1.1 Conditioning on background evidence

a) Given a background evidence E, show that $P(X,Y|E) = P(X|Y,E)P(Y|E)$

If we create the variable $Z = X,Y$ and use Bayes rule, we get

$P(Z|E) = \frac{P(E|Z)P(Z)}{P(E)}$. If we use the Product rule on the nominator, we get

$P(Z|E) = \frac{P(Z,E)}{P(E)}$. We change back from Z to X,Y to get

$$P(X,Y|E) = \frac{P(X,Y,E)}{P(E)} \overset{\text{chain rule}}{=} \frac{P(X|Y,E)P(Y,E)}{P(E)} \overset{\text{Product rule}}{=} P(X|Y,E)\frac{\cancel{P(E)}P(Y|E)}{\cancel{P(E)}} = P(X|Y,E)P(Y|E)$$ QED.

b) Prove the conditionalized version of Bayes rule $P(X|Y,E) = \frac{P(Y|X,E)P(X|E)}{P(Y|E)}$.

From a) we have that

$P(X|Y,E) = \frac{P(X,Y|E)}{P(Y|E)}$. We take the nominator and write it as

$$P(X,Y|E) = \frac{P(X,Y,E)}{P(E)} \overset{\text{chain rule}}{=} \frac{P(Y|X,E)P(X,E)}{P(E)} \overset{\text{product rule}}{=} P(Y|X,E)\frac{\cancel{P(E)}P(X|E)}{\cancel{P(E)}} = P(Y|X,E)P(X|E).$$ (Product rule)

If we insert this, we get

$$P(X|Y,E) = \frac{P(X,Y|E)}{P(Y|E)} = \frac{P(Y|X,E)P(X|E)}{P(Y|E)}$$ QED.

c) Prove the conditionalized version of marginalization $P(X|E) = \sum_y P(X,Y=y|E)$.

$$P(X|E) \overset{\text{product rule}}{=} \frac{P(X,E)}{P(E)}$$

$$\overset{\text{marginalization}}{=} \frac{\sum_y P(X,Y=y,E)}{P(E)}$$

$$\overset{\text{product rule}}{=} \frac{\sum_y P(X,Y=y|E)\cancel{P(E)}}{\cancel{P(E)}}$$

$$= \sum_y P(X,Y=y|E)$$ QED

1.2 Conditional independence

We have the three following statements about random variables X, Y, E :

(1) $P(X,Y|E) = P(X|E)P(Y|E)$

(2) $P(X|Y,E) = P(X|E)$

(3) $P(Y|X,E) = P(Y|E)$

Show that the three statements are equivalent, i.e. (1) implies (2) & (3), (2) implies (1) & (3) and (3) implies (1) & (2).

We start with showing that (1) implies (2) and (3).

$P(X,Y|E) = P(X|E)P(Y|E) \qquad (1)$

We use what we proved in 1.1a) on the left hand side to get (No need of proof again as it has been done already)

$P(X|Y,E)\cancel{P(Y|E)} = P(X|E)\cancel{P(Y|E)}$

$\Rightarrow P(X|E) = P(X|Y,E)$ which proves (1) impies (2). QED

(1)⇒(2)

Likewise, we could have written $P(X,Y|E)$ as

$$P(X,Y|E) = \frac{P(X,Y,E)}{P(E)} \overset{\text{Chain rule}}{=} \frac{P(Y|X,E)P(X,E)}{P(E)} \overset{\text{Product rule}}{=} P(Y|X,E)\frac{\cancel{P(E)}P(X|E)}{\cancel{P(E)}} = P(Y|X,E)\cancel{P(X|E)} \overset{(1)}{=} \cancel{P(X|E)}P(Y|E)$$

Thus, we have $P(Y|E) = P(Y|X,E)$, which proves (1) implies (3). QED

(1)⇒(3)

Hence, we have proven that (1) implies (2) & (3). QED.

We now want to show that (2) implies (1) and (3).

$P(X|Y,E) = P(X|E) \qquad (2)$

We use the conditionalized Bayes rule on the left hand side (derived in 1.1b) to get

$$\frac{P(Y|X,E)P(X|E)}{P(Y|E)} = P(X|E)$$

$\Rightarrow P(Y|E) = P(Y|X,E)$ which proves (2) implies (1). QED

(2)⇒(1)

If we use the conditionalized product rule on the right hand side of (2) for P(X|E), we get

$$P(X|Y,E) = \frac{P(X,Y|E)}{P(Y|X,E)} \Leftrightarrow P(X|Y,E)P(Y|X,E) = P(X,Y|E)$$

If we use the conditionalized product rule on the left hand side, we get

$$\frac{P(X,Y|E)}{P(Y|E)} \cdot \frac{\cancel{P(X,Y|E)}}{P(X|E)} = \cancel{P(X,Y|E)}$$

$\Rightarrow P(X,Y|E) = P(Y|E)P(X|E)$ which proves (2)⇒(1). Thus, (2) ⇒ (1) & (3). QED.

