Probabilistic Modelling and Reasoning Project

Empirical Independence Tests

Francisca Mulebya

April 29, 2024

Introduction

1 Graphical Models

Graphical Models are structural models that represent data in a graph i.e. nodes as instances connected through the edges as the interactions. Generally, the structure can either be directed (edges with directions) or undirected (undirectional edges) for instance concept maps, communication networks etc. A graphical model which represents the conditional dependencies assumption(s) of variables as the probabilistic interaction is known as the Probabilistic Graphical Model.

In Machine Learning, Probabilistic Graphical Models are used as a tool for modelling as well as inference. ([Murphy, 2023]) These models can be used specifically in Bioinformatics (for example in analysing gene interactions), Natural Language Processing (NLP - modelling and inferring words and phrases), Epidemiology (modelling and inference), Recommender Systems (predicting preferences based on past inputs) etc. The framework of probabilistic graphical models provides a mechanism for exploiting structures in complex distributions in any dimensional space to describe them compactly, and in a way that allows them to be constructed and utilised effectively[Koller and Friedman, 2009].

Some popular graphs for modelling are belief networks, Markov networks, chain graphs and influence diagrams whereas for inference, suitable GM for which an algorithm can be readily applied in factor graphs, junction trees [Barber, 2012].

Since the variable(s) dependency is the main property in probabilistic models then it's important to check that in the data to form a good model that can best fit the data and for better inference (if applicable). To achieve this we need to do 'Structure Learning' from the data.

1.1 Structure Learning

Structure Learning in graphical models means learning of model structure as well as parameter(s). Consider the case in which the dataset $\mathcal{D}$ is complete (i.e. there are no missing observations) as the simple starting point; according to [Barber, 2012], for D variables, there is an exponentially large number (in D) of structures which we can't possibly search all making structure learning a computationally challenging problem thus we must rely on constraints (e.g. having at most 1 parent i.e. a tree) and heuristics search/algorithms to help guide the search. Also for inference, we need score-based algorithms to determine the 'best' model.

pgm suggest that we need Structure Learning for

- Density Estimation
- Knowledge Discovery
- Optimization.

In Learning we rely on the probabilistic interaction between variables which is the conditional independence, and likely we expect to experience strong dependency for data with few variables which have many interactions and on the other hand, weak dependency for data with many variables which have few interactions.

For big structures though, we can't decide on the property without an evidence to suggest so. In forming a graphical model, it is crucial to observe conditional independence.

1.2 Conditional Independence

As discussed above, this is a core property of probabilistic graphical models whether they are directed or undirected graphs. With this property, we see that, having information about the variables' descendant (common effect) couples its parents' variables. With this property also, we can simplify the distribution by factorising the conditional distribution (based on dependency assumption).

