Philipp Lintl
12152498
philipp.lintl@student.uva.nl

Homework Assignment 2
Machine Learning 2, 19/20

2019-09-16

As disclosed to the professor, I work remotely for the first three weeks and therefore have not collaborated with anyone. Group E.

Problem 1

1)Given discrete random variables X,Y,Z. Give mutual information I(X,Y) and the conditional mutual information I(X,Y—Z). Explain what the conditional mutual information measures.

Mutual information of two random variables X,Y is defined both in terms of the entropy and Kullback-Leibler divergence:

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

The conditional mutual information is also defined in terms of entropy and KL divergence :

$$\begin{aligned} I(X,Y|Z) &= I(X,Y,Z) - I(X,Z) \\ &= H(X) - H(X|Y,Z) - (H(X) - H(X|Z)) \\ &= H(X|Z) - H(X|Y,Z) \\ \text{(similarly:)} \qquad &= H(Y|Z) - H(Y|X,Z) \\ &= \mathbb{E}_Z[\mathcal{KL}(p(X,Y|Z)\|p(X|Z)\cdot p(Y|Z))] \end{aligned}$$

The mutual information is a measure to quantify the information that is shared between random Variables. In other words, it measures the amount of information that can be derived abut one variable after observing the other. The conditional mutual information also measures the mutual information between two variables but given a third random variable is already observed. In other words, Conditional Mutual Information represents an expected value for the mutual information between two random variables given a third.

Consider $x, y, z \in \{0, 1\}$, with joint distribution $p(x, y, z)$ given.

2) Evaluate the quantity I(X,Y) and show it being greater zero.

As we deal with discrete random variables, we draw on the following mutual information definition:

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

We at first have to find the properties $p(x, y), p(x), p(y)$ with the help of Table 1 we thus obtain:

x	y	$p(x,y)$	$p(x)$	$p(y)$	$\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$
0	0	0.336	0.6	0.592	-0.0556
0	1	0.264	0.6	0.408	0.0755
1	0	0.256	0.4	0.592	0.0780
1	1	0.144	0.4	0.408	-0.1250

Just as an example: for $x = 0, y = 0$ we get $p(x = 0, y = 0) = p(x = 0, y = 0|z = 0) + p(x = 0, y = 0|z = 1) = 0.336$. Similarly, we end up with the values $p(x = 0) = 0.192 + 0.144 + 0.048 + 0.216 = 0.6$ and $p(x = 1) = 1 - p(x = 0) = 0.4$. For y analogously. With the values from our new table, we arrive at the mutual information:

$$I(x,y) = 0.336\cdot{-0.0556}+0.264\cdot 0.0755+0.256\cdot 0.0780+0.144\cdot{-0.1250} \approx -0.0187+0.0199+0.0199-0.018 = 0.0031$$

As the result is greater than zero, we can conclude that the two random variables x, y are dependent. In other words, given the value of one variable we have to some extent additional information on the value of the other random variable. Looking at the definition in terms of the KL divergence, MI can only be zero, if $p(x, y) = p(x)\cdot(y)$ which is the definition of independence.

3) Evaluate $I(x, y|z)$ and show that its equal to zero. What does it mean?
Again, conditional mutual information can be formulated as following according to its formulation with respect to KL divergence for discrete random variables:

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

x	y	z	$p(z)$	$p(x,z)$	$p(x,y\|z)$	$p(x\|z)$	$p(y\|z)$	$\log\left(\frac{p(x,y\|z)}{p(x\|z)p(y\|z)}\right)$
0	0	0	0.480	0.24	0.4	0.5	0.8	0
0	1	0			0.1		0.2	0
1	0	0		0.24	0.4	0.5		0
1	1	0			0.1			0
0	0	1	0.520	0.36	0.277	0.692	0.4	0
0	1	1			0.415		0.6	0
1	0	1		0.16	0.123	0.308		0
1	1	1			0.185			0

The values for $p(z), p(x, z)$ are obtained analogously to the previous task. Conditional probabilities follow the Bayes Theorem, e.g. $p(x = 0|Z = 0) = \frac{p(x=0,z=0)}{p(z=0)}$. Similarly, $p(x = 0y = 0|Z = 0) = \frac{p(x=0,z=0,y=0)}{p(z=0)}$. As all the log values are zero, all the terms in the sum all multiplied by zero and we arrive at

$$I(x, y|z) = 0$$

So, given Z, X and Y are independent.

4) Show that $p(x, y, z) = p(x)p(z|x)p(y|z)$ and draw the graph that corresponds to this factorization.

$$p(x,y,z) \overset{\text{Bayes}}{=} p(x,z|y)p(y) \overset{\text{Bayes}}{=} \frac{p(y|x,z)p(x,z)}{p(y)}\cdot p(y) \overset{\text{Bayes}}{=} p(y|x,z)p(x,z) \overset{\text{Bayes}}{=} p(y|x,z)p(x)p(z|x)$$

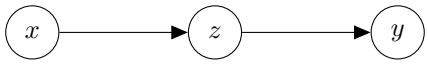
Now, we draw on the conditional independence we derived earlier: X and Y are independent given Z, which in formulas and according to Bishops 8.20 looks like:

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

Plugging it back in yields:

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

which is the sought factorization.
The corresponding graph follows as such:



Problem 2: I found five more general clusters, for which perturbations exist, that would have other independence relations. Those would be analogous to the ones given and therefore are saved here.

