# Machine Learning 2 - Homework 2

Pascal M. Esser

April 15, 2018

*Collaborators: Sindy Löwe, Alex Geenen, Andrew Sklyar, Linda Petrini, Davide Belli*

## 1

### 1.1

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

In general we can write the mutual information as (where the second version shows the case for discrete variables):

$$\begin{aligned} I(X;Y) &= \mathcal{D_{KL}}(p(x,y)||p(x)p(y)) && (1)\\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) && (2)\end{aligned}$$

As we can express the mutual information using a KL divergence it also gives us insight into how the distance between the two distributions given by $p(x,y)$ and $p(x)p(y)$ is.

For the conditional mutual information we can write in general and for the discrete case:

$$\begin{aligned} I(X;Y|Z) &= \mathbb{E}_z[\mathcal{D_{KL}}(x,y|z)||p(x|z)p(y|z)] && (3)\\ &= \sum_{z\in Z}\sum_{x\in X}\sum_{y\in Y} p(z)p(x,y|z)\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) && (4)\end{aligned}$$

In general the conditional mutual information gives us a measure of the mutual information of two variables given a third one.

The forms for the discrete cases are used in the subsections below.

### 1.2

*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 $I(X;Y) > 0$?*

Table 1 gives the values for the different parts of the equation. If we sum up the final values in the last row we get $I(X;Y) \approx 0.0033$.

As $(X;Y) = 0$ holds if and only if X and Y are independent random variables and we see that in this case $(X;Y) > 0$ we can conclude that X and Y are not independent.

Table 1: $I(X;Y)$

| $i$ | $p(x_i,y_i)$ | $p(x_i)$ | $p(y_i)$ | $log(p(x_i,y_i)) - log(p(x_i)p(y_i))$ | $p(x_i,y_i) * (log(p(x_i,y_i)) - log(p(x_i)p(y_i)))$ |
|---|---|---|---|---|---|
| 1 | 0.336 | 0.6 | 0.592 | -0.0556 | -0.0187 |
| 2 | — | — | — | — | — |
| 3 | 0.264 | 0.6 | 0.408 | 0.0755 | 0.0200 |
| 4 | — | — | — | — | — |
| 5 | 0.256 | 0.4 | 0.592 | 0.0780 | 0.0200 |
| 6 | — | — | — | — | — |
| 7 | 0.144 | 0.4 | 0.408 | -0.125 | -0.0180 |
| 9 | — | — | — | — | — |

Table 2: $I(X;Y|Z)$ with $p(z=0)=0.48, p(z=1)=0.52$. Further we abbriviate: $\alpha = \log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)$,

| $i$ | $p(x,y\|z)$ | $p(x\|z)$ | $p(y\|z)$ | $\alpha$ | $p(x,y\|z)\alpha$ | $p(z)p(x,y\|z)\alpha$ |
|---|---|---|---|---|---|---|
| 1 | 0.4 | 0.5 | 0.8 | 0 | 0 | 0 |
| 2 | 0.277 | 0.692 | 0.5 | 0.00072 | 0.00020 | 0.00010 |
| 3 | 0.1 | 0.5 | 0.2 | 0 | 0 | 0 |
| 4 | 0.515 | 0.692 | 0.6 | -0.00048 | -0.00020 | -0.00010 |
| 5 | 0.4 | 0.5 | 0.8 | 0 | 0 | 0 |
| 6 | 0.123 | 0.308 | 0.4 | -0.00162 | -0.00020 | -0.00010 |
| 7 | 0.1 | 0.5 | 0.2 | 0 | 0 | 0 |
| 8 | 0.185 | 0.308 | 0.6 | 0.00108 | 0.00020 | 0.00010 |

## 1.3

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

In Table 2 the numerical values of the different parts of $I(X;Y|Z)$ are displayed and by summing up the last column we see that it sums up to zero.

The fact that we can show that $I(X;Y|Z)=0$ means that X and Y are independent given Z.

### 1.4

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

We start from:

$$p(x,y,z) = p(x)p(z|x)p(y|x,z) \tag{5}$$

$$\text{using the results from 1.3} \tag{6}$$

$$= p(x)p(z|x)p(y|z) \tag{7}$$

In general we can only see, that the first equality holds ($p(x,y,z) = p(x)p(z|x)p(y|x,z)$) as this is how the joint probability mass function of three discrete random variables is defined. But in this special case if we have a corresponding directed graph like the on shown in Figure 1 we see the independence of $p(y|x,z)$ from $z$ and therefore we can replace it by $p(y|x)$.

From the given formula alone we can see that we are dealing with a chain case as displayed in Figure 1. We start from a probability of $p(x)$ with an directed edge towards z which give us for $z$: $p(z|x)$ and similar for $y$: $p(y|x)$.

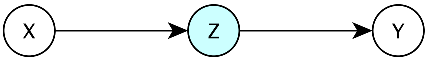

Figure 1: Chain case.