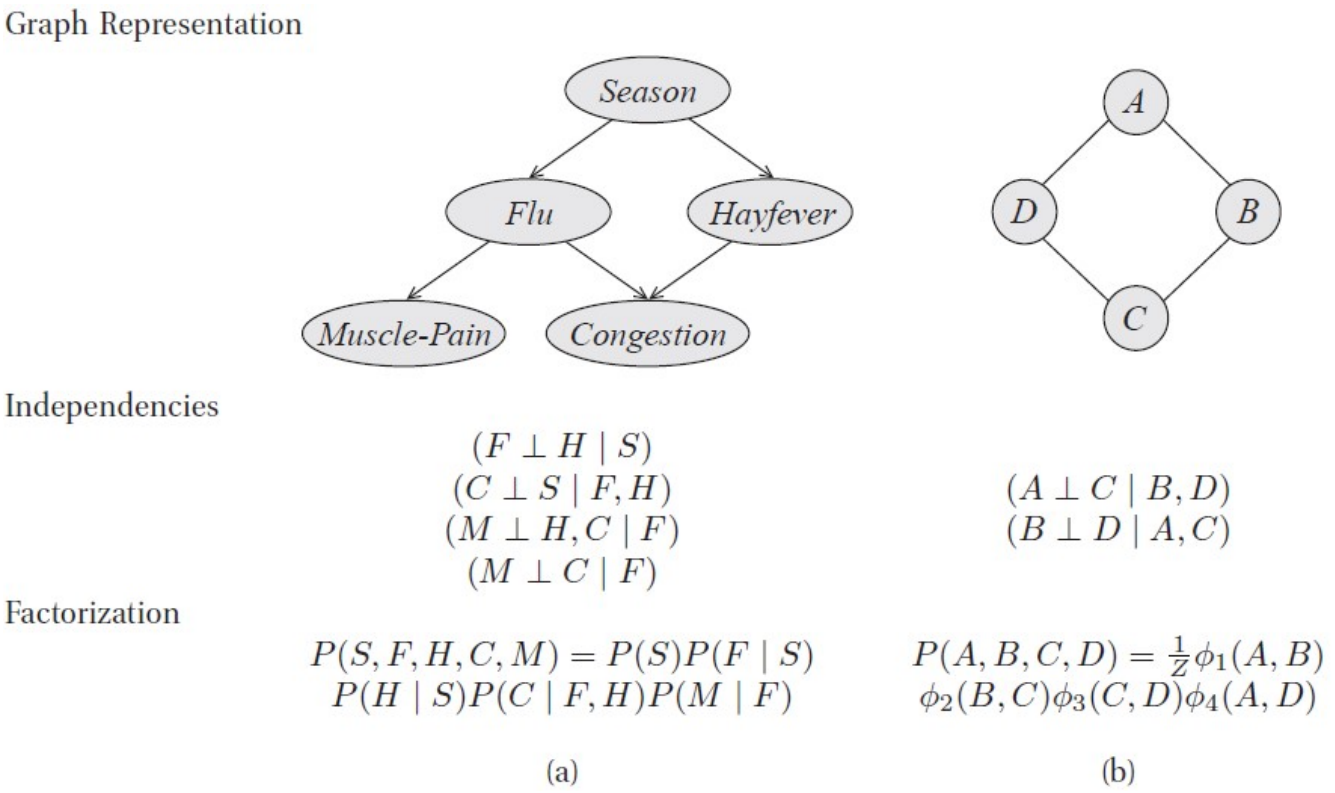

Figure 1: Conditional Independence [Koller and Friedman, 2009]

The figure 1 above shows probabilistic graphical models (directed on the left and undirected on the right), the conditional independence assumption that holds in the model and probability density simplification.

The dependency assumption in a graph can be represented using *d-separation* and *Markov blanket*. Given a set of data, we can test whether this dependency assumption holds or not in the model.

2 Empirical Independence Tests

Empirical Independence property in graphical models is used to observe data relationships. In the context of Machine Learning, this can be used also to uncover biases (checking whether the predictions rely on certain features) and can be opted as a measure for model performance assessment along with other appropriate metrics.

There are several tests that can be used to check whether the independence assumption prevails among variables in the model. The most common ones are such as:

- Constraint-based Methods
- Score-based Methods
- Kernel-based Tests
- Others

The choice of the test mainly depends on the goal and data available, although more than 1 can also be performed at once. These are discussed in depth in the next subsections.

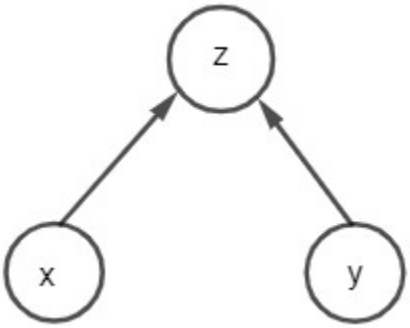

Figure 2: $x \perp y|z$

For simplicity, say we have a model with variables x, y and z we are interested to see whether our model is structured as demonstrated in the figure 2 such that it justifies the conditional independence assumption.

2.1 Constraint-based Methods

In this '*frequentist*' approach, the tests are limited to only a single model at a time, i.e. checking the conditional independence assumption of variables in the model. Since there are no graph-based dependency tests then we employ some appropriate statistical test such as chi-square, correlation, fisher's test. We can use the idea of relative entropy (Kullback-Leibler Divergence as a distribution distance measure) as well. A common test in this case is the Mutual Information as a dependency test which evaluates the common uncertainty of variables.

2.1.1 Mutual Information

Empirical Conditional Independence can be evaluated using Conditional Mutual Information (simply Mutual information, MI). This test uses the concept of relative entropy to find the distance between the joint distribution to the product of their marginal distribution.

