University of Stockholm
MT7037-VT19 Statistical Information Theory
Final Examination

Marina Herrera Sarrias

March 19, 2019

1 Exercise 1

(a) Let X and Y be discrete random variables. Show that H ((X+Y)|X) = H(Y|X). Argue that if X and Y are independent then H(X+Y) ≥ H(Y).

We wish to prove that the conditional entropy of Z ($Z = g(X,Y) = X + Y$) is the same as the conditional entropy of Y ($Y = Y'$ where $Y' = Z - X$) given X.

As Z is a function of (X,Y) and the probability of Z taking a value is linked to the probability of X having one. Both probabilities are the same as Z will happen if and only if X happens.

$$p(Z = z|X = x) = p(Y = z - x|X = x) = p(Y = y'|X = x)$$

We will now prove the equality:

$$\begin{aligned} H(Z|X) &= \sum p(x) H(Z|X = x) \\ &= \sum_{x\in\chi} p(X = x) \sum_{z\in\mathcal{Z}} p(Z = z|X = x) \log p(Z = z|X = x) \\ &= \sum_{x\in\chi} p(X = x) \sum_{y'\in\mathcal{Y}'} p(Y = y'|X = x) \log p(Y = y'|X = x) \qquad (1) \\ &= \sum p(x) H(Y|X = x) \\ &= H(Y|X) \end{aligned}$$

Referring to their entropy we can say that:

$$H(X,Y) = H(X) + H(Y|X) \geq H(g(X,Y)) = H(Z)$$

and consequently, knowing that conditioning reduces entropy:

$$H(Z) \geq H(Z|X) = H(Y|X)$$

In the case that X and Y are independent discrete random variables. We can say that the addition of independent random variables adds uncertainty:

$$H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$$

or similarly,

$$H(Z) \geq H(Z|Y) = H(X|Y) = H(X)$$

So,

$$\max\{H(X), H(Y)\} \leq H(Z)$$

■

(b) Show that among all N - valued random variables (i.e. with values k = 1,2,...) with expected value μ ,the geometric distributed random variable with expected value μ has the maximum value of Shannon entropy. Reminder: The probability function in a geometric distribution is $p(k) = p(1-p)^{k-1}, k = 1, 2, ...$

Assuming that

$$P(X = k) = p(1-p)^{k-1}$$

Where k and $X \in \mathbb{N}_+$. Knowing that:

$$\sum_{k=1}^{+\infty}(k-1)(1-p)^{k-1} = \frac{1-p}{(1-(1-p))^2}, \qquad \sum_{k=1}^{+\infty}(1-p)^{k-1} = \frac{1}{(1-(1-p))}$$

The entropy of the geometric distribution is derived as:

$$\begin{aligned}
H(X) &= -\sum_{k=1}^{+\infty} p(1-p)^{k-1} \log p(1-p)^{k-1} \\
&= -\sum_{k=1}^{+\infty} p(1-p)^{k-1} \log p - \sum_{k=1}^{+\infty}(k-1)p(1-p)^{k-1} \log(1-p) \\
&= -[\sum_{k=0}^{+\infty} p(1-p)^{k} \log p - \sum_{k=0}^{+\infty} kp(1-p)^{k} \log{(1-p)}] \\
&= -\frac{p}{1-(1-p)} \log p - \frac{(1-p)}{(1-(1-p))^2} \log p(1-p) \\
&= -\log p - \frac{(1-p)}{p} \log p(1-p)
\end{aligned}$$

We are now going to solve the maximization problem:

$$\arg\max_{p} \Big(-\sum_{k=1}^{+\infty} p_k \log p_k \Big)$$

Subject to the constraints:

$$\sum_{k=1}^{+\infty} p_k = 1$$

$$\sum_{k=1}^{+\infty} kp_k = \mu$$

We will use the Lagrange multipliers to obtain the general form of the p_k distribution.

$$L(p_k, \lambda_1, \lambda_2) = -\Big(\sum_{k=1}^{+\infty} p_k \log p_k\Big) + \lambda_1\Big(\sum_{k=1}^{+\infty} p_k - 1\Big) + \lambda_2\Big(\sum_{k=0}^{+\infty} kp_k - \mu\Big)$$