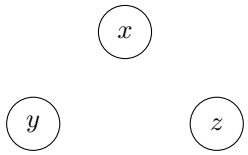


Figure 1: Cluster 1

Independence relations for cluster 1:
No dependece relationships

$$X \perp\!\!\!\perp Y \wedge X \perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z$$

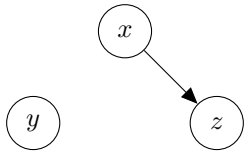


Figure 2: Cluster 2

Independence relations for cluster 2:

$$X \perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z$$

Also holds for the oppositely directed edge. All other perturbations with one dependence relation between x,y or z depend to the same cluster but would have respectively different independence relations.

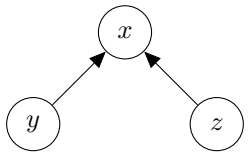


Figure 3: Cluster 3

Independence relations for cluster 3:

$$Z \perp\!\!\!\perp Y \wedge Z \not\perp\!\!\!\perp Y|X \wedge X \not\perp\!\!\!\perp Z \wedge X \not\perp\!\!\!\perp Y$$

Same cluster would entail graphs with all edges going to y or z.

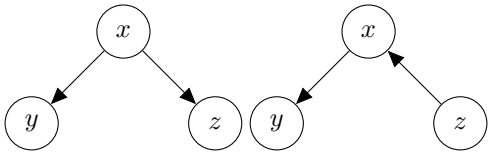


Figure 4: Cluster 4

Independence relations for cluster 4:

$$X \not\perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp Z \wedge Y \perp\!\!\!\perp Z|X \wedge Y \not\perp\!\!\!\perp Z$$

The chain graph (on the right) with oppositely directed edges would be in the cluster too. The same holds for the the variants with once the edges directing to y and x from z and once to z and x from y.

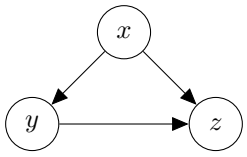


Figure 5: Cluster 5

Independence relations for cluster 5:

$$X \not\perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp Z \wedge Y \not\perp\!\!\!\perp Z$$

Problem 3
1) Given distributions p and q of a continuous random variable, Kullback-Leibler divergence of q from p is defined as

$$\mathcal{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}$$

Evaluate the Kullback-Leibler divergence when $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})$, by applying (380) from the matrix cookbook:

$$\mathbb{E}\left[(\mathbf{x}-\mathbf{m}')^T \mathbf{A} (\mathbf{x}-\mathbf{m}')\right]_{p(\mathbf{x})} = (\boldsymbol{\mu}-\mathbf{m}')^T \mathbf{A} (\boldsymbol{\mu}-\mathbf{m}') + \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma})$$

$$\begin{aligned}
\mathcal{KL}(p\|q) &= -\int p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \\
&= \int p(\mathbf{x}) \log(q(\mathbf{x}) - p(\mathbf{x})) d\mathbf{x} \\
&= \int p(\boldsymbol{x}) \log p(\boldsymbol{x}) dx - \int p(\boldsymbol{x}) \log q(\boldsymbol{x}) dx \\
\text{(fill in Gaussian)} \qquad &= \int p(\mathbf{x}) \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} \\
&\quad + \int p(\mathbf{x}) \left[\frac{D}{2}\log(2\pi) + \frac{1}{2}\log(|\mathbf{L}|) + \frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \right] d\mathbf{x} \\
\text{(Use } \int p(x)dx = 1) \qquad &= -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \frac{1}{2}\log(|\mathbf{L}|) + \left[\int p(x) \frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \right. \\
&\quad \left. - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x} \right] \\
\text{(move terms indep. of x)} \qquad &= \frac{1}{2}\log\left(\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|}\right) \int p(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \right] d\mathbf{x} \\
\text{(Exp. wrt p)} \qquad &= \frac{1}{2}\log\left(\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|}\right) - \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] + \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x}-\mathbf{m})^T \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right]
\end{aligned}$$

As the expectation is according to $p(x) \sim N(\mu, \boldsymbol{\Sigma})$, we can draw on the given rulfe from the matrix cookbook. Applying it to the two expectations yields:

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

Here, the following property was applied: $Tr(I) = \sum_{i=1}^{D} 1 = D$ and the brought back together into one bracket.

2) Entropy of a distribution p is given by

$$\mathcal{H}(p) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$

Derive the entropy of the multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and apply (1)
As the term $-\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$ already appeared above, the derivations look very similar:

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