According to [Vu, nd], conditional mutual information is a measure of how much uncertainty is shared by x and y, but not by z.

Mathematically,

$$\begin{aligned} MI(x, y|z) &= KL(p(x; y|z)||p(x|z)p(y|z)) \text{ for variables } x, y \text{ and } z \text{ in the model.} \\ &= -\sum_{x,y,x} p(x, y|z) log \frac{p(x|z)p(y|z)}{p(x, y|z)} \text{ for discrete case.} \\ &= -\int\int p(x, y|z) log \frac{p(x|z)p(y|z)}{p(x, y|z)} \text{ for continuous variables.} \end{aligned}$$

If x and y are conditionally independent given z, then $p(x, y|z) = p(x|z).p(y|z)$

Then,

$$\begin{aligned} MI(x, y|z) &= -\sum_{x,y,x} p(x, y|z) log(1) \\ MI(x, y|z) &= 0 \end{aligned}$$

If conditional mutual information is 0 then the parents' variables are independent given their descendant variable meaning that there is no shared uncertainty of the variable given their common effect.

In terms of expectation, the test equation can be expressed as

$$MI(x, y|z) = -E_{p(x,y|z)}\left[log\frac{p(x|z)p(y|z)}{p(x, y|z)}\right]$$

using Jensen's Inequality for Concave functions log(.)

$$= -log\left(E_{p(x,y|z)}\left[\frac{p(x|z)p(y|z)}{p(x, y|z)}\right]\right)$$

Considering a discrete case, as for continuous we can replace $\sum$ with $\int$

$$= -log\left(\sum p(x, y|z)\frac{p(x|z)p(y|z)}{p(x, y|z)}\right)$$

$$= -log\left(\sum p(x|z)p(y|z)\right)$$

$$MI(x, y|z) \in [0, +\infty)$$

This shows that MI is non-negative and the value is positive if the conditional independence assumption doesn't hold true for variables in the model. This is valid theoretically, but in reality, we need a threshold value to justify that the value obtained is truly sufficient as evidence for the conditional independence assumption. We can use a statistical decision rule for hypothesis testing i.e. using a $p-value$ at a certain significance level, α as the probability of false rejection.

We might need a hypothesis for which our test will be based upon and in our case, the formulation is as follows:

$$\begin{aligned} H_o &: p(x, y|z) = p(x|z)p(y|z) \\ H_1 &: p(x, y|z) \neq p(x|z)p(y|z) \end{aligned} \tag{1}$$

Suppose we have discrete-valued, independent random variables X, Y, Z. Typically, we expect that the counts $N[x, y|z]$ in the data are close to $N \cdot \hat{p}(x|z).\hat{p}(y|z)$ (where N is the number of samples). This is the expected value of the count, and, as we know, the deviances from this value are improbable for large N. Based on this intuition, we can measure the deviance of the data from H0 in terms of these distances. A common approach is using X^2 statistic measure of this type.

Recall the X^2 statistic formula for Independence testing gives as follows,

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

whereby, O_i is an observed count and E_i is an expected value.

With probability distribution context, the equation 2 above becomes

$$X^2 = \sum \frac{(N[x, y|z] - N \cdot \hat{p}(x|z).\hat{p}(y|z))^2}{N \cdot \hat{p}(x|z).\hat{p}(y|z)}$$

So, using this measure as the empirical mutual information independence test with the relative probability defined as the ratio of event count to the total count; then,

$$MI^2_X = -\sum_{x,y}\left(\frac{N[x, y|z]}{N} log\frac{\frac{N(x|z)}{N}\cdot\frac{N(y|z)}{N}}{\frac{N[x,y|z]}{N}}\right) \quad \text{[Koller and Friedman, 2009]}$$

This equation is proportional to the log-likelihood ratio; also the degree of freedom associated with it is $(x-1)(y-1)z$; it's preferable to consider chi-square test for discrete data and log-likelihood ratio for continuous.

The decision rule is to reject the 'null hypothesis' H_o if the $p - value < \alpha$ whereby $pvalue = p(MI > threshold|H_o)$

The empirical mutual information as the conditional independence test can be computed based on other measures such as the Monte-Carlo permutation test.