(2)⇒(3)

We now want to show that (3) implies (1) and (2).

$$P(Y|X,E) = P(Y|E) \qquad (3)$$

By the same way as before, we deduce

$$P(Y|X,E) = \frac{P(X,Y|E)}{P(X|Y,E)} \Leftrightarrow P(X|Y,E)P(Y|X,E) = P(X,Y|E)$$

If we use the conditionalized product rule on the left hand side, we get

$$\frac{P(X,Y|E)}{P(Y|E)} \cdot \frac{\cancel{P(X,Y|E)}}{P(X|E)} = \cancel{P(X,Y|E)}$$

$\Rightarrow P(X,Y|E) = P(Y|E)P(X|E)$ which proves $(3) \Rightarrow (1)$. QED.

$(3) \Rightarrow (1)$

To show $(3) \Rightarrow (2)$, we use unconditionalized Bayes rule on the left side of (3)

$$\frac{P(X|Y,E)\cancel{P(Y|E)}}{P(X|E)} = \cancel{P(Y|E)}$$

$(3) \Rightarrow (2)$

$\Rightarrow P(X|Y,E) = P(X|E)$ which proves $(3) \Rightarrow (2)$ and thus $(3) \Rightarrow (1)$ & (2). QED.

Thus, we have proven that the statements are equivalent.

1.3 Creative writing

Attach events to the binary events X, Y, Z that are consistent with the following patterns of commonsense reasoning.

a) Cumulative evidence : $P(X) < P(X|Y) < P(X|Y,Z)$

We let X, Y and Z be as follows:

X: A plane crashes

Y: The pilots are intoxicated

Z: The plane has technical issues

With events Y and Z, the probability of event X will increase drastically.

b) Explaining away: $P(X|Y) > P(X)$ but $P(X|Y,Z) < P(X|Y)$

We let X, Y and Z be as follows:

X: The plane crashes during landing

Y: The pilot is intoxicated

Z: The plane's autopilot is active

An autopilot can land a plane without a lot of assistance, so it will decrease the crash probability.

c) Conditional independent $P(X,Y) \neq P(X)P(Y)$ but $P(X,Y|Z) = P(X|Z)P(Y|Z)$

We imagine that we have two dice, where one of them is a regular one, and the other has sixes on all sides, and we pick one at random and roll it twice.

We let X, Y and Z be as follows:

X: First throw results in a six.

Y: Second throw results in a six.

Z: We know that we have chosen the first die.

Explanation: If we roll the first six, there's a better chance for the second throw also being one as there are more total sixes across the two dice than other numbers. (Dependent)

If we know that the first die is chosen, the probability is uniform, so the two throws become independent.

1.4 Bayes Rule

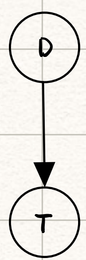
1% of cyclists are using performance-enhancing drugs

A drug test has 5% false positive rate and 10% false negative rate:

$D \in \{0,1\}$ indicates whether a cyclist is doping, and

$T \in \{0,1\}$ indicates the outcome of the test.

a) Draw the belief network for the random variables D and T, and deduce probability tables for $P(D)$ and $P(T|D)$.



	prob
D = 0	0.99
D = 1	0.01

		prob
D = 0	T = 0	0.95
D = 0	T = 1	0.05
D = 1	T = 0	0.1
D = 1	T = 1	0.9

These come from the info about false positives and negatives

b) Cyclist A tests negative for drug use. What is the probability that Cyclist A is not using drugs?

What we are looking for is $P(D=0 \mid T=0)$.