Numerically for the given example we see the prove in Table 3. We calculate $p(x), p(z|x), p(y|z)$ and use this to calculate $p(x,y,z) = p(x)p(z|x)p(y|x)$, which we can use to compare to the values for $p(x,y,z)$ given in the exercise. We see that they match up an therefor see that $p(x,y,z) = p(x)p(z|x)p(y|x)$ holds.

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

| $i$ | $p(x)$ | $p(z\|x)$ | $p(y\|z)$ | $p(x,y,z) = p(x)p(z\|x)p(y\|z)$ |
|---|---|---|---|---|
| 1 | 0.6 | 0.4 | 0.8 | 0.192 |
| 2 | 0.6 | 0.48 | 0.5 | 0.144 |
| 3 | 0.6 | 0.4 | 0.2 | 0.048 |
| 4 | 0.6 | 0.6 | 0.6 | 0.216 |
| 5 | 0.4 | 0.6 | 0.8 | 0.192 |
| 6 | 0.4 | 0.32 | 0.5 | 0.064 |
| 7 | 0.4 | 0.6 | 0.2 | 0.048 |
| 8 | 0.4 | 0.4 | 0.6 | 0.096 |

## 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.*

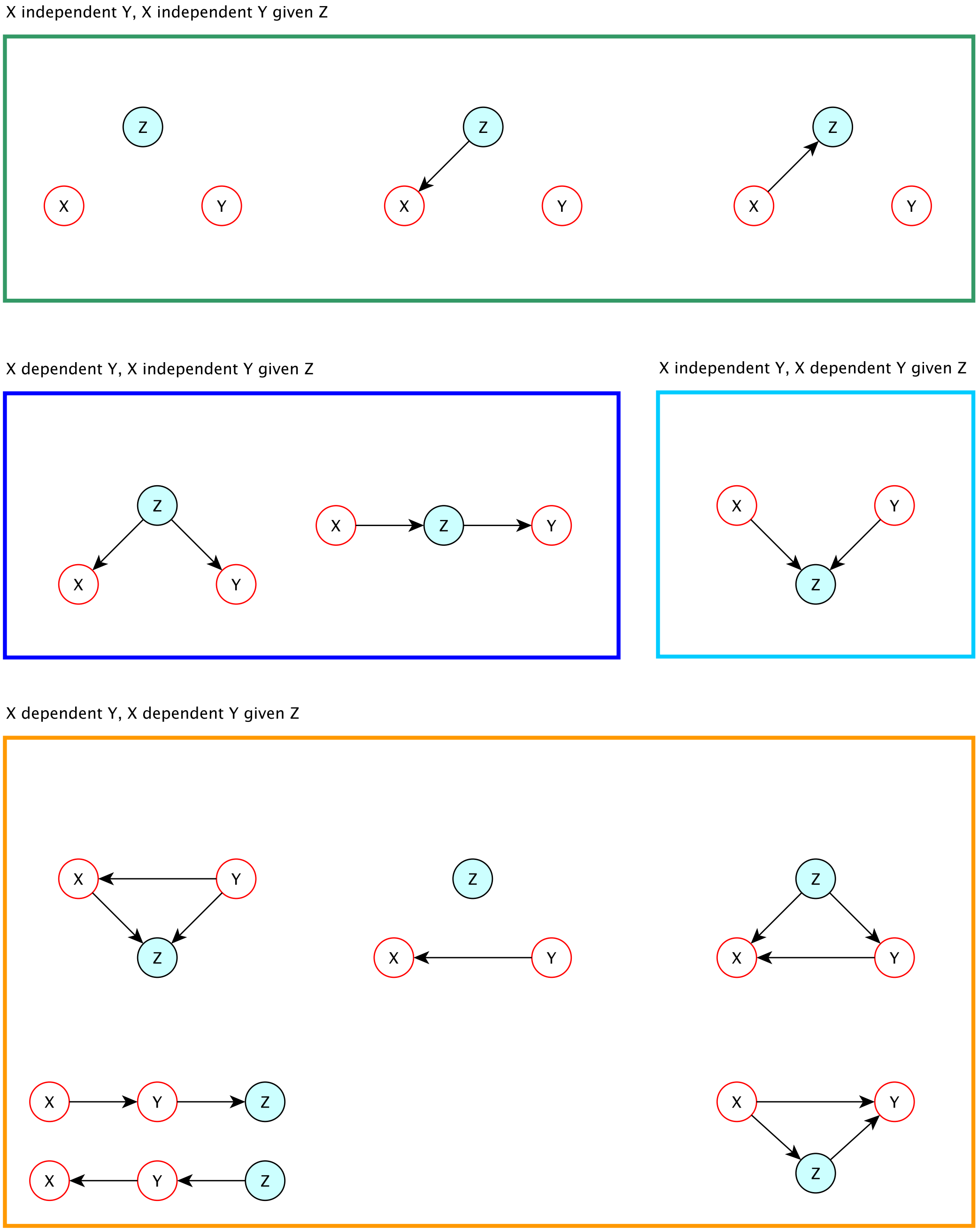
X independent Y, X independent Y given Z
X dependent Y, X independent Y given Z
X independent Y, X dependent Y given Z
X dependent Y, X dependent Y given Z

Figure 2: All the Bayesian networks consisting of three vertices X, Y and Z. Shown is one possible permutation (the other ones are possible to archive by permuting the edges with red borders.) each but the other ones hold wlog.

