# Machine Learning 2 - Homework 2

Andrii Skliar, 11636785

deadline: April 16, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Gabriele Bani    Linda Petrini    Pascal Esser    Gautier Dagan

*NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.*

**Problem 1:**

1. Given three discrete random variables $X$, $Y$ and $Z$. Give the denition of the mutual information $I(X;Y)$ and the conditional mutual information $I(X;Y|Z)$. Explain what the (conditional) mutual information measures.

**Solution:** Mutual information is defined as:

$$\begin{aligned} I(X;Y) &= KL(p(x,y)||p(x)p(y)) \\ &= \begin{cases} \text{discrete case: } \sum_{x\in X}\sum_{y\in Y} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)} \\ \text{continuous case: } \int_{x\in X}\int_{y\in Y} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy \end{cases} \end{aligned}$$

Intuitively, mutual information measures how much one random variable tells us about another, or, in other words, how much would we reduce uncertainty about $x$ if we knew $y$.

Conditional mutual information is defined as:

$$\begin{aligned} I(X;Y|Z) &= KL(p(x,y,z)||p(x|z)p(y|z)p(z)) \\ &= \begin{cases} \text{discrete case: } \sum_{z\in Z} p(z)\sum_{x\in X}\sum_{y\in Y} p(x,y|z)\ln\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ \text{continuous case: } \int_{z\in Z} p(z)\int_{x\in X}\int_{y\in Y} p(x,y|z)\ln\frac{p(x,y|z)}{p(x|z)p(y|z)}dxdy \end{cases} \end{aligned}$$

Intuitively, conditional mutual information measures how much one random variable tells us about another given conditioning on the third random variable, or, in other words, how much would we reduce uncertainty about $x$ if we knew $y$ given conditioning on $z$.

2. Evaluate the quantity $I(X;Y)$ and show that it is greater than zero. Hint: Compute the tables for $p(x,y), p(x)$ and for $p(y)$. Moreover, remember that we use the convention that $0\ln(0) := 0$. Interpret this result, i.e. what does it mean that $I(X;Y) > 0$?

**Solution:** First, as suggested, I would like to calculate tables for $p(x,y), p(x)$ and for $p(y)$. I will show general formulas, which will then be used to calculate results. Calculations themselves are avoided.

$$\begin{aligned} p(x,y) &= \sum_{z\in Z} p(x,y,z) = p(x,y,z=0) + p(x,y,z=1) \\ p(x) &= \sum_{y\in Y} p(x,y) = p(x,y=0) + p(x,y=1) \\ p(y) &= \sum_{x\in X} p(x,y) = p(x=0,y) + p(x=1,y) \\ I(X,Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy \end{aligned}$$

Using these formulas, we can calculate all the necessary values, which are summarized in following table:

| x | y | $p(x)$ | $p(y)$ | $p(x,y)$ | $I(x,y)$ |
|---|---|---|---|---|---|
| 0 | 0 | 0.6 | 0.592 | 0.336 | -0.018671 |
| 0 | 1 | 0.6 | 0.408 | 0.264 | 0.019934 |
| 1 | 0 | 0.4 | 0.592 | 0.256 | 0.019958 |
| 1 | 1 | 0.4 | 0.408 | 0.144 | -0.018023 |

These results give us following value for mutual information:
$I(X,Y) = -0.018671 + 0.019934 + 0.019958 + -0.018023 = 0.003197 > 0$.
It means, that random variables $X$ and $Y$ are not independent.

3. Evaluate $I(X;Y|Z)$ and show that it is equal to zero. Hint: Compute the tables for $p(x,y|z), p(x|z)$ and for $p(y|z)$. Interpret this result, i.e. what does it mean that $I(X;Y|Z) = 0$?

**Solution:** Again, as suggested, I would like to first calculate tables for $p(x,y|z), p(x|z)$ and for $p(y|z)$. To do so, we first need to calculate $p(z)$. Similarly to the way we calculated $p(x)$ and $p(y)$, we get following results:

| z | p(z) |
|---|---|
| 0 | 0.48 |
| 1 | 0.52 |

Now we can calculate $p(x,y|z), p(x|z) and p(y|z)$ using following formulas:

$$\begin{aligned}
p(x,y|z) &= \frac{p(x,y,z)}{p(z)} \\
p(x|z) &= \sum_{y \in Y} p(x,y|z) = p(x, y=0|z) + p(x, y=1|z) \\
p(y|z) &= \sum_{x \in X} p(x,y|z) = p(x=0, y|z) + p(x=1, y|z) \\
I(X,Y|Z) &= \sum_{z \in Z} p(z) \sum_{x \in X} \sum_{y \in Y} p(x,y|z) \ln \frac{p(x,y|z)}{p(x|z)p(y|z)}
\end{aligned}$$