We use the product rule to get

$$P(D=0 \mid T=0) \overset{\text{P.R.}}{=} \frac{P(T=0, D=0)}{P(T=0)}$$

$$\overset{\text{P.R.}}{=} \frac{P(T=0 \mid D=0) \cdot P(D=0)}{P(T=0)}$$

$$\overset{\text{marginalization}}{=} \frac{P(T=0 \mid D=0) \cdot P(D=0)}{\sum_d P(T=0, D=d)}$$

$$= \frac{P(T=0 \mid D=0) \cdot P(D=0)}{P(T=0, D=0) + P(T=0, D=1)}$$

$$\overset{\text{P.R.}}{=} \frac{P(T=0 \mid D=0) \cdot P(D=0)}{P(D=0)P(T=0 \mid D=0) + P(D=1)P(T=0 \mid D=1)}$$

$$\overset{\text{tables}}{=} \frac{0.95 \cdot 0.99}{0.99 \cdot 0.95 + 0.01 \cdot 0.1}$$

$$P(D=0 \mid T=0) = 0.99894$$

⇒ The probability that Cyclist A is not using drugs is approx. 99.9%.

c) Cyclist B tests positive for drug use. What is the probability that Cyclist B is using drugs?

We are looking for $P(D=1|T=1)$.

$$P(D=1|T=1) \overset{\text{P.R.}}{=} \frac{P(T=1, D=1)}{P(T=1)}$$

$$\overset{\text{P.R.}}{=} \frac{P(T=1|D=1) \cdot P(D=1)}{P(T=1)}$$

$$\overset{\text{marginalization}}{=} \frac{P(T=1|D=1) \cdot P(D=1)}{\sum_d P(T=1, D=d)}$$

$$= \frac{P(T=1|D=1) \cdot P(D=1)}{P(T=1, D=0) + P(T=1, D=1)}$$

$$\overset{\text{P.R.}}{=} \frac{P(T=1|D=1) \cdot P(D=1)}{P(D=0)P(T=1|D=0) + P(D=1)P(T=1|D=1)}$$

$$\overset{\substack{\text{tables} \\ \text{in a)}}}{=} \frac{0.9 \cdot 0.01}{0.99 \cdot 0.05 + 0.01 \cdot 0.9}$$

$$P(D=1|T=1) = 0.15385$$

⇒ The probability that Cyclist B is using drugs is approx. 15.4 %.

1.5 Entropy

Let X be a discrete random variable with $P(X = x_i) = p_i$ for $i \in \{1, 2, \dots, n\}$.

The entropy $\mathcal{H}[X]$ of the random variable X is a measure of its uncertainty.

It is defined as $\mathcal{H}[X] = -\sum_{i=1}^{n} p_i \log p_i$.

a) Show that $\mathcal{H}[X]$ is maximized when $p_i = \frac{1}{n}$ $\forall i$.

To find the gradient, we take the partial derivative of H for each element j in the gradient vectors.

We could have written out the gradient vectors, but we look at one and one of the elements.

Lagrange gives us $\nabla H = \lambda \nabla g$ where g is our constraint $\sum_{i=1}^{n} p_i$.

We look at H_j and g_j where $g_j = p_j$ and the sum of all p_j's is 1 as in the task.

$$\frac{\partial \mathcal{H}_j}{\partial p_j} \overset{\text{Lagrange}}{=} \lambda \frac{\partial g_j}{\partial p_j} = -\frac{\partial}{\partial p_j} p_j \log p_j$$

$$= -(1 + \log p_j) = \lambda \cdot \frac{\partial}{\partial p_j} p_j = \lambda \cdot 1$$

Thus, we now have

$$\log p_j = -(1 + \lambda)$$

$$\Rightarrow p_j = e^{-(1+\lambda)}$$

By summing over all p_j's, we get

$$\sum_{j=1}^{n} p_j = n e^{-(1+\lambda)} \overset{\text{constraint}}{=} 1$$

$$\Rightarrow e^{-(1+\lambda)} = \frac{1}{n}$$

But we already know that $p_j = e^{-(1+\lambda)}$, thus we get the maximized entropy for

$$e^{-(1+\lambda)} = p_j = \frac{1}{n} \quad \forall j \quad \text{QED.}$$

The joint entropy of n discrete random variables $(X_1, X_2, \dots, X_n)$ is defined as

$H(X_1, X_2, \dots, X_n) = -\sum_{x_1}\sum_{x_2}\dots\sum_{x_n} P(x_1, x_2, \dots, x_n) \log P(x_1, x_2, \dots, x_n)$ where the sums range over all possible instantiations of $(X_1, X_2, \dots, X_n)$

b) Show that $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P(x_i)$ implies $H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i)$