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_1} = 0 &= \sum_{k=1}^{+\infty} p_k - 1 \\
&= \exp^{(-1+\lambda_1)} \sum_{k=1}^{+\infty} \exp^{(\lambda_2)k} \\
&= \exp^{(-1+\lambda_1)} \Big(\frac{1}{1-\exp^{\lambda_2}} - 1\Big) = 1 \\
&= \exp^{(-1+\lambda_1)} \Big(\frac{\exp^{\lambda_2}}{1-\exp^{\lambda_2}}\Big) = 1 \\
\frac{\partial L}{\partial \lambda_2} = 0 &= \sum_{k=0}^{+\infty} kp_k - \mu \\
&= \exp^{(-1+\lambda_1)} \sum_{k=1}^{+\infty} k\exp^{(\lambda_2)k} \\
&= \exp^{(-1+\lambda_1)} \frac{\exp^{\lambda_2}}{(1-\exp^{\lambda_2})^2} = \mu \\
\frac{\partial L}{\partial p_k} = 0 &= -\log p_k - 1 + \lambda_1 + \lambda_2 k
\end{aligned}$$

Where,

$$\mu = \frac{1}{1-\exp^{\lambda_2}}$$
$$\lambda_2 = \log\left(\frac{\mu-1}{\mu}\right)$$
$$\lambda_1 = \log\left(\frac{1}{\mu-1}\right)+1$$

Finally, the p_k distribution is derived as:

$$\begin{aligned}
p_k &= \exp^{(-1+\lambda_1+\lambda_2 k)} \\
&= \exp^{(-1+\lambda_1+\lambda_2)}\exp^{(k-1)\lambda_2} \\
&= \frac{1}{\mu}\exp^{(k-1)\lambda_2} \\
&= \frac{1}{\mu}\exp^{(k-1)\log\left(\frac{\mu-1}{\mu}\right)} \\
&= \left(\frac{1}{\mu}\right)\left(\frac{\mu-1}{\mu}\right)^{k-1}
\end{aligned} \tag{2}$$

We also know that the mean of the geometric distribution is:

$$\begin{aligned}
\mu = E[X] &= \sum_{k=1}^{+\infty} k(1-p)^{k-1}p \\
&= p\sum_{k=1}^{+\infty} k(1-p)^{k-1} \\
&= p\frac{1}{(1-(1-p))^2} \\
&= \frac{1}{p}
\end{aligned} \tag{3}$$

If we now plug the results obtained in eq.(3) into eq.(2) we can verify that p_k is the geometric distribution i.e, the distribution that maximizes the entropy.

We could also prove it using the Gibb's inequality eq.(5), which is just the difference of the Kullback–Leibler divergence eq.(4). Knowing that the random variable X distributed as $p_k \neq 0$ and for the random variable Y

distributed as q_k, where $k \in \mathbb{Z}^+$

$$\begin{aligned}\mathbb{D}_{KL}(Y|X) &= \sum_{k=1}^{+\infty} q_k \log \frac{q_k}{p_k} \geq 0 \\ &= \sum_{k=1}^{+\infty} q_k \log q_k - \sum_{k=1}^{+\infty} q_k \log p_k\end{aligned} \tag{4}$$

$$-\sum_{k=1}^{+\infty} q_k \log q_k \leq -\sum_{k=1}^{+\infty} q_k \log p_k \tag{5}$$

and knowing that:

$$H(Y) = -\sum_{k=1}^{+\infty} q_k \log q_k$$

$$\begin{aligned}-\sum_{k=1}^{+\infty} q_k \log p_k &= -\sum_{k=1}^{+\infty} q_k(\log p + (1-k)\log(1-p)) \\ &= -\log p \sum_{k=1}^{+\infty} q_k - \sum_{k=1}^{+\infty} q_k(1-k)\log(1-p)) \\ &= -\log p - \sum_{k=1}^{+\infty} q_k(1-k)\log(1-p) \\ &= -\log p - \log(1-p)\sum_{k=1}^{+\infty} q_k(1-k) \\ &= -\log p - \log(1-p)\sum_{k=1}^{+\infty} q_k + \log(1-p)\sum_{k=1}^{+\infty} kq_k \\ &= -\log p - \log(1-p) + \log(1-p)\sum_{k=1}^{+\infty} kq_k \\ &= -\log p - \log(1-p) + \log(1-p)\mathbb{E}[Y] \\ &= -\log p + \log(1-p)(\mathbb{E}[Y]-1) \\ &= -\log p - \frac{(1-p)}{p}\log p(1-p) = \mathbb{H}(X)\end{aligned}$$

$$\mathbb{H}(Y) \leq \mathbb{H}(X)$$

■

(c) Show that
(i) $\mathrm{H}(X_1,X_2,X_3) \leq \frac{1}{2}$ $[\mathrm{H}(X_1,X_2)$ $+\mathrm{H}(X_2,X_3)$ $+\mathrm{H}(X_1,X_3)]$

