Machine Learning 2 - Homework 2

Sindy Löwe (11594969)

April 15, 2018

Collaborators: Pascal Esser

Problem 1

1.

$$I(X;Y) = \mathcal{KL}(p(x,y) \mid\mid p(x) \cdot p(y))$$
$$I(X;Y|Z) = \mathbb{E}[\mathcal{KL}(p(x,y|z) \mid\mid p(x|z) \cdot p(y|z))]$$

where:

$$p(x,y|z) = \begin{cases} \frac{p(x,y,z)}{p(z)} & \text{if } p(z) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

The mutual information revolves around the information that is shared between two variables, i.e. the intersection of their information. We can view it as the reduction in uncertainty about x by virtue of being told y (and vice versa). The conditional mutual information again measures the intersection between the information of x and y, but is conditioned on the information in z, which describes the context of prior information.

2. For calculating $I(X;Y)$ from discrete values, we use the equation:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y)\ log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

The resulting values for $p(x), p(y), p(x,y)$ and $log\left(\frac{p(x,y)}{p(x)p(y)}\right)$ are:

x	y	p(x)	p(y)	p(x,y)	$\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$
0	0	0.6	0.592	0.336	- 0.0556
0	1		0.408	0.264	0.0755
1	0	0.4		0.256	0.0780
1	1			0.144	- 0.1252

Therefore:

$$\begin{aligned} I(X;Y) &= 0.336 \cdot (-0.0556) + 0.264 \cdot 0.0755 + 0.256 \cdot 0.0780 + 0.144 \cdot (-0.1252) \\ &= 0.0032 \end{aligned}$$

Since $I(X;Y) = 0.0032 > 0$, we can conclude that X and Y are dependent.

3. For calculating $I(X;Y|Z)$ from discrete values, we use the equation:

$$I(X;Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p(z)\ p(x,y|z)\ log \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right)$$

The resulting values for $p(z), p(x,y|z), p(x|z), p(y|z)$ and $log \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right)$ are:

x	y	z	p(z)	$p(x,y\|z)$	$p(x\|z)$	$p(y\|z)$	$log \left(\frac{p(x,y\|z)}{p(x\|z)p(y\|z)} \right)$
0	0	0	0.480	0.4	0.5	0.8	0
0	0	1	0.520	0.2770	0.6923	0.4	0
0	1	0		0.1		0.2	0
0	1	1		0.4153		0.6	0
1	0	0		0.4	0.5		0
1	0	1		0.1231	0.3077		0
1	1	0		0.1			0
1	1	1		0.1846			0

Therefore:

$$I(X;Y|Z) = 0$$

Since $I(X;Y|Z) = 0$, we can conclude that X and Y are independent, given Z.

4. In general:

$$p(x,y,z) = p(x) \cdot p(z|x) \cdot p(y|x,z)$$

But since we showed in subtask 3 that X and Y are independent given Z, we can set $p(y|x,z) = p(y|z)$ (see Bishop 8.20 and 8.22). Therefore, we can simplify this equation and get

$$p(x,y,z) = p(x) \cdot p(z|x) \cdot p(y|z)$$

This equation can be depicted by the graph shown in fig. 1, which describes the chain case.

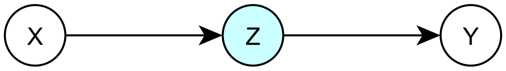

Figure 1: The corresponding directed graph depicting $p(x,y,z) = p(x) \cdot p(z|x) \cdot p(y|z)$

Problem 2

Considering the dependency relation between X and Y given Z, we can create four clusters. The first cluster, that describes the situation where X and Y are independent from one another, no matter whether Z is given or not, contains all Bayesian Networks, in which there is simply no connection between the variables X and Y.

The more interesting cases arise when there is no direct connection between X and Y, but a connection between X and Z and Y and Z. Here, the fork and chain case both give rise to a situation in which X is dependent on Y, but independent on it when Z is given. The collider case describes the exact opposite, where X and Y are independent, but become dependent when Z is given.

Lastly, whenever X and Y share a direct connection, they are dependent on one another, not matter whether Z is given or not.

The resulting clusters are depicted in fig. 2.

Problem 3

1.

$$
\begin{aligned}
\mathcal{KL}(p||q) &= -\int p(\mathbf{x}) \log \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \\
&= \int p(\mathbf{x}) \log(q(\mathbf{x}) - p(\mathbf{x})) d\mathbf{x} \\
&= \int p(\mathbf{x}) [-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&+ \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\mathbf{L}|) + \frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})] d\mathbf{x} \\
&= \frac{1}{2}\log\left(\frac{|\mathbf{L}|}{|\mathbf{\Sigma}|}\right) \int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{x}) [-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&+ \frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})] d\mathbf{x}
\end{aligned}
$$

We use $\int p(\mathbf{x}) d\mathbf{x} = 1$ and the law of the unconscious statistician to get

$$
= \frac{1}{2}\log\left(\frac{|\mathbf{L}|}{|\mathbf{\Sigma}|}\right) - \frac{1}{2}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] + \frac{1}{2}\mathbb{E}[(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})]
$$