2.2 Score-based Approach

This is another approach to test the conditional independence assumption in the probabilistic graphical model but unlike the above discussed approach; we can compare more than 1 model at once with the 'network score' and select the structure with the highest score. One way to do so is by using the Bayes Factor.

If there is only 1 model then we can restructure it and compare the scores using data likelihood.

2.2.1 Bayesian Conditional Independence Test

In this test, we find the Bayes Factor for the model likelihood ratio associated with the hypothesis that can be formulated as 1 in a Bayesian way.

Suppose we have a graphical model for such hypotheses are as follows:

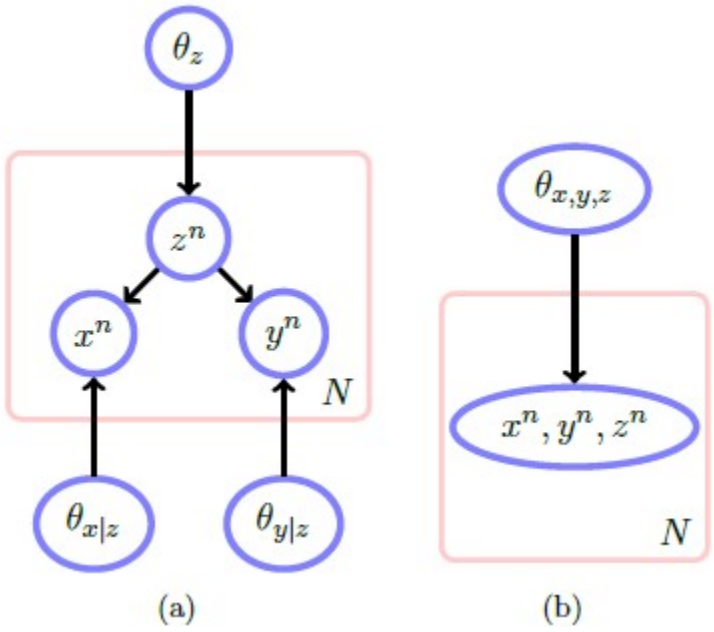

Figure 3: Plate notion for Model under Hypothesis [Barber, 2012]

In figure 3 above, (a) indicates the model under null hypothesis (H_o) while (b) is the model under alternative hypothesis (H_1) with θ_* as the global parameter (ignorimg the states).

The sample model joint distribution is as follows:

$$p(x, y, z, \theta|H_o) = p(x|z, \theta_{x|z})p(y|z, \theta_{y|z})p(z|\theta_z)p(\theta_{x|z})p(\theta_{y|z})p(\theta_z)$$
$$p(x, y, z, \theta|H_1) = p(x, y, z|\theta_{xyz})p(\theta_{xyz})$$

For a set of assumed *i.i.d.* data $\mathcal{D} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = (x^{(n)}, y^{(n)}, z^{(n)})$; $n = 1, \cdots N$, the marginal likelihood (marginalizing over the parameter) is then given by integrating over the parameters θ:

$$p(x, y, z) = \int_\theta p(\theta)p(x, y, z|\theta)$$
$$p(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = \int_\theta p(\theta) \cdot \prod_{n=1}^{N} p(x^{(n)}, y^{(n)}, z^{(n)}|\theta)d\theta$$

Mostly in these graphical model, we use categorical distribution so the conjugate prior suitable in this case is the

Dirichlet distribution for the parameter (with β as hyper-parameter usually treated as 1)

$$\theta \sim Dir(\beta) = p(\theta|\beta)$$
$$p(\theta) \propto \prod_m \theta_m^{\beta_m - 1} \text{ for m categories}$$
$$p(\theta) = \frac{\Gamma(\sum_{m=1}^{M} \beta_m)}{\prod_{m=1}^{M} \Gamma(\beta_m)} \prod_m \theta_m^{\beta_m - 1}$$
$$p(\theta) = \frac{((\sum_{m=1}^{M} \beta_m) - 1)!}{\prod_{m=1}^{M} (\beta_m - 1)!} \prod_m \theta_m^{\beta_m - 1}$$

