CSE 150A / 250A - Homework 1 (Solutions) - *48 pts total*

1.1 Conditioning on background evidence *(3 pts)*

(a) **Product rule with background evidence** *(1 pt)*

By the standard product rule, we have:

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

(b) **Bayes rule with background evidence** *(1 pt)*

By the product rule in part (a), we have:

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

Dividing the last two expression by $P(Y|E)$ yields the desired form of Bayes rule:

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

(c) **Marginalization with background evidence** *(1 pt)*

By the product rule and marginalization:

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

1.2 Conditional independence *(3 pts)*

We wish to show that the following three statements are equivalent:

$$\begin{array}{lrcl} (1) & P(X,Y|E) & = & P(X|E)P(Y|E) \\ (2) & P(X|Y,E) & = & P(X|E) \\ (3) & P(Y|X,E) & = & P(Y|E) \end{array}$$

To do this, we must show that each statement implies the other two:

$$\begin{array}{lcll} (1) & \rightarrow & (2),(3) & \textit{(1 pt)} \\ (2) & \rightarrow & (1),(3) & \textit{(1 pt)} \\ (3) & \rightarrow & (1),(2) & \textit{(1 pt)} \end{array}$$

We start by showing that the first implies the second. In particular, dividing statement (i) by $P(Y|E)$ we obtain

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

Next we apply the conditionalized product rule to the numerator:

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

Finally, substituting the right hand side of this expression into the numerator yeilds statement (ii):

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

Inverting these steps allows us to deduce (i) from (ii). Transposing X and Y , the same steps show that (i) implies (iii) and (iii) implies (i). It follows (by chaining) that (ii) implies (iii) and (iii) implies (ii). Thus the three statements are equivalent.

1.3 Creative writing *(3 pts)*

We are asked to assign events to the binary random variables X, Y, and Z that are consistent with the following patterns of commonsense reasoning. Of course there are many possible solutions. Below are some examples.

(a) **Cumulative evidence** *(1 pt)* Consider a single effect X with multiple causes Y, Z. For example:

Let X indicate whether I am late.

Let Y indicate whether there is traffic.

Let Z indicate whether I oversleep.

Then $P(X=1) < P(X=1|Y=1) < P(X=1|Y=1, Z=1)$

(b) **Explaining away** *(1 pt)* Consider a single effect Y with independent unlikely causes X, Z. For example:

Let Y indicate whether you have a stomach ache.

Let X indicate whether you have an ulcer.

Let Z indicate whether you have food poisoning.

Then $P(X=1|Y=1) > P(X=1)$ and $P(X=1|Y=1, Z=1) < P(X=1|Y=1)$

(c) **Conditional independence** *(1 pt)* Consider a single cause Z of two effects X, Y. For example:

Let Z indicate whether an exam is too long to finish.

Let X indicate whether one student finishes.

Let Y indicate whether another student finishes.

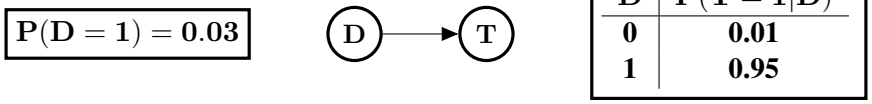
Then $P(X=1, Y=1) \neq P(X=1)P(Y=1)$
and $P(X=1, Y=1|Z=1) = P(X=1|Z=1)P(Y=1|Z=1)$

1.4 Bayes Rule *(6 pts)*

Suppose that 3% of competitive cyclists use performance-enhancing drugs and that a particular drug test has a 1% false positive rate and a 5% false negative rate.

(a) **Belief Network** *(2 pts)*

Let $D \in \{0, 1\}$ indicate the use of drugs, and let $T \in \{0, 1\}$ indicate the test result. Since 3% of cyclists use drugs, we have $P(D = 1) = 0.03$. Since the test has a 1% false positive rate, we have $P(T = 1|D = 0) = 0.01$. Since the test has a 5% false negative rate, we have $P(T = 0|D = 1) = 0.05$. The belief network is shown below.



D	P(T = 1|D)
0	0.01
1	0.95

(b) **False Negative test** *(2 pts)* Cyclist A test negative for drug use. What is the probability that Cyclist A is using drugs?

$$
\begin{aligned}
P(D = 1|T = 0) &= \frac{P(T = 0|D = 1)P(D = 1)}{P(T = 0)} && \text{Bayes rule} \\
&= \frac{P(T = 0|D = 1)P(D = 1)}{\sum_d P(T = 0, D = d)} && \text{marginalization} \\
&= \frac{P(T = 0|D = 1)P(D = 1)}{\sum_d P(T = 0|D = d)P(D = d)} && \text{product rule} \\
&= \frac{(0.05)(0.03)}{(0.05)(0.03) + (1 - 0.03)(1 - 0.01)} && \text{substitution} \\
&\approx \mathbf{0.00156}
\end{aligned}
$$