We know from a) that we have $\sum_{i=1}^{n} P(x_i) = 1$ and that $H(X_i) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$

$$H(X_1, X_2, \dots, X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} \underbrace{P(x_1, x_2, \dots, x_n)}_{\prod_{i=1}^{n} P(x_i)} \log \underbrace{P(x_1, x_2, \dots, x_n)}_{\prod_{i=1}^{n} P(x_i)} = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} \prod_{i=1}^{n} P(x_i) \log \prod_{i=1}^{n} P(x_i)$$

log rule
$\log(a \cdot b) = \log a + \log b$

$$= -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} \prod_{i=1}^{n} P(x_i) \sum_{i=1}^{n} \log P(x_i)$$

$$= -\sum_{x_1} \log P(x_1) \cdot \prod_{i=1}^{n} P(x_i) - \sum_{x_2} \log P(x_2) \prod_{i=1}^{n} P(x_i) - \dots - \sum_{x_n} \log P(x_n) \prod_{i=1}^{n} P(x_i)$$

$$= -\sum_{x_1} \log P(x_1) \cdot P(x_1) \cdot \overbrace{\sum_{x_2} P(x_2)}^{=1} \cdot \ldots \cdot \overbrace{\sum_{x_n} P(x_n)}^{=1} - \sum_{x_2} \log P(x_2) \cdot P(x_2) \cdot \overbrace{\sum_{x_1} P(x_1)}^{=1} \cdot \overbrace{\sum_{x_3} P(x_3)}^{=1} \cdot \ldots \cdot \overbrace{\sum_{x_n} P(x_n)}^{=1} - \dots - \sum_{x_n} \log P(x_n) \cdot P(x_n) \cdot \overbrace{\sum_{x_1} P(x_1)}^{=1} \cdot \overbrace{\sum_{x_2} P(x_2)}^{=1} \cdot \ldots \cdot \overbrace{\sum_{x_{n-1}} P(x_{n-1})}^{=1}$$

$$= -\sum_{x_1} \log P(x_1) \cdot P(x_1) - \sum_{x_2} \log P(x_2) \cdot P(x_2) - \dots - \sum_{x_n} \log P(x_n) \cdot P(x_n)$$

$$= \mathcal{H}[X_1] + \mathcal{H}[X_2] + \dots + \mathcal{H}[X_n]$$

$$= \sum_{i=1}^{n} \mathcal{H}[X_i] \qquad \text{QED}$$

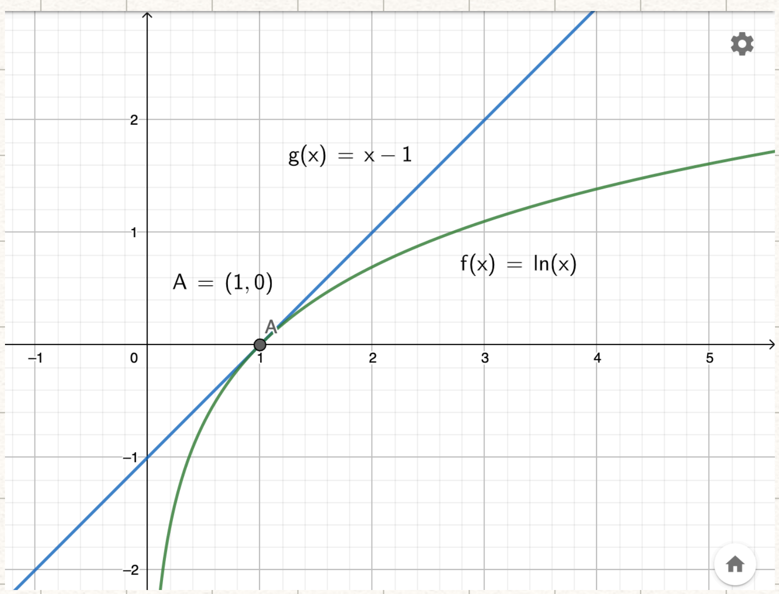
1.6 Kullback-Leibler distance

We let $p_i = P(X=i|E)$ and $q_i = P(X=i|E')$ denote the conditional distributions over the RV X for different evidences $E \neq E'$

We note that $\sum_i p_i = \sum_i q_i = 1$.

The KL distance between these distributions is defined as $KL(p,q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$

a) Sketch the graphs of $\log(x)$ and $x-1$ to verify $\log(x) \leq x-1$ with equality iff. $x=1$. Confirm by differentiation of $\log(x)-(x-1)$



We observe that the only intersection is at point A where $x=1$. We also observe that $\log(x)$ never goes over $x-1$.