(ii)$\mathrm{H}(X_1,X_2,X_3) \geq \frac{1}{2}$ $[\mathrm{H}(X_1,X_2\ |X_2)$+$\mathrm{H}(X_2,X_3|X_1)$+$\mathrm{H}(X_1,X_3|X_2)]$

Using the chain rule of entropy we can prove that a collection of random variables are the sum of the conditional entropies, such that:

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1)$$

We also know that as conditioning reduces entropy:

$$H(X_n) \geq H(X_n|X_1) \geq H(X_n|X_{n-1}, X_1)$$

(i)

The left hand side of the inequality is:

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \end{aligned} \tag{6}$$

While , the right hand side is:

$$\begin{aligned} H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3) &= H(X_1) + H(X_2|X_1) + H(X_2) \\ &= +H(X_3|X_2) + H(X_1) + H(X_3|X_1) \\ &= 2H(X_1) + H(X_2|X_1) + H(X_2) \\ &= +H(X_3|X_2) + H(X_3|X_1) \end{aligned} \tag{7}$$

Now, referring to the results obtained in eq.(7) we could also say that:

$$\begin{aligned} &\frac{1}{2}[H(X_2) + H(X_2|X_1) + \\ &H(X_3|X_2) + H(X_3|X_1)] \geq +\frac{1}{2}[H(X_2|X_1) + H(X_2|X_1) + \\ &\qquad\qquad H(X_3|X_2, X_1) + H(X_3|X_2, X_1)] \\ &\qquad\qquad \geq H(X_2|X_1) + H(X_3|X_2, X_1) \end{aligned} \tag{8}$$

and, because of the above, we can finally prove that:

$$H(X_1, X_2, X_3) \leq H(X1) + \frac{1}{2}[H(X_2) + H(X_2|X_1) + H(X_3|X_2) + H(X_3|X_1)]$$

■

(ii)

The left hand side of the inequality is the same as in eq.(6) while the right hand side is as follows:

$$\begin{aligned}
&= \frac{1}{2}[H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_1, X_3|X_2)] \\
&= \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) + H(X_2|X_1) \\
&+ H(X_3|X_2, X_1) + H(X_1|X_2) + H(X_3|X_2, X_1)] \qquad (9) \\
&= H(X_3|X_2, X_1) + \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
&+ H(X_2|X_1) + H(X_1|X_2)]
\end{aligned}$$

Following the same intuition as in eq.(8) we can say about the results obtained in eq.(9) that:

$$\begin{aligned}
&\frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
&\quad + H(X_2|X_1) + H(X_1|X_2)] \leq \frac{1}{2}[H(X_2|X_1) + H(X_2|X_1) + H(X_1) + H(X_1)] \\
&\quad \leq H(X_1) + H(X_2|X_1)
\end{aligned}$$

We can now verify that:

$$\begin{aligned}
H(X_1, X_2, X_3) \geq{}& H(X_3|X_2, X_1) + \frac{1}{2}[H(X_1|X_3) + H(X_2|X_3, X_1) \\
&+ H(X_2|X_1) + H(X_1|X_2)]
\end{aligned}$$

■

(d) Suppose that (X, Y, Z) are jointly normal distributed and that $X \to Y \to Z \to$ forms a Markov chain. Let X and Y have the correlation coefficient $\rho 1$ and let Y and Z have the correlation coefficient $\rho 2$. Find mutual information $I(X; Z)$.

The mutual information $I(X; Z)$ is the reduction in the uncertainty of X due to the knowledge of Z and it is represented as:

$$\begin{aligned} I(X;Z) &= H(X) + H(Z) - H(X,Z) \\ &= H(X) + H(X|Z) \end{aligned} \tag{10}$$