Using these formulas, we can calculate all the necessary values, which are summarized in following table:

| x | y | z | p(x) | p(y) | p(z) | $p(x\|z)$ | $p(y\|z)$ | $p(x,y)$ | $p(x,y\|z)$ | $I(X,Y\|Z)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0.6 | 0.592 | 0.48 | 0.5 | 0.8 | 0.336 | 0.4 | 0.0 |
| 0 | 0 | 1 | 0.6 | 0.592 | 0.52 | 0.692308 | 0.4 | 0.336 | 0.276923 | 0.0 |
| 0 | 1 | 0 | 0.6 | 0.408 | 0.48 | 0.5 | 0.2 | 0.264 | 0.1 | 0.0 |
| 0 | 1 | 1 | 0.6 | 0.408 | 0.52 | 0.692308 | 0.6 | 0.264 | 0.415385 | 0.0 |
| 1 | 0 | 0 | 0.4 | 0.592 | 0.48 | 0.5 | 0.8 | 0.256 | 0.4 | 0.0 |
| 1 | 0 | 1 | 0.4 | 0.592 | 0.52 | 0.307692 | 0.4 | 0.256 | 0.123077 | 0.0 |
| 1 | 1 | 0 | 0.4 | 0.408 | 0.48 | 0.5 | 0.2 | 0.144 | 0.1 | 0.0 |
| 1 | 1 | 1 | 0.4 | 0.408 | 0.52 | 0.307692 | 0.6 | 0.144 | 0.184615 | 0.0 |

These results give us following value for mutual information: $I(X,Y|Z) = 0$.
It means, that random variables $X$ and $Y$ are independent given $Z$.

4. Show that $p(x,y,z) = p(x)p(z|x)p(y|z)$, and draw the corresponding directed graph?

**Solution:** It can be proved both numerically and analytically. First, I would like to prove it numerically. The fact, that it is also correct numerically can be easily seen by using previously calculated values. Doing so, we get following table:

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

As you can see, numerically equation $p(x,y,z) = p(x)p(z|x)p(y|z)$ is satisfied.

Now, I would like to also prove it analytically:

$$
\begin{aligned}
p(x,y,z) = p(z,y|x)p(x) &= p(y|x,z)p(z|x)p(x) \\
&= \{\text{as we have proven in the previous point, that } X \text{ and } Y \text{ are independent given } Z, \\
&\text{we can rewrite } p(y|x,z) = p(y|z)\} \\
&= p(y|z)p(z|x)p(x)
\end{aligned}
$$

The corresponding graph looks as suggested in fig. 1:

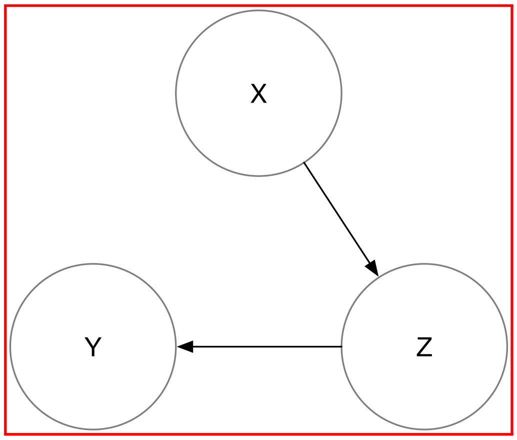


Figure 1: Graph for task 1

**Problem 2:** Consider all the Bayesian networks consisting of three vertices $X$, $Y$ and $Z$. Group them into clusters such that all the graphs in each cluster encode the same set of independence relations. Draw those clusters and write down the set of independence relations for each cluster.

**Solution:** Here, I would like to draw all possible clusters and corresponding Bayesian networks. In total, we have 25 possible BNs and 11 clusters. Independence relations are written as captions to each cluster.

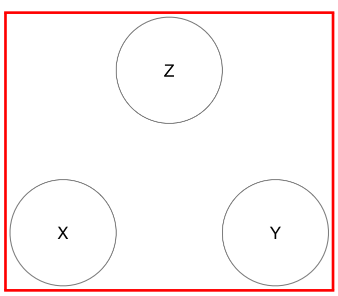


Figure 2: $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Y|Z, X \perp\!\!\!\perp Z|Y, Y \perp\!\!\!\perp Z|X$

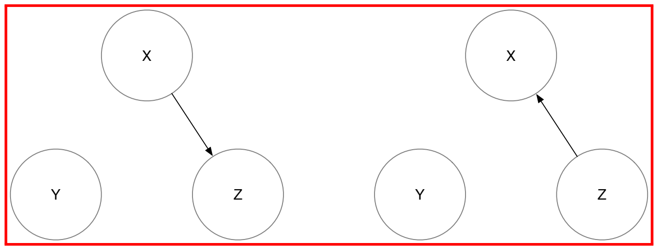


Figure 3: $X \perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \perp\!\!\!\perp Z|X$

