p(x,y^-|\mathbf{z}) = p(x|\mathbf{z})(1-p(y^-|\mathbf{z})) \Leftrightarrow p(x,y^-|\mathbf{z}) = p(x|\mathbf{z})p(y^-|\mathbf{z}) \end{aligned}$$

□

2 Proof of Theorem 2

Theorem 2 (Power of PU conditional test of independence). *In the positive unlabelled scenario, under the selected completely at random assumption, when a variable X is dependent on the class label Y given a subset of features $\mathbf{Z}$, $X \not\perp\!\!\!\perp Y|\mathbf{Z}$, we have: $I(X;Y|\mathbf{Z}) > I(X;S_P|\mathbf{Z})$.*

Proof. The conditional mutual information between variables X, Y given $\mathbf{Z}$ (for binary Y) is:

$$I(X;Y|\mathbf{Z}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x,y,\mathbf{z}) \ln \frac{p(x,y|\mathbf{z})}{p(x|\mathbf{z})p(y|\mathbf{z})} =$$

$$= \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x,y^+,\mathbf{z}) \ln \frac{p(x,y^+|\mathbf{z})}{p(x|\mathbf{z})p(y^+|\mathbf{z})} + \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x,y^-,\mathbf{z}) \ln \frac{p(x,y^-|\mathbf{z})}{p(x|\mathbf{z})p(y^-|\mathbf{z})}.$$

Using the fact that: $p(x,y^-|\mathbf{z}) = p(x|\mathbf{z}) - p(x,y^+|\mathbf{z}) = p(x|\mathbf{z}) - p(x|y^+,\mathbf{z})p(y^+|\mathbf{z})$ we can re-write the conditional mutual information as:

$$I(X;Y|\mathbf{Z}) = \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x|y^+,\mathbf{z})p(y^+|\mathbf{z}) \ln \frac{p(x|y^+,\mathbf{z})}{p(x|\mathbf{z})}$$
$$+ \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} (p(x|\mathbf{z}) - p(x|y^+,\mathbf{z})p(y^+|\mathbf{z})) \ln \frac{p(x|\mathbf{z}) - p(x|y^+,\mathbf{z})p(y^+|\mathbf{z})}{p(x|\mathbf{z})(1-p(y^+|\mathbf{z}))}.$$

In order to explore the relationship between $I(X;Y|\mathbf{Z})$ and $I(X;S_P|\mathbf{Z})$, we introduce the following function:

$$f(\tilde{p}) = \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x|s_P^+,\mathbf{z})\tilde{p} \ln \frac{p(x|s_P^+,\mathbf{z})}{p(x|\mathbf{z})}$$
$$+ \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} (p(x|\mathbf{z}) - p(x|s_P^+,\mathbf{z})\tilde{p}) \ln \frac{p(x|\mathbf{z}) - p(x|s_P^+,\mathbf{z})\tilde{p}}{p(x|\mathbf{z})(1-\tilde{p})}. \tag{7}$$

When $\tilde{p} = p(y^+|\mathbf{z})$, and using Lemma 1, this is exactly $I(X;Y|\mathbf{Z})$. Alternatively, when $\tilde{p} = p(s_P^+|\mathbf{z})$, this is exactly $I(X;S_P|\mathbf{Z})$. So, in order to explore the relationship between $I(X;Y|\mathbf{Z})$ and $I(X;S_P|\mathbf{Z})$, we should explore the monotonicity of the function $f(\tilde{p})$ with the following Lemma.

Lemma 2. *Under the selected completely at random assumption f is a non-decreasing function of $\tilde{p}$, and it is strictly increasing when $X \not\perp\!\!\!\perp Y|\mathbf{Z}$.*

Proof (of Lemma 2). By taking the first derivative of f with respect to $\tilde{p}$ we have

$$\frac{d}{d\tilde{p}} f(\tilde{p}) = -\sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x|s_P^+,\mathbf{z}) \ln \frac{\frac{p(x|\mathbf{z})}{p(x|s_P^+,\mathbf{z})} - \tilde{p}}{1-\tilde{p}}.$$