Furthermore the entropy of a normal distribution is derived as:

$$h(X) = -\int_{-\infty}^{\infty} \phi(X) \log \phi(X) dx$$

Where the density function $\phi(X)$ is:

$$\phi(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{11}$$

and then, plugging eq.(11) in eq.(10) we obtain:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \phi(X) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right) dx \\ &= -\int_{-\infty}^{\infty} \phi(X) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx + \log(e) \int_{-\infty}^{\infty} \phi(X) (\frac{-(x-\mu)^2}{2\sigma^2}) dx \\ &= -\int_{-\infty}^{\infty} \phi(X) \log \frac{1}{\sqrt{2\pi\sigma^2}} dx - \log(e) \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 \phi(X) dx \\ &= \frac{1}{2} \log\big(2\pi\sigma^2\big) - \log e \frac{1}{2} \\ &= \frac{1}{2} [\log 2\pi\sigma^2 - \log e] \\ &= -\frac{1}{2} \log[2\pi e\sigma^2] \end{aligned}$$

Since (X, Y, Z) are jointly normal distributed X and Z will also be. Their covariance matrix $\Sigma_{x,z}$ is defined as:

$$\Sigma_{x,z} = \begin{bmatrix} {\sigma_x}^2 & \sigma_x \sigma_z \rho_{xz} \\ \sigma_x \sigma_z \rho_{xz} & {\sigma_z}^2 \end{bmatrix}$$

Now, going back to the mutual information $I(X;Z)$ defined on eq.(10) we obtain:

$$\begin{aligned}I(X;Z) &= \frac{1}{2}\log\bigl(2\pi e{\sigma_x}^2\bigr) + \frac{1}{2}\log\bigl(2\pi e{\sigma_z}^2\bigr) - \frac{1}{2}\log\bigl(2\pi e^2|\Sigma_{x,z}|\bigr)\\ &= \frac{1}{2}\log\bigl(2\pi e{\sigma_x}^2\bigr) + \frac{1}{2}\log\bigl(2\pi e{\sigma_z}^2\bigr) - \log(2\pi e) - \frac{1}{2}|\Sigma_{x,z}|\end{aligned}$$

therefore;

$$I(X;Z) = -\frac{1}{2}|\Sigma_{x,z}|$$

Where,

$$\begin{aligned}|\Sigma_{x,z}| &= {\sigma_x}^2{\sigma_z}^2 - {\sigma_x}^2{\sigma_z}^2{\rho_{xz}}^2\\ &= {\sigma_x}^2{\sigma_z}^2(1 - {\rho_{xz}}^2)\end{aligned}$$

If we now assume that $(X,Y,Z) \sim \mathcal{N}(0,1)$:

$$I(X;Z) = -\frac{1}{2}\log\bigl(1 - {\rho_{xz}}^2\bigr) \tag{12}$$

The Markov chain $X \to Y \to Z$ implies that the conditional distribution of Z depends only on Y and is conditionally independent of X. The joint probability of the random variables (X,Y,Z) is:

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

and it implies conditional independence as:

$$\begin{aligned}p(x,z|y) &= \frac{p(x,y,z)}{p(y)}\\ &= \frac{p(x)p(x|y)p(z|y)}{p(y)}\\ &= \frac{p(x,y)p(z|y)}{p(y)}\\ &= \frac{p(x|y)p(y)p(z|y)}{p(y)}\\ &= p(x|y)p(z|y)\end{aligned} \tag{13}$$

Then the correlation coefficient ρ_{xz} is obtained as:

$$\begin{aligned}
\rho_{xz} &= \frac{E[XZ] - E[X]E[Z]}{\sigma_x \sigma_z} \\
&= \frac{E[XZ]}{\sigma_x \sigma_z} \\
&= \frac{E[E[XZ|Y]]}{\sigma_x \sigma_z} \\
&= \frac{E[E[X|Y]E[Z|Y]]}{\sigma_x \sigma_z} \\
&= \frac{E[\frac{E[Y]-E[X]cov(X,Y)}{\sigma_y} + \frac{cov(X,Y)}{\sigma_y}Y][\frac{E[Y]-E[Z]cov(Z,Y)}{\sigma_y} + \frac{cov(Z,Y)}{\sigma_y}Y]}{\sigma_x \sigma_z} \\
&= \frac{E[\frac{cov(X,Y)}{\sigma_y}Y][\frac{cov(Z,Y)}{\sigma_y}Y]}{\sigma_x \sigma_z} \\
&= \frac{E[\sigma_x \rho_{xy} Y][\sigma_z \rho_{zy} Y]}{\sigma_x \sigma_z} \\
&= \rho_{xy}\rho_{zy}E[Y^2] \\
&= \rho_{xy}\rho_{zy}
\end{aligned}$$

Plugging the results into eq.(12), we can now conclude that:

$$\begin{aligned}
I(X;Z) &= -\frac{1}{2}\log\bigl(1 - {\rho_{xy}}^2 {\rho_{zy}}^2\bigr) \\
&= -\frac{1}{2}\log\bigl(1 - {\rho_1}^2 {\rho_2}^2\bigr)
\end{aligned}$$

■

2 Exercise 2

(a) Define the transfer entropy, starting from Schreiber's definition in Phys. Rev. Lett., 85, 461, 2000, in terms of mutual information and Shannon entropy.