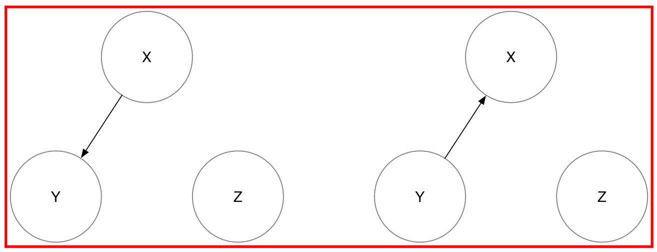


Figure 4: $X \not\perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \perp\!\!\!\perp Z|Y, Y \perp\!\!\!\perp Z|X$

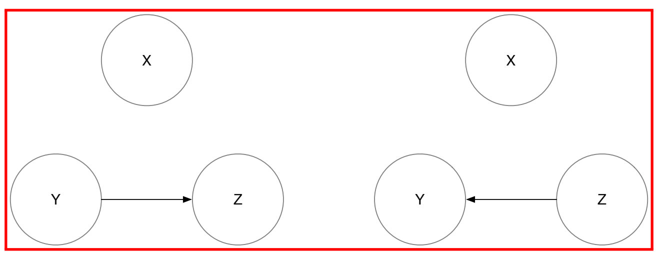


Figure 5: $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \perp\!\!\!\perp Y|Z, X \perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

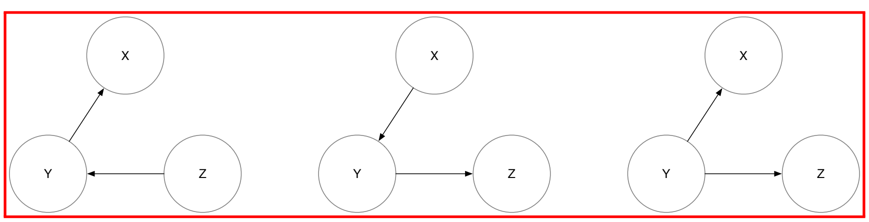


Figure 6: $X \not\perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

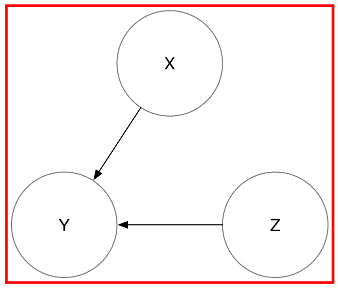


Figure 7: $X \not\perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

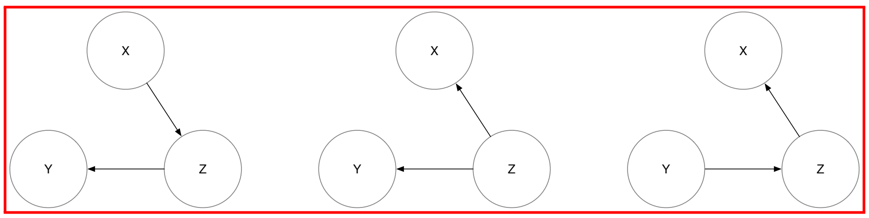


Figure 8: $X \not\perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

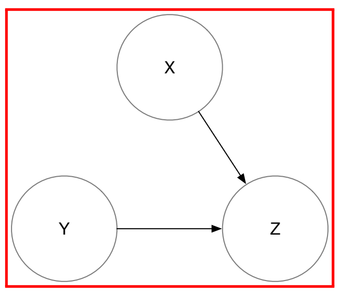


Figure 9: $X \perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

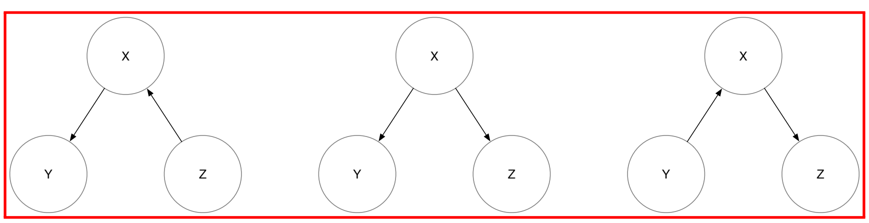


Figure 10: $X \not\perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \perp\!\!\!\perp Z|X$