We now differentiate

$\frac{d}{dx}(\log(x)-(x-1)) = \frac{1}{x} - 1$ with $x \in (0^+, \infty)$

We set it equal to zero to find the function's extrema.

$\frac{1}{x} - 1 = 0 \Rightarrow x = 1$

For $x \in (0,1)$, the function $\log(x)-(x-1)$ is a growing function, while it's a decreasing function for $x \in (1, \infty)$.

Thus, $x=1$ maximizes the function, making the max value $\log(1)-(1-1) = 0$.

$\Rightarrow \log(x) \leq x-1$ QED

b) Use the previous result to prove that $KL(p,q) \geq 0$, with equality only if p_i and q_i are equal.

We have

To simplify later

$$KL(p,q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) = -\sum_i p_i \log\left(\frac{q_i}{p_i}\right) \qquad \left(\log\left(\frac{a}{b}\right) = -\log\left(\frac{b}{a}\right)\right)$$

From a) we have that $\log(x) \leq x-1$.

If we let $x = \frac{q_i}{p_i}$, we have

$$KL(p,q) = -\sum_i p_i \log\left(\frac{q_i}{p_i}\right) \geq -\sum_i p_i \left(\frac{q_i}{p_i} - 1\right) = -\sum_i (q_i - p_i)$$

$$= -\left(\underbrace{\sum_i q_i}_{=1} - \underbrace{\sum_i p_i}_{=1}\right)$$

$$= 0$$

Thus, we have proven that $KL(p,q) \geq 0$ QED

We have the inequality

$$-\sum_i p_i \underbrace{\log\left(\frac{q_i}{p_i}\right)} \geq -\sum_i p_i \underbrace{\left(\frac{q_i}{p_i} - 1\right)}$$

From a) we know that $\log(x) = x-1$ only for $x=1$.

Thus, we need our $x = \frac{p_i}{q_i} = 1$ to have an equality in this case, as the rest of the functions are identical.

Thus, $\frac{p_i}{q_i} = 1$ gives us an equality only when $p_i = q_i \ \forall i$. QED.

c) Using the inequality in a), as well as $\log x = 2\log\sqrt{x}$, derive the tighter lower bound $KL(p,q) \geq \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$.

We have $\log(x) \leq x-1$ and $\log x = 2\log\sqrt{x}$.

$$KL(p,q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) = -\sum_i p_i \log\left(\frac{q_i}{p_i}\right)$$

given equality

$$= -\sum_i 2p_i \log\left(\sqrt{\frac{q_i}{p_i}}\right) \geq -\sum_i 2p_i \left(\frac{\sqrt{q_i}}{\sqrt{p_i}} - 1\right)$$

$$= -\sum_i 2\left(\sqrt{p_i}\sqrt{q_i} - p_i\right)$$

$$= \sum_i \left(2p_i - 2\sqrt{p_i}\sqrt{q_i}\right)$$

$$= 2\sum_i p_i - \sum_i 2\sqrt{p_i}\sqrt{q_i}$$

Because we know that $\sum_i p_i = \sum_i q_i$, and that we have $2\sum_i p_i$, we can write this as

square formula

$$KL(p,q) \geq \sum_i p_i - \sum_i 2\sqrt{p_i}\sqrt{q_i} + \sum_i q_i = \sum_i \left(p_i - 2\sqrt{p_i}\sqrt{q_i} + q_i\right) = \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \qquad \text{QED}$$

d) Provide a counterexample to show that the KL distance is not symmetric in function of its arguments: $KL(p,q) \neq KL(q,p)$.

To find a counterexample, we would need sets of probabilities for p and q.

If the set for p is $X_1 \sim \{0,1\}$ and $P(X_1=0) = 0.8$ & $P(X_1=1) = 0.2$, and

if the set for q is $X_2 \sim \{0,1\}$ and $P(X_2=0) = 0.4$ & $P(X_2=1) = 0.6$, we have

$$KL(p,q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) = P(X_1=0)\cdot\log\left(\frac{P(X_1=0)}{P(X_2=0)}\right) + P(X_1=1)\cdot\log\left(\frac{P(X_1=1)}{P(X_2=1)}\right)$$

$$= 0.8\cdot\log\left(\frac{0.8}{0.4}\right) + 0.2\log\left(\frac{0.2}{0.6}\right)$$

$$= \underline{0.3348}$$

$$KL(q,p) = \sum_i q_i \log\left(\frac{q_i}{p_i}\right) = P(X_2=0)\cdot\log\left(\frac{P(X_2=0)}{P(X_1=0)}\right) + P(X_2=1)\cdot\log\left(\frac{P(X_2=1)}{P(X_1=1)}\right)$$

$$= 0.4\cdot\log\left(\frac{0.4}{0.8}\right) + 0.6\log\left(\frac{0.6}{0.2}\right)$$

$$= \underline{0.3819}$$

As we didn't get the same result, we have found a counterexample to show $KL(p,q) \neq KL(q,p)$ QED.