The transfer entropy $T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t)$ from the source variable $\mathbb{Y}$ to the target variable $\mathbb{X}$ is the information shared between $\mathbb{Y}$'s past and $\mathbb{X}$'s present given the knowledge of $\mathbb{X}$'s past. it measures how much information the source variable provides about state transitions in the target variable.

The random time series processes $\mathbb{X}$ and $\mathbb{Y}$ are a join of sequential processes that evolve over time t, with a history length of k for the target variable and j for the source, in which most of the cases $l = 1$ or $l = k$ i.e $\mathbb{X}_t^{(k)} = (X_t, X_{t-1}, ..., X_{t-k+1})$ and $\mathbb{Y}_t^{(l)} = (Y_t, Y_{t-1}, ..., Y_{t-l+1})$.

The state of the process Y depends only on its own past and do not depend in any matter on X, i.e there is a zero information transfer between $X \to Y$. Unlike Y, the current state of X do not depends on its own past if not it depends probabilistically on the state of Y at the previous time step. i.e $X_t = Y_{t-1}$.

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = \sum_{X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}} p(x_t, {\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)}) \log_2 \frac{p(x_t|{\mathbb{x}_{t-1}}^{(k)}, {\mathbb{y}_{t-1}}^{(l)})}{p(x_t|{\mathbb{x}_{t-1}}^{(k)})}$$

$$\begin{aligned} I(X_t, {\mathbb{Y}_{t-1}}^{(L)}|{\mathbb{X}_{t-1}}^{(k)}) &= \sum_{X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)}} \mathbb{E}_{p(X_t, {\mathbb{X}_{t-1}}^{(k)}, {\mathbb{Y}_{t-1}}^{(l)})} \log \frac{p(x_t, {\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})}{p(x_t|{\mathbb{x}_{t-1}}^{(k)})p({\mathbb{y}_{t-1}}^{(l)}|{\mathbb{x}_{t-1}}^{(k)})} \\ &= H(X_t|{\mathbb{X}_{t-1}}^{(k)}) - H(X_t|{\mathbb{Y}_{t-1}}^{(L)}, {\mathbb{X}_{t-1}}^{(k)}) \end{aligned}$$

$$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t) = I(X_t, {\mathbb{Y}_{t-1}}^{(l)}|{\mathbb{X}_{t-1}}^{(k)})$$

$T^{(k,l)}_{\mathbb{Y}\to\mathbb{X}}(t)$ measures the deviation from the Markov property which equals the conditional mutual information.

(b) Why is mutual information not a good measure for an information transfer?

(c) Does Schreiber's definition of transfer entropy coincide with the definition of mutual information?

(d) What is the meaning of local transfer entropy?

(e) Is the Granger causality concept the same as transfer entropy? Explain possible similarities and differences.

(f) To complete this part you need to read the paper by Lizier and Prokopenko in Eur. Phys. J. B 73, 605-615, 2010. *i*) Does information transfer have the same meaning as information flow in the opinion of the authors of the paper? Explain possible similarities and differences.
ii) What is the difference between interventional and standard conditional probabilities? *iii*) Explain the difference between local transfer entropy and local information flow.
iv)What are the advantages of the definition of local information flow according to the authors of the paper.

3 Exercise 3

To understand better the concepts of multi-information and its decomposition, this exercise requires you to work out some details in the paper by Schneidman et al., Phys. Rev. Lett., 91, 238701, 2003 discussed in the class.

