Machine Learning 2 — Homework 2

Maurice Frank
11650656
maurice.frank@posteo.de

September 16, 2019

Problem 1.

1.

We have the discrete random variables X, Y, Z. Mutual information than is:

$$\begin{aligned} I(X;Y) &= \mathcal{KL}(p(x,y)||p(x)p(y)) \\ &= \mathcal{H}(X) - \mathcal{H}(X|Y) \end{aligned}$$

The conditional mutual information is:

$$\begin{aligned} I(X;Y|Z) &= \mathbb{E}_{p(z)}[\mathcal{KL}(p(x,y|z)||p(x|y)p(y|z))] \\ &= \mathcal{H}(X|Z) - \mathcal{H}(X|Y,Z) \end{aligned}$$

We see that the conditional mutual information measures the expected mutual information between X and Y given Z.

2.

We have $x, y, z \in \{0, 1\}$ with $p(x, y, z)$. First we will write down $p(x, y)$ (see Table 1), $p(x)$ and $p(y)$ (see Table 2).

X	Y	p(x,y)
0	0	0.336
0	1	0.264
1	0	0.256
1	1	0.144

Table 1: Marginalizing Z.

Y X	0	1	p(x)
0	0.336	0.264	0.6
1	0.256	0.144	0.4
p(y)	0.592	0.408	1

Table 2: Marginalizing X and Y.

y x	0	1	p(x | z=0)
0	0.4	0.1	0.5
1	0.4	0.1	0.5
p(y | z=0)	0.8	0.2	1.0

Table 3: p(x,y | z=0) and p(x | z=0), p(y | z=0)

$$\begin{aligned} I(X;Y) &= \mathcal{KL}(p(x,y)||p(x)p(y)) \\ &= -\sum_{x,y} p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right) \\ &= 3.197 \cdot 10^{-3} \\ &> 0 \end{aligned}$$

As the mutual information between X and Y is bigger than zero (it is symmetric!) we showed that having knowledge about one of the values we gain knowledge about the possible distribution of values of the second variable. We see this with $I(X;Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) > 0 \implies \mathcal{H}(X|Y) < \mathcal{H}(X)$, the entropy of one given the other is lower than without this conditional information.

3.

See Table 3 and Table 4 for intermediate calculation tables.

y x	0	1	p(x | z=1)
0	0.277	0.415	0.692
1	0.123	0.185	0.308
p(y | z=1)	0.4	0.6	1.0

Table 4: p(x,y | z=1) and p(x | z=1), p(y | z=1)

x	$p(x)$	z	$p(z\|x)$	y	$p(y\|z)$	$p(x) \cdot p(z\|x) \cdot p(y\|z)$
0	0.6	0	0.4	0	0.8	0.192
0	0.6	0	0.4	1	0.2	0.048
0	0.6	1	0.6	0	0.4	0.144
0	0.6	1	0.6	1	0.6	0.216
1	0.4	0	0.6	0	0.8	0.192
1	0.4	0	0.6	1	0.2	0.048
1	0.4	1	0.4	0	0.4	0.064
1	0.4	1	0.4	1	0.6	0.096

Table 5: Computation of $p(x) \cdot p(z|x) \cdot p(y|z)$

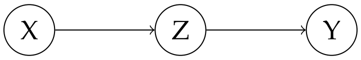


Figure 1: The directed graph to the factorization $p(x)p(z|x)p(y|z)$

$$\begin{aligned}
I(X;Y|Z) = \quad & p(z=0) \cdot \mathcal{KL}(p(x,y|z=0)||p(x|z=0)p(y|z=0)) \\
& + p(z=1) \cdot \mathcal{KL}(p(x,y|z=1)||p(x|z=1)p(y|z=1)) \\
& = -\sum_{z \in Z} p(Z=z) \cdot \sum_{x,y} p(x,y|z) \cdot \log\left(\frac{p(x|z)p(y|z)}{p(x,y|z)}\right) \\
& = -(0.48 \cdot 0 + 0.52 \cdot 0) \\
& = 0
\end{aligned}$$

That now the conditional mutual information is zero tells us that given that we know about the value of Z than having information about X or Y will not tell us anything about the respective third variable. $I(X;Y|Z) = 0 \implies \mathcal{H}(X|Z) = \mathcal{H}(X|Y,Z) \wedge \mathcal{H}(Y|Z) = \mathcal{H}(Y|X,Z)$.

4.