1.7 Mutual information

The mutual information $I(X,Y)$ between two discrete random variables X and Y is defined as

$$I(X,Y) = \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right],$$ where the sum is over all possible values of the RV's X and Y.

a) Show that the mutual information is nonnegative. $(I(X,Y) \geq 0)$

$$I(X,Y) = \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right]$$

We use the same method as in the previous problem to cancel out one of the $P(X,Y)$ terms later

$$I(X,Y) = \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right] = -\sum_x \sum_y P(x,y) \log\left[\frac{P(x)P(y)}{P(x,y)}\right]$$

$$\geq -\sum_x \sum_y P(x,y)\left[\frac{P(x)P(y)}{P(x,y)} - 1\right] = -\sum_x \sum_y \left[P(x)P(y) - P(x,y)\right] = \sum_x \sum_y P(x,y) - \sum_x P(x) \sum_y P(y)$$

We use marginalization $\sum_x \sum_y P(x,y) = \sum_x P(x)$ and the fact that $\sum_y P(y) = 1$ to get

$$I(X,Y) \geq \sum_x \sum_y P(x,y) - \sum_x P(x) \overbrace{\sum_y P(y)}^{=1} \overset{\text{marginalization}}{=} \sum_x P(x) - \sum_x P(x) = 0 \qquad \text{QED}$$

b) Show that the mutual information $I(X,Y)$ vanishes iff. X and Y are independent random variables.

We need to show that $I(X,Y) = 0$ implies that X and Y are independent, and that X and Y independent implies $I(X,Y) = 0$. We start with the latter.

X and Y independent means that $P(X,Y) = P(X)P(Y)$. Thus, we have

$$I(X,Y) = \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right] = \sum_x \sum_y P(x,y) \log\left[\frac{P(x)P(y)}{P(x)P(y)}\right] = \sum_x \sum_y P(x,y) \log(1) = 0 \qquad \text{QED}$$

For the first thing we wanted to prove, we need to look at $I(X,Y) = 0$.

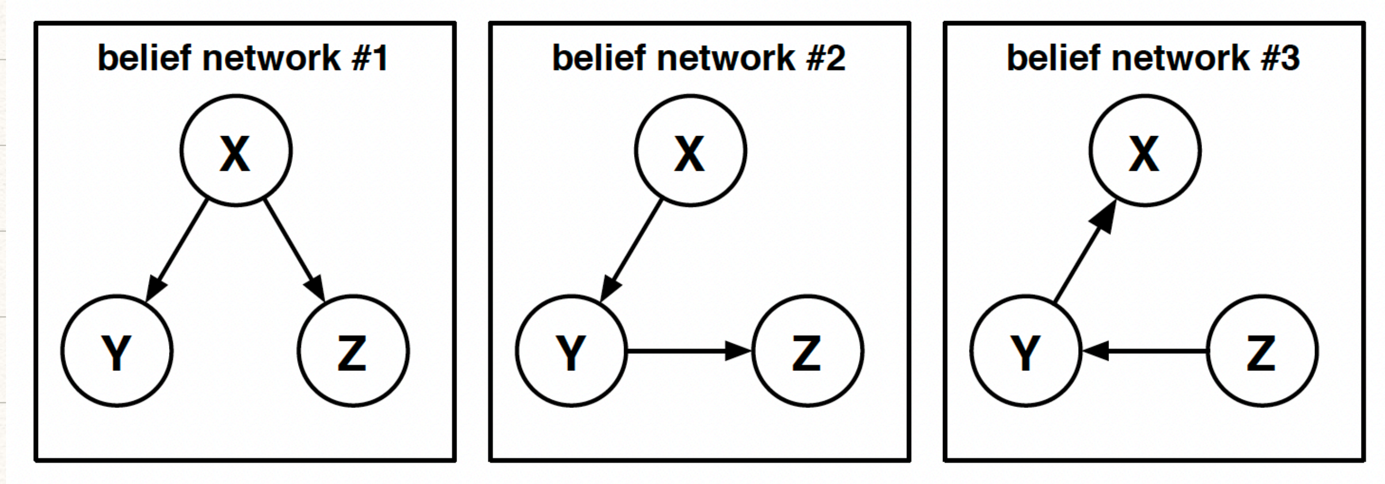
$$I(X,Y) = \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right] = 0$$