. So,

$$p(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = \frac{\Gamma(\sum_{m=1}^{M} \beta_m)}{\prod_{m=1}^{M} \Gamma(\beta_m)} \int_\theta \prod_m \theta_m^{\beta_m - 1} \cdot \prod_{n=1}^{N} \theta_m^{\#_m^v} d\theta \quad ; \text{with } v = \{x, y\}$$
$$p(\mathcal{D}) = \frac{\Gamma(\sum_{m=1}^{M} \beta_m)}{\prod_{m=1}^{M} \Gamma(\beta_m)} \int_\theta \prod_m \theta_m^{\#_m^v + \beta_m - 1} d\theta$$

Simply, $p(\mathcal{D}) = \frac{f(\#_m^v + \beta_m)}{f(\beta_m)}$ with $f(\beta_m) = \frac{\prod_{m=1}^{M} \Gamma(\beta_m)}{\Gamma(\sum_{m=1}^{M} \beta_m)}$

For a belief/bayesian network,

$$p(\mathcal{D}) = \prod_k \prod_n p(x_k^n, y_k^n | z_k^n) = \prod_k \prod_j \frac{f(\beta^{new}(x_k, y_k; j))}{f(\beta(x_k, y_k; j))} = \prod_k \prod_n p(x_k^n, y_k^n | z_k^n) = \prod_k \prod_j \frac{\Gamma(\sum_m \beta_m(x_k, y_k; j))}{\Gamma(\sum_m \beta_m^{new}(x_k, y_k; j))} \prod_m \frac{\Gamma(\beta_m^{new}(x_k, y_k; j))}{\Gamma(\beta_m(x_k, y_k; j))}$$

For, $j \in [1, \text{no. of parent states}]$ and the updated hyperparameter over the observed counts $\beta_m^{new} = \beta_m + \#_m^v$ is the number of times that variable v is in state m and $\#(v, z)$ is the number of times v and z are in certain states simultaneously.

$\beta_{v|pa(v)}$ is the hyperparameter matrix of pseudo counts for each state of v given each state of its parent$z = pa(v)$.

In $H_0 \& H_1, \theta = \theta_{x|z}, \theta_{y|z}, \theta_z$ and $\theta = \theta_{xyz}$

Then respective equations for H_0 and H_1 are as follows ,

$$p(\mathcal{D}) = \frac{f(\#(z) + \beta_z)}{f(\beta_z)} \prod_z \frac{f(\#(x, z) + \beta_{x|z})}{f(\beta_{x|z})} \frac{f(\#(y, z) + \beta_{y|z})}{f(\beta_{y|z})} \tag{3}$$

$$p(\mathcal{D}) = \frac{f(\#(x, y, z) + \beta_{x,y,z})}{f(\beta_{x,y,z})} \tag{4}$$

Assuming that all hypotheses 1 are equally likely, then the Bayes Factor (BF) as the ratio of equation 3 to 4

If the value is greater than unity, then the test favours the Independence hypothesis, H_0; the following table is thorough conclusions for the BF as per Jeffrey for a null over alternative ratio:

Bayes factor	Interpretation
> 100	Extreme evidence for H_0
30 – 100	Very strong evidence for H_0
10 – 30	Strong evidence for H_0
3 – 10	Moderate evidence for H_0
1 – 3	Anecdotal evidence for H_0
1	No evidence
1/3 – 1	Anecdotal evidence for H_1
1/10 – 1/3	Moderate evidence for H_1
1/30 – 1/10	Strong evidence for H_1
1/100 – 1/30	Very strong evidence for H_1
< 1/100	Extreme evidence for H_1

Another score used is a BIC (Bayesian Information Criterion) score which is similar to 'Minimum Description Length' (MDL)

Mathematically, $BIC_{score} = N(\sum_{i=1}^{n} MI_{p(v,pa(v))}(v_i, pa(v_i)) - \sum_{i=1}^{n} H_p(v, pa(v))(v_i) - \frac{logN}{2} dim(\mathcal{G})$ whereby, $H \sim Entropy$, $dim(\mathcal{G})$ is the degree of freedom or the number of independent model parameters.

According to [Koller and Friedman, 2009], the score trade-off model complexity with data reducing the chances for overfitting; as for simple structures it could be biased to simpler structures. Also, there's an 'Extended BIC' for dense structures. This score can be effective even for a model with latent variable(s).

It is consistent and approximates the Bayes Factor i.e. $2logBF_{H_0H_1} \approx BIC_{H_0} - BIC_{H_1}$.