Applying Jensen's inequality to the strictly convex function $-ln(\cdot)$, we get

$$\frac{d}{d\tilde{p}}f(\tilde{p}) \geq -\ln \sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}} p(x|s_P^+,\mathbf{z})\frac{\frac{p(x|\mathbf{z})}{p(x|s_P^+,\mathbf{z})}-\tilde{p}}{1-\tilde{p}}$$
$$= -\ln\left(\frac{1}{1-\tilde{p}}\sum_{x\in\mathcal{X}}\sum_{\mathbf{z}\in\boldsymbol{\mathcal{Z}}}\big(p(x|\mathbf{z})-\tilde{p}p(x|s_P^+,\mathbf{z})\big)\right) = -\ln\frac{1-\tilde{p}}{1-\tilde{p}} = 0.$$

So $f(\tilde{p})$ is a non-decreasing function of $\tilde{p}$. Furthermore we will have equality if and only if $\frac{\frac{p(x)}{p(x|s_P^+,\mathbf{z})}-\tilde{p}}{1-\tilde{p}}$ is constant for all $x\in\mathcal{X}$ and $\mathbf{z}\in\boldsymbol{\mathcal{Z}}$. This implies that we will have $\frac{d}{d\tilde{p}}f(\tilde{p})=0$ if and only if $p(x|\mathbf{z})=p(x|s_P^+,\mathbf{z})\ \ \forall\ \ x\in\mathcal{X},\ \mathbf{z}\in\boldsymbol{\mathcal{Z}}$, or in other words when $X\perp\!\!\!\perp Y|\mathbf{Z}$. So when $X\not\perp\!\!\!\perp Y|\mathbf{Z}$, f is strictly increasing function of $\tilde{p}$. This finishes the proof of Lemma 2. □

Given Lemma 2, and combining it with the fact that in PU data $p(y^+|\mathbf{z}) > p(s_P^+|\mathbf{z})$, when $X\not\perp\!\!\!\perp Y|\mathbf{Z}$ we get

$$f(p(y^+|\mathbf{z})) > f(p(s_P^+|\mathbf{z})) \Leftrightarrow I(X;Y|\mathbf{Z}) > I(X;S_P|\mathbf{Z})$$

and this two quantities are equal only when $X\perp\!\!\!\perp Y|\mathbf{Z}$. This finishes the proof of Theorem 2. □

3 Proof of Theorem 3

Theorem 3 (Correction factor for PU test).
The non-centrality parameter of the conditional G-test between X and S_P given a subset of features $\mathbf{Z}$ takes the form:

$$\lambda_{G(X;S_P|\mathbf{Z})} = \kappa_P\lambda_{G(X;Y|\mathbf{Z})} = \kappa_P 2NI(X;Y|\mathbf{Z}),$$

where $\kappa_P = \frac{1-p(y^+)}{p(y^+)}\frac{p(s_P^+)}{1-p(s_P^+)} = \frac{1-p(y^+)}{p(y^+)}\frac{N_{S_P^+}}{N-N_{S_P^+}}$.

Proof. By using the chain rule of the mutual information [1] the non-centrality parameter can be written as:

$$\lambda_{G(X;S_P|\mathbf{Z})} = 2NI(X;S_P|\mathbf{Z}) = 2NI(X\mathbf{Z};S_P) - 2NI(\mathbf{Z};S_P) = \lambda_{G(X\mathbf{Z};S_P)} - \lambda_{G(\mathbf{Z};S_P)}.$$

Using Theorem 3 from [3], we can associate the non-centrality parameters of the G-tests X, S_P and X, Y, so we have:

$$\lambda_{G(X;S_P|\mathbf{Z})} = \kappa_P\lambda_{G(X\mathbf{Z};Y)} - \kappa_P\lambda_{G(\mathbf{Z};Y)} =$$
$$= \kappa_P 2NI(X\mathbf{Z};Y) - \kappa_P 2NI(\mathbf{Z};Y) = \kappa_P 2N(I(X\mathbf{Z};Y) - I(\mathbf{Z};Y)).$$

And, by using again the chain rule, the last expression can be written as:

$$\lambda_{G(X;S_P|\mathbf{Z})} = \kappa_P 2NI(X;Y|\mathbf{Z}) = \kappa_P\lambda_{G(X;Y|\mathbf{Z})}.$$

□

References

1. Cover, T.M., Thomas, J.A.: Elements of information theory. J. Wiley & Sons (2006)
2. Sechidis, K., Brown, G.: Markov blanket discovery in positive-unlabelled and semi-supervised data. In: Machine Learning and Knowledge Discovery in Databases (ECML/PKDD). Springer Berlin Heidelberg (2015)
3. Sechidis, K., Calvo, B., Brown, G.: Statistical hypothesis testing in positive unlabelled data. In: Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pp. 66–81. Springer Berlin Heidelberg (2014)