The only way this equality holds is if $\log\left[\frac{P(x,y)}{P(x)P(y)}\right] = 0$. We know that $\log(1) = 0$, so we need

$$\frac{P(x,y)}{P(x)P(y)} = 1 \Leftrightarrow P(X,Y) = P(X)P(Y) \Rightarrow X \text{ and } Y \text{ are independent.} \qquad \text{QED}$$

1.8 Compare and contrast

We consider three belief networks (BNs) shown below for the discrete RVs X, Y and Z.



a) Does BN1 imply a statement of marginal or conditional independence that is not implied by BN2?

Yes, BN1 shows that Y and Z are conditionally independent given X.

$P(X,Y,Z) = P(Z)P(Y|X)P(Z|X) \Rightarrow P(Y,Z|X) = P(Y|X)P(Z|X)$ is implied in BN1 but not in BN2.

b) Does BN2 imply a statement of marginal or conditional independence that is not implied by BN3?

No, they convey the same information regarding marginal or conditional independence between the three variables

c) Does BN3 imply a statement of marginal or conditional independence that is not implied by BN1?

$P(X,Y,Z) = P(Z)P(Y|Z)P(X|Y) \Rightarrow P(X,Z|Y) = \frac{P(Z)}{P(Y)} P(Y|Z)P(X|Y) \overset{\text{Bayes}}{=} P(Z|Y)P(X|Y)$ which shows Z and X are conditional independent given Y. We know that BN1 is not conveying the same.

1.9 Hangman

a) From the counts in the file compute the prior probability $P(\omega) = \frac{\text{COUNT}(\omega)}{\sum_{\omega'} \text{COUNT}(\omega')}$.

Print out the 15 most and 14 least frequent words. Sorted by prior prob with sort_values function from Pandas.

```python
# Read lines from file and save as pandas dataframe. Separate lines into two columns.
data = pd.read_csv('hw1_word_counts_05.txt', sep=" ", header=None, names=['Word', 'Count'])

# Compute prior probability and add as third column of dataframe for each word
data['P(W=w)'] = data['Count']*1.0/data['Count'].sum()

# Sort the dataframe by the word's prior probability (Second column of dataframe)
data_sorted = data.sort_values(by = ['P(W=w)'], ascending=False)

# Sanity check. 15 most frequent and 14 least frequent words.
print(data_sorted.head(15))
print(data_sorted.tail(14))
```

	0	1	2
5821	THREE	273077	0.035627
5102	SEVEN	178842	0.023333
1684	EIGHT	165764	0.021626
6403	WOULD	159875	0.020858
18	ABOUT	157448	0.020542
5804	THEIR	145434	0.018974
6320	WHICH	142146	0.018545
73	AFTER	110102	0.014365
1975	FIRST	109957	0.014346
1947	FIFTY	106869	0.013943
4158	OTHER	106052	0.013836
2073	FORTY	94951	0.012388
6457	YEARS	88900	0.011598
5806	THERE	86502	0.011286
5250	SIXTY	73086	0.009535

15 most frequent words

	0	1	2
5093	SERNA	7	9.132590e-07
5872	TOCOR	7	9.132590e-07
3978	NIAID	7	9.132590e-07
1842	FABRI	7	9.132590e-07
4266	PAXON	7	9.132590e-07
2041	FOAMY	7	9.132590e-07
6443	YALOM	7	9.132590e-07
977	CCAIR	7	9.132590e-07
1107	CLEFT	7	9.132590e-07
3554	MAPCO	6	7.827935e-07
895	CAIXA	6	7.827935e-07
4160	OTTIS	6	7.827935e-07
5985	TROUP	6	7.827935e-07
712	BOSAK	6	7.827935e-07

14 least frequent words (bottom to top)

b) Fill in the table.