With BIC, we choose the model with the lowest value, i.e. Model 0 is selected over 1 iff $BIC_0 - BIC_1 \in (-\infty, 0)$.

Unlike conditional independence tests, network scores focus on the (Directed Acyclic Graph) DAG as a whole; they are goodness-of-fit statistic measuring how well the DAG mirrors the dependence structure of the data.

2.3 Kernel-based Tests

For the above parametric & semi-parametric tests discussed, we might need a nonparametric approach where the distribution is learnt from the data.

One approach is using Hilbert-Schmidt Independence Criterion (HSIC) which also captures non-linearity. It uses the notion of Maximum Mean Discrepancy (MMD)[l1,].

$$
\begin{aligned}
MMD(p,q) &= ||E_a \sim p\phi(a) - E_b \sim q\phi(b)|| \text{ with } \phi \text{ as feature maps of a specific kernel} \\
HSIC(x,y|z) &= MMD(p(x,y|z), p(x|z)p(y|z)) \\
HSIC(x,y|z) &= ||E_{x,y|z} \sim p(x,y|z)\phi(x,y|z) - E_{x|z} \sim p(x|z)\phi(x|z) \cdot E_{y|z} \sim p(x|z)\phi(y|z)||_{\mathcal{H}_\parallel}.
\end{aligned}
$$

whereby $\mathcal{H}_\parallel$ is the Hilbert-Schmidt norm for the k^{th} kernel.

$HSIC(x,y|z) = 0$ iff $x \perp y|z$.

3 Example

Suppose we have a simple dataset,

Asbestos	1	1	0	0	1	0	1
Smoking	1	0	1	1	1	0	0
Cancer	1	0	1	0	1	0	1

Using hypotheses stated above 1 and using R package, the following results were obtained:

```
        Mutual Information (disc.)

data:  asbestos ~ smok | cancer_l
mi = 1.7261, df = 2, p-value = 0.4219
alternative hypothesis: true value is greater than 0


        Mutual Information (disc.)

data:  smok ~ asbestos | cancer_l
mi = 1.7261, df = 2, p-value = 0.4219
alternative hypothesis: true value is greater than 0


        Pearson's X^2

data:  asbestos ~ smok | cancer_l
x2 = 1.1944, df = 2, p-value = 0.5503
alternative hypothesis: true value is greater than 0
```

By default, $\alpha = 0.05$; we see that p-value¿0.05 $\implies x \perp y|z$ for the Empirical Mutual Information Conditional Independence tests.

The following plot was used for the hypotheses

The figure on the left represent model with conditional independence assumption while that on the right doesn't.
On the Bayesian approach the following scores were computed;

```
{r BF & BIC}
##BF
model_indep=model2network('[asbestos][smok][cancer_l|smok:asbestos]')#H_0
#Model=model2network('[cancer_l][smok|cancer_l][asbestos |cancer_l]')6params
Model_dep=model2network('[cancer_l][smok|cancer_l][asbestos]')#H1
BF(Model_dep,model_indep,data=data)
1/BF(Model_dep,model_indep,data=data)
score(Model_dep, data, type='bic') - score(model_indep, data, type='bic')
#for log-likelihood ratio
score(Model_dep, data, type='loglik')-score(model_indep, data, type='loglik')
```

```
[1] 0.685179
[1] 1.459473
[1] 0.5596158
[1] -1.386294
```

the BF and BIC results suggests considering the H_0

4 Conclusion

Determining the conditional independence of variables in a graphical model is crucial as one of the steps in structure learning. In ML we use these models in different contexts such as Reinforcement Learning, Computer vision and graphics, Bioinformatics and Computational Biology, Genetics and medical diagnosis/prognosis, Natural language processing etc. It is important though to obtain a more graph-based approach that would best work to determine the dependency assumption in the model than using statistical approaches.

References

[l1,] Probabilistic graphical models introduction to gm.

[Barber, 2012] Barber, D. (2012). Bayesian reasoning and machine learning.

[Koller and Friedman, 2009] Koller, D. and Friedman, N. (2009). Probabilistic graphical models: Principles and techniques.

[Murphy, 2023] Murphy, K. P. (2023). *Probabilistic Machine Learning : Advanced Topics*. MIT Press.

[Vu, nd] Vu, M. (n.d). Lecture 1: Entropy and mutual information.