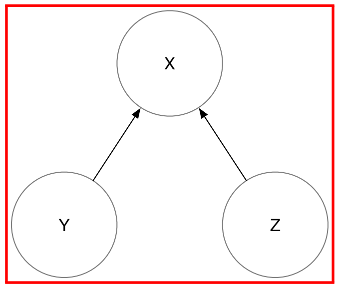


Figure 11: $X \not\perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

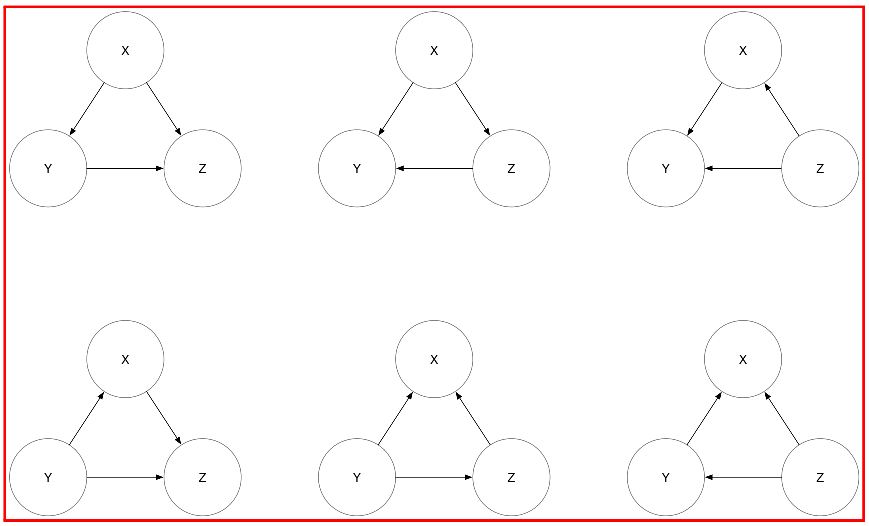


Figure 12: $X \not\perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Y|Z, X \not\perp\!\!\!\perp Z|Y, Y \not\perp\!\!\!\perp Z|X$

**Problem 3:**

1. Given distributions $p$ and $q$ of a continuous random variable, Kullback-Leibler divergence of $q$ from $p$ is defined as:

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

Evaluate the Kullback-Leibler divergence when $p(x) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(x) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})$.

**Solution:** Note: here I will be using following equation (380) from Matrix Cookbook: Assuming that $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, we get following equation: $\mathbb{E}[(\mathbf{x} - \mathbf{m}')\mathbf{A}(\mathbf{x} - \mathbf{m}')] = (\mathbf{m} - \mathbf{m}')^T \mathbf{A}(\mathbf{m} - \mathbf{m}') + Tr(\mathbf{A}\boldsymbol{\Sigma})$

$$
\begin{aligned}
KL(p\|q) = & \int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} = -\int p(\mathbf{x})(\ln q(\mathbf{x}) - \ln p(\mathbf{x})) d\mathbf{x} \\
& = \int p(\mathbf{x})(\ln p(\mathbf{x}) - \ln q(\mathbf{x})) d\mathbf{x} \\
& = \{\text{Using law of unconscious statistician}\} \\
& = \mathbb{E}[\ln p(\mathbf{x}) - \ln q(\mathbf{x})] \\
& = \mathbb{E}[-\frac{1}{2}(D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}|) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) + \frac{1}{2}(D\ln(2\pi) + \ln(|\mathbf{L}|) + (\mathbf{x}-\mathbf{m})^T\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}))] \\
& = \frac{1}{2}\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] + \frac{1}{2}\mathbb{E}[(\mathbf{x}-\mathbf{m})^T\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}))] \\
& = \frac{1}{2}\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2}[(\boldsymbol{\mu}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) + Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma})] \\
& + \frac{1}{2}[(\boldsymbol{\mu}-\mathbf{m})^T\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\boldsymbol{\Sigma})] \\
& = \frac{1}{2}\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2}\cdot[\mathbf{0} + Tr(\mathbf{I})] + \frac{1}{2}[(\boldsymbol{\mu}-\mathbf{m})^T\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\boldsymbol{\Sigma})] \\
& = \frac{1}{2}\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{D}{2} + \frac{1}{2}[(\boldsymbol{\mu}-\mathbf{m})^T\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + Tr(\mathbf{L}^{-1}\boldsymbol{\Sigma})], \text{ where } D \text{ is the dimensionality of } \mathbf{x}.
\end{aligned}
$$

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 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

**Solution:** Here, we can use the derivation from previous part of the task:

$$
\begin{aligned}
\mathcal{H}(p) &= -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x} = -\int -\frac{1}{2}(D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}|) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))p(\mathbf{x})d\mathbf{x} \\
&= \mathbb{E}[\frac{1}{2}(D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}|) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&= \frac{1}{2}[D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}|) + \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]] \\
&= \frac{1}{2}[D\ln(2\pi) + \ln(|\boldsymbol{\Sigma}|) + D], \text{ where } D \text{ is the dimensionality of } \mathbf{x}.
\end{aligned}
$$