correctly guessed	incorrectly guessed	best next guess ℓ	$P(L_i = \ell \text{ for some } i \in \{1,2,3,4,5\} \mid E)$
– – – – –	{}	E	0.5394
– – – – –	{E, A}	O	0.5340
A – – – S	{}	E	0.7715
A – – – S	{I}	E	0.7127
– – O – –	{A, E, M, N, T}	R	0.7454
– – – – –	{E, O}	I	0.6366
D – – I –	{}	A	0.8207
D – – I –	{A}	E	0.7521
– U – – –	{A, E, I, O, S}	Y	0.6270

c) Print source code.

```python
import numpy as np
import pandas as pd
### Problem 1a ###

# Read lines from file and save as pandas dataframe. Separate lines into two columns.
data = pd.read_csv('hw1_word_counts_05.txt', sep=" ", header=None, names=['Word', 'Count'])

# Compute prior probability and add as third column of dataframe for each word
data['P(W=w)'] = data['Count']*1.0/data['Count'].sum()

# Sort the dataframe by the word's prior probability (Second column of dataframe)
data_sorted = data.sort_values(by = ['P(W=w)'], ascending=False)

# Sanity check. 15 most frequent and 14 least frequent words.
print(data_sorted.head(15))
print(data_sorted.tail(14))

### Problem 1b ###

# Make words and prior probabilities from pandas dataframe to lists
words = data['Word'].tolist()
priors = data['P(W=w)'].to_numpy()
alphabet = ['A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z']

# Check if word is possible given the evidence
def isWordPossible(word, correct_evidence, incorrect_evidence):
    for i, letter in enumerate(word):
        if correct_evidence[i] == None:
            if letter in correct_evidence:
                return 0
        if correct_evidence[i] != None:
            if letter != correct_evidence[i]:
                return 0
        if letter in incorrect_evidence:
            return 0
    return 1

# Evaluate the denominator in posterior probability
def posteriorDenominator(words, priors, correct_evidence, incorrect_evidence):
    denominator = 0.0
    for i, word in enumerate(words):
        denominator += isWordPossible(word,correct_evidence,incorrect_evidence)*priors[i]
    return denominator

# Evaluate posterior probability for all words
def posteriorProbability(words, priors, correct_evidence, incorrect_evidence):
    posteriors = []
    denominator = posteriorDenominator(words, priors, correct_evidence, incorrect_evidence)
    for i, word in enumerate(words):
        nominator = isWordPossible(word,correct_evidence,incorrect_evidence)*priors[i]
        posteriors.append(nominator*1.0/denominator*1.0)
    return posteriors

# Check if a letter is in a word
def isLetterInWord(letter, word):
    for char in word:
        if letter == char:
            return 1
    return 0

# Predict the probability of a single letter
def letterPredictiveProbability(posteriors, letter, words, priors, correct_evidence, incorrect_evidence):
    posteriors = posteriorProbability(words, priors, correct_evidence, incorrect_evidence)
    letter_prob = 0.0
    for i, word in enumerate(words):
        letter_in_word = isLetterInWord(letter,word)
        letter_prob += letter_in_word*posteriors[i]
    if letter in correct_evidence or letter in incorrect_evidence:
        letter_prob = 0
    return letter_prob

```

```python
# Predict probability of all letters, returning the letter with largest probability
def nextGuess(alphabet, words, priors, correct_evidence, incorrect_evidence):
    print("Correct evidence: ", correct_evidence)
    print("Incorrect evidence: ", incorrect_evidence)
    max_probability = 0.0
    max_letter = ''
    posteriors = posteriorProbability(words,priors,correct_evidence,incorrect_evidence)
    for i, alpha in enumerate(alphabet):
        letter_probability = letterPredictiveProbability(posteriors,alpha,words,priors,correct_evidence,incorrect_evidence)
        if letter_probability > max_probability:
            max_probability = letter_probability
            max_letter = alphabet[i]
    print("Next letter: ", max_letter, " with probability: ", max_probability, '\n')

# TEST CASES
print("Test case 1")
correct_evidence = [None,None,None,None,None]
incorrect_evidence = []
nextGuess(alphabet,words,priors,correct_evidence,incorrect_evidence)

print("Test case 2")
correct_evidence = [None,None,None,None,None]
incorrect_evidence = ['E','A']
nextGuess(alphabet,words,priors,correct_evidence,incorrect_evidence)

print("Test case 3")
correct_evidence = ['A',None,None,None,'S']
incorrect_evidence = []
nextGuess(alphabet,words,priors,correct_evidence,incorrect_evidence)

print("Test case 4")
correct_evidence = ['A',None,None,None,'S']
incorrect_evidence = ['I']
nextGuess(alphabet,words,priors,correct_evidence,incorrect_evidence)

print("Test case 5")
correct_evidence = [None,None,'O',None,None]
incorrect_evidence = ['A','E','M','N','T']
nextGuess(alphabet,words,priors,correct_evidence,incorrect_evidence)
```