Since p is considered to be the original distribution, it follows that $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$. Therefore, we can use equation 380 from the matrix cookbook to get:

$$
\begin{aligned}
&= \frac{1}{2}\log\left(\frac{|\mathbf{L}|}{|\mathbf{\Sigma}|}\right) - \frac{1}{2}[(\boldsymbol{\mu}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) + Tr(\mathbf{\Sigma}^{-1}\mathbf{\Sigma})] \\
&+ \frac{1}{2}[(\boldsymbol{\mu}-\mathbf{m})^T \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\mathbf{\Sigma})] \\
&= \frac{1}{2}\left[\log\left(\frac{|\mathbf{L}|}{|\mathbf{\Sigma}|}\right) - Tr(I) + (\boldsymbol{\mu}-\mathbf{m})^T \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\mathbf{\Sigma})\right] \\
&= \frac{1}{2}\left[\log\left(\frac{|\mathbf{L}|}{|\mathbf{\Sigma}|}\right) - D + (\boldsymbol{\mu}-\mathbf{m})^T \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\mathbf{\Sigma})\right]
\end{aligned}
$$

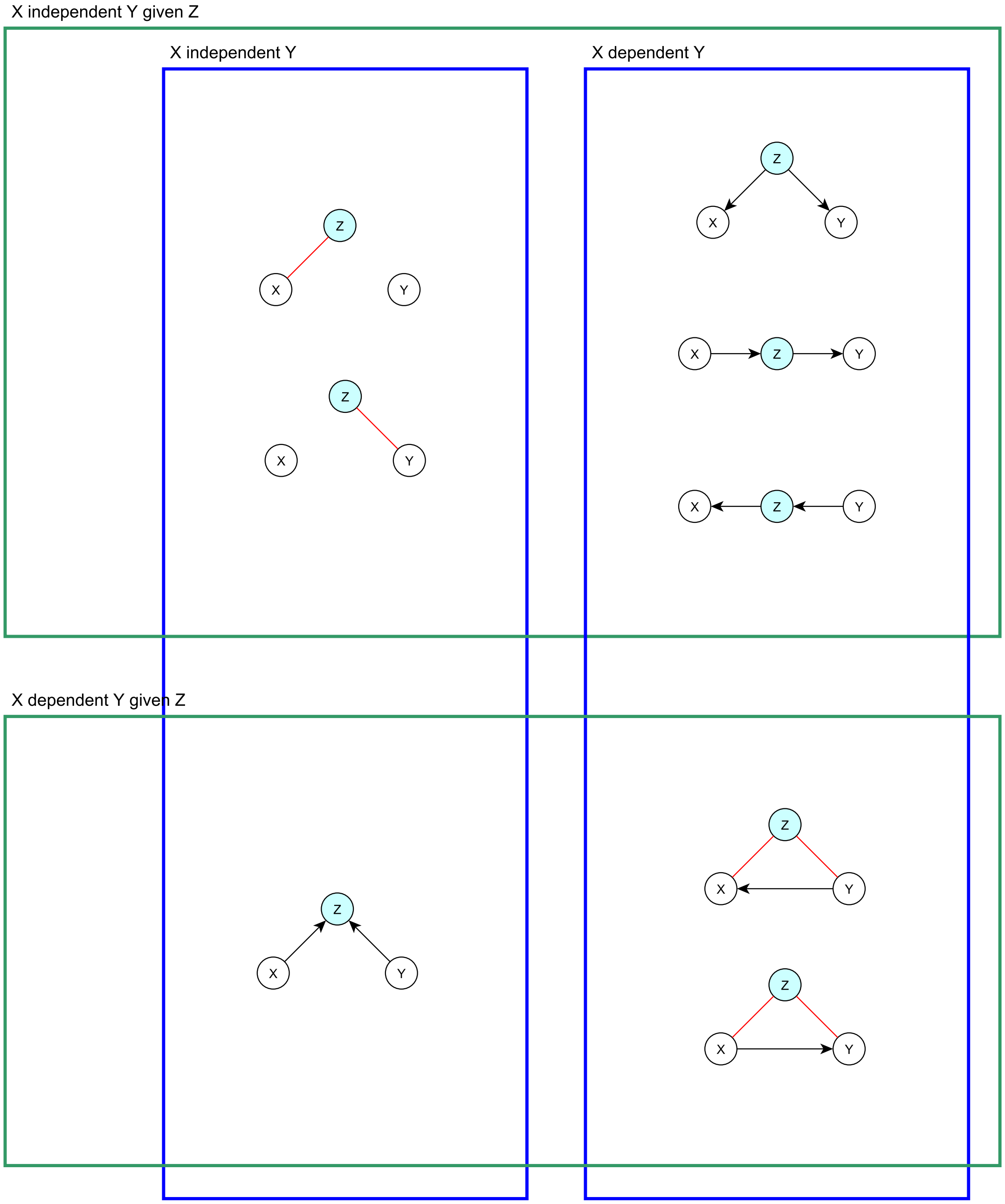

Figure 2: Depicting the independence relations between all Bayesian networks consisting of three vertices X, Y and Z. Arrows indicate dependencies between variables. Red edges indicate that any kind of relationship could be present or absent between the variables (as long as the graph stays acyclic) without changing the considered dependency relation. Without loss of generality, we consider the dependency X on Y given Z. All permutations of this follow accordingly.

2.

$$\begin{aligned}
\mathcal{H}(p) &= -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\
&= -\int p(\mathbf{x}) \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} \\
&= -\int p(\mathbf{x}) d\mathbf{x} \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) \right] + \int p(\mathbf{x}) \left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x}
\end{aligned}$$

We use $\int p(\mathbf{x}) d\mathbf{x} = 1$ to get:

$$= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \int p(\mathbf{x}) \left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x}$$

Again, we can use the law of the unconscious statistician and make use of equation 380 from the matrix cookbook to get:

$$\begin{aligned}
&= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \frac{1}{2}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] \\
&= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \frac{1}{2}\left[(\boldsymbol{\mu}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) + Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) \right] \\
&= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \frac{D}{2}
\end{aligned}$$