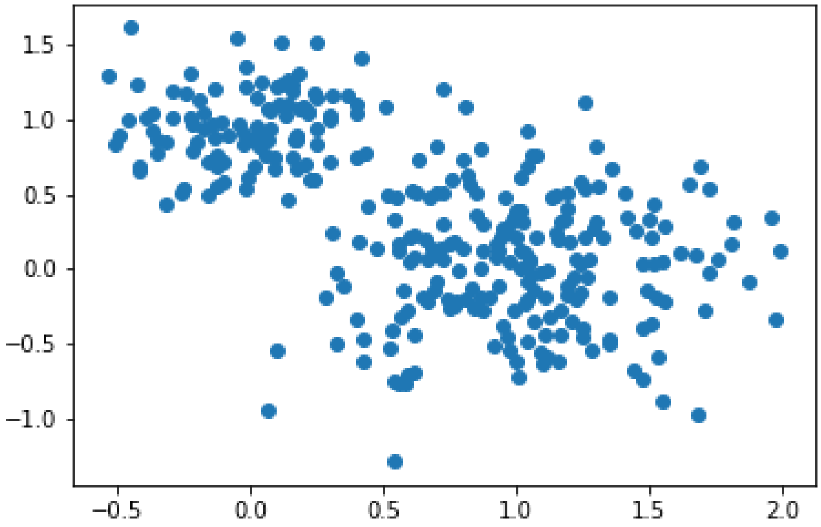
(a) For the case of three binary variables (i.e., x_1, x_2, x_3 equal to either 0 or1) with all pairwise marginals known, derive with clear steps the maximum entropy distribution in terms of the Lagrange multipliers. Hint: You do not need to solve for the Lagrange multipliers and your answer should look like Eq. 1, namely, the Ising model, in Schneidman et al., Nature, 440, 1007, 2006.

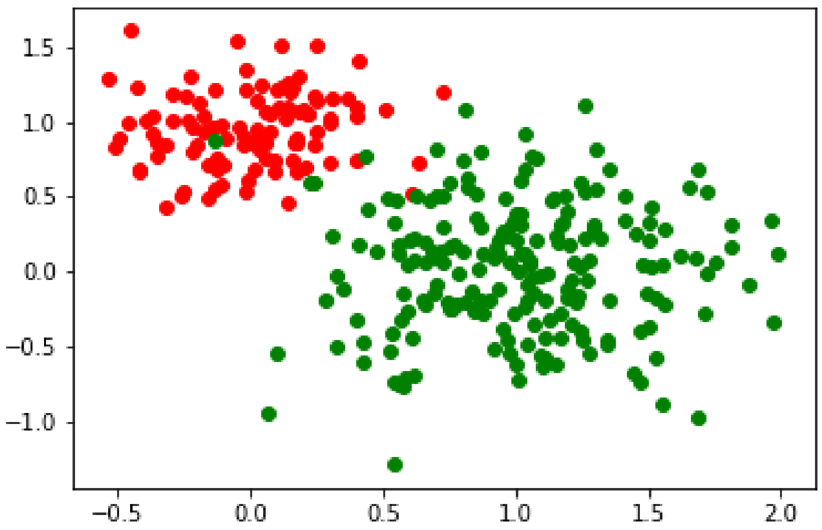
(b) Can you derive *Eq.*(8) of the paper and correct the typos? Hint: Use the Venn diagram. This exercise shows that 1) although the area-information correspondence of the Venn diagram does not hold in the 3-variable case, the Venn diagram is still useful when deriving relations between different information quantities; 2) mistakes exist in published papers, so always be critical in reading them.

(c) Show that the connected information of order k in *Eq.*(6) of the paper can be written as a relative entropy. When does it equal to zero?

4 Exercise 4

(a) Plot the data points on the x-y plane to see how they look like and prepare the element-to-element distance matrix $d(\vec{x}_i, \vec{x}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ $(i, j = 1, 2, ..., 300)$ that will be the input for the clustering algorithm. Also assume each data point carries the same weight, i.e., $p(\vec{x}_i) = 1/300$ for all i.





(b) Write a code (no restriction on the program language) to implement the Blahut-Arimoto algorithm discussed in the class to evaluate the clustering membership probability,, with fixed number of clusters, and compression distortion tradeoff parameter. Your code should implement a multiple run each starting with random initial conditions. Note: Your source code should include clear comments/documentations to describe what is evaluating. We may later randomly ask a few students, especially those without clear documentations, to demonstrate how their code works.

(c) Run your program to construct the information curves for $N_c = 2, 3$ and 4. Hint: choose different values of β in between 1 to 50.

(d) As we have already known that the correct number of clusters is 2, propose a reasonable way using the quantities evaluated from your code and the information curves to correctly identify the number of clusters. You should clearly explain your rationale and show explicitly which quantities, graphs and/or curves are used in the identification.