For the Computation of $p(x) \cdot p(z|x) \cdot p(y|z) = p(x,y,z)$ see Table 5. The directed graph is show in Figure 1.

Problem 2.

We strictly plot not the permutated directed graphs. See Figures 3 until 7 for the plotted clusters. Any not written conditional relationship is independent. As for the definition of permutation consider the graph in Figure 2. This is the same graph as in Figure 4 as we can permute the variable names: $X \to Y \wedge Z \to X \wedge Y \to Z$. This is a permutation and proves that those two clusters are the same.

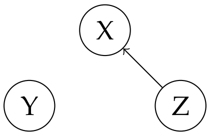
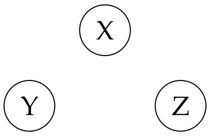
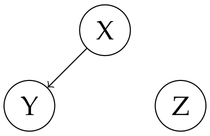


Figure 2: Example DAG



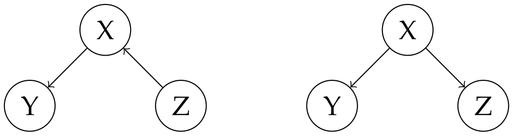
$$\boxed{X \perp\!\!\!\perp Y \wedge X \perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z}$$

Figure 3: Cluster I



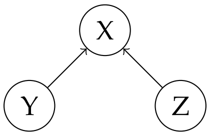
$$\boxed{X \not\!\perp\!\!\!\perp Y \wedge X \perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z}$$

Figure 4: Cluster II



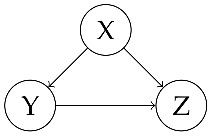
$$\boxed{X \not\!\perp\!\!\!\perp Z \wedge X \not\!\perp\!\!\!\perp Y \wedge Y \not\!\perp\!\!\!\perp Z|\varnothing \wedge Y \perp\!\!\!\perp Z|X}$$

Figure 5: Cluster III



$$\boxed{X \not\!\perp\!\!\!\perp Y \wedge X \not\!\perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z \wedge Y \not\!\perp\!\!\!\perp Z|X}$$

Figure 6: Cluster IV



$$\boxed{X \not\perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp Z \wedge Y \not\perp\!\!\!\perp Z}$$

Figure 7: Cluster V

Problem 3.

1.

We have given $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{L})$. $\boldsymbol{x} \in \mathbb{R}^k$. We are using the result of Section 2. in this computation.

$$\begin{aligned}
\ln q(\boldsymbol{x}) &= -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{L}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m}) \\
\mathcal{KL}(p||q) &= -\int p(\boldsymbol{x})\ln(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})})d\boldsymbol{x} \\
&= -\int p(\boldsymbol{x})\left[\ln q(\boldsymbol{x}) - \ln p(\boldsymbol{x})\right]d\boldsymbol{x} \\
&= \int p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x} - \int p(\boldsymbol{x})\ln q(\boldsymbol{x})d\boldsymbol{x} \\
&= -\mathcal{H}(p(\boldsymbol{X})) - \int p(\boldsymbol{x})\ln q(\boldsymbol{x})d\boldsymbol{x} \\
&= -\mathcal{H}(p(\boldsymbol{X})) + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{L}| + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})d\boldsymbol{x} \\
&= -\frac{k}{2} - \frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{L}| + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})d\boldsymbol{x} \\
&= -\frac{k}{2} + \frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})d\boldsymbol{x} \\
&= -\frac{k}{2} + \frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + \frac{1}{2}\mathbb{E}[(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})]_{p(\boldsymbol{x})} \qquad \text{(using (1))} \\
&= -\frac{k}{2} + \frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + \frac{1}{2}[(\boldsymbol{\mu}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + \operatorname{Tr}(\boldsymbol{L}^{-1}\boldsymbol{\Sigma})] \\
&= \frac{1}{2}\left[\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - k + (\boldsymbol{\mu}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + \operatorname{Tr}(\boldsymbol{L}^{-1}\boldsymbol{\Sigma})\right]
\end{aligned}$$

2.

$$
\begin{aligned}
\ln p(\boldsymbol{x}) &= -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
\mathcal{H}(p) &= -\int p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\left[(\boldsymbol{\mu}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) + \mathrm{Tr}\,(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma})\right] \\
&\qquad\qquad \text{(using (1))} \\
&= \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}\,(\mathbb{1}_k) \\
&= \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{k}{2}
\end{aligned}
$$