Figure 2 shows all the Bayesian networks consisting of three vertices X, Y and Z. The plot is structured into the three combinations of independence between X and Y as well as the independence between X and Y given Z that are possible. The middle column shows the three mayor building blocks of BN. In the blue cases we see the fork and chain case where X is dependent on Y but X is not dependent on Y given Z. The light blue case shows the collision case where X is independent of Y but X is dependent on Y given Z. The upper block shows cases where there are no directed edges between X, Z and Y, Z. Naturally in those cases X is independent of Z and also independent of Y given Z. In the bottom black we see versions of the fork case and the collision case where there is and additional edge between X and Y which gives as an direct dependency between X and Y as well as one for X and Y given Z. Furthermore we see versions of the chain case with direct combinations between X and Z and a version where there is only one edge between X and Y but none to Z. These combinations give us dependency between X and Y and also dependencies between X and Y given Z.

Plotted are each one possible permutation of the case of interest. The edges with red boundaries can be perpetuated in each graph to create a new BN but with the same independence relationship.

# 3

## 3.1

*Given distributions p and q of a continuous random variable, Kullback-Leibler divergence of q from p is defined as* $\mathcal{D}_{\mathcal{KL}}(p||q) = -\int p(\boldsymbol{x})ln\left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right] dx$ *Evaluate the Kullback-Leibler divergence when* $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{L})$

$$\mathcal{D}_{\mathcal{KL}}(p||q) = -\int p(\boldsymbol{x})ln\left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right] dx \tag{8}$$

$$= -\int p(\boldsymbol{x})\left[\ln q(\boldsymbol{x}) - \ln p(\boldsymbol{x})\right] dx \tag{9}$$

$$= -\int p(\boldsymbol{x})\ln q(\boldsymbol{x})dx \int p(\boldsymbol{x})\ln p(\boldsymbol{x})dx \tag{10}$$

$$= -\int p(\boldsymbol{x})\ln\left[\frac{1}{2\pi^{D/2}}|\boldsymbol{L}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right]\right] dx \tag{11}$$

$$+\int p(\boldsymbol{x})\ln\left[\frac{1}{2\pi^{D/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right] dx \tag{12}$$

$$=\frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|}\int p(\boldsymbol{x})dx + \frac{1}{2}\int(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})p(\boldsymbol{x})dx - \frac{1}{2}\int(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})p(\boldsymbol{x})dx \tag{13}$$

$$\text{using the law of the unconscious statistician and } \int p(\boldsymbol{x})dx = 1 \tag{14}$$

$$=\frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right] - \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{15}$$

$$\text{using Matrix Cookbook 380 gives us} \tag{16}$$

$$\mathbb{E}[\boldsymbol{x}-\boldsymbol{m}']^\top\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}')] = (\boldsymbol{m}-\boldsymbol{m}')^\top\boldsymbol{A}(\boldsymbol{m}-\boldsymbol{m}') + Tr(\boldsymbol{A\Sigma}) \tag{17}$$

$$\text{for } \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \Sigma)\text{. As with } \mathcal{D}_{\mathcal{KL}}(p||q) \text{ we are trying to match } p \text{ we get } m \equiv \mu\text{. Therefore:} \tag{18}$$

$$=\frac{1}{2}\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + \frac{1}{2}\left[(\boldsymbol{\mu}-\boldsymbol{m})^\top\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + Tr(\boldsymbol{L}^{-1}\boldsymbol{\Sigma})\right] - \frac{1}{2}\left[(\boldsymbol{\mu}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) + Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma})\right] \tag{19}$$

$$=\frac{1}{2}\left[\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} + (\boldsymbol{\mu}-\boldsymbol{m})^\top\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + Tr(\boldsymbol{L}^{-1}\boldsymbol{\Sigma}) - D\right] \tag{20}$$

## 3.2

*Entropy of a distribution p is given by $\mathcal{H}(\boldsymbol{x}) = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) dx$. Derive the entropy of the multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$*

We use the same approach on decomposition and application of the matrix cookbook as shown above. Therefore this version is presented slightly shorter.

$$
\begin{aligned}
\mathcal{H}(\boldsymbol{x}) &= -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) dx && (21)\\
&= -\int p(\boldsymbol{x}) \ln \left[ \frac{1}{2\pi^{D/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp \left[ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right] \right] dx && (22)\\
&\quad \text{using the law of the unconscious statistician} && (23)\\
&= \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \mathbb{E} \left[ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right] && (24)\\
&\quad \text{using Matrix Cookbook 380} && (25)\\
&= \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \left[ -\frac{1}{2} (\boldsymbol{\mu}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\boldsymbol{\mu}) \right] Tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) && (26)\\
&= \frac{1}{2} \left[ D \ln(2\pi) + \ln |\boldsymbol{\Sigma}| + D \right] && (27)
\end{aligned}
$$