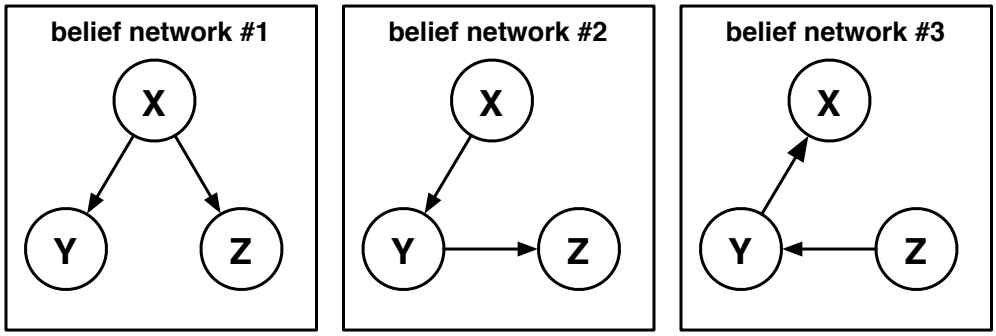
(c) **False Positive test** *(2 pts)*

Cyclist B tests positive for drug use. What is the probability that Cyclist B is not using drugs?

$$
\begin{aligned}
P(D = 0|T = 1) &= \frac{P(T = 1|D = 0)P(D = 0)}{P(T = 1)} && \text{Bayes rule} \\
&= \frac{P(T = 1|D = 0)P(D = 0)}{\sum_d P(T = 1, D = d)} && \text{marginalization} \\
&= \frac{P(T = 1|D = 0)P(D = 0)}{\sum_d P(T = 1|D = d)P(D = d)} && \text{product rule} \\
&= \frac{(0.01)(1 - 0.03)}{(0.01)(1 - 0.03) + (0.03)(1 - 0.05)} && \text{substitution} \\
&\approx \mathbf{0.25393}
\end{aligned}
$$

1.5 Compare and contrast *(3 pts)*

Consider the different belief networks (BNs) shown below for the discrete random variables X, Y, and Z.



(a) Does the first belief network imply a statement of marginal or conditional independence that is not implied by the second? If yes, provide an example.

Yes - BN #1 implies that $P(Y, Z|X) = P(Y|X)P(Z|X)$

(b) Does the second belief network imply a statement of marginal or conditional independence that is not implied by the third? If yes, provide an example.

No - the independence relations implied by BN #2 are also implied by BN #3.

(c) Does the third belief network imply a statement of marginal or conditional independence that is not implied by the first? If yes, provide an example.

Yes - BN #3 implies that $P(X, Z|Y) = P(X|Y)P(Z|Y)$

1.6 Hangman *(12 pts)*

(a) **Most and least frequent words** *(1 pt)*

1. **Most frequent:** THREE, SEVEN, EIGHT, WOULD, ABOUT, THEIR, WHICH, AFTER, FIRST, FIFTY, OTHER, FORTY,YEARS,THERE,SIXTY.
2. **Least frequent:** BOSAK, CAIXA, MAPCO, OTTIS, TROUP, CCAIR, CLEFT, FABRI, FOAMY, NIAID, PAXON, SERNA, TOCOR, YALOM.

(b) **Best next guesses and their probabilities** *(5 pts)*

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	$\{$I$\}$	E	0.7127
$-----$	$\{$E, O$\}$	I	0.6366
D$--$I$-$	$\{$A$\}$	E	0.7521
$-$U$---$	$\{$A, E, I, O, S$\}$	Y	0.6270

(c) **Source code with output** *(6 pts)*

1.7 Entropy *(6pts)*

(a) **Maximum Entropy** *(2pts)*

We wish to maximize $-\sum_{i=1}^{n} p_i \log p_i$ subject to the constraint $\sum_{i=1}^{n} p_i = 1$. Let λ denote a Lagrange multipler for this constraint. Then the Lagrangian is given by:

$$\mathcal{L}(p, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

Setting the gradient of the Lagrangian to zero yields:

$$-\log p_i - 1 - \lambda = 0 \quad \forall i,$$
$$\sum_{i=1}^{n} p_i = 1.$$

Thus, $p_i = e^{\lambda - 1}$ regardless of i. Since all p_i are equal and sum to unity, it follows that $p_i = \frac{1}{n}$ is the distribution with maximum entropy.

Bonus: strictly speaking, one must also show that the above solution corresponds to the global maximum of the entropy (although this is not required for full credit). The simplest proof of this comes from the next problem in KL divergence. Specifically, if $KL(p, q) \geq 0$ for all distribution p and q, what is implied by $p_i = \frac{1}{n}$?

(b) **Joint Entropy** *(4pts)*

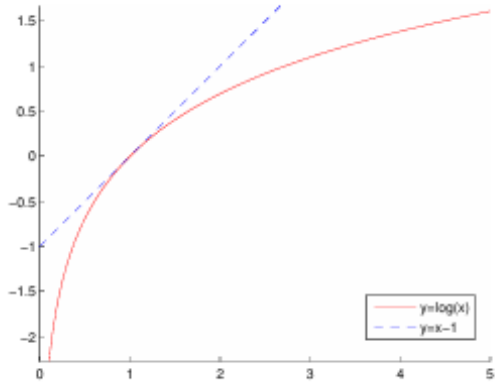
$$
\begin{aligned}
H(X_1, ..., X_n) &= -\sum_{x_1} \cdots \sum_{x_n} P(x_1, ..., x_n) \log P(x_1, ..., x_n) \quad \boxed{\textbf{Definition of joint entropy}} \\
&= -\sum_{x_1} \cdots \sum_{x_n} P(x_1, ..., x_{n-1}) P(x_n) \log[\log P(x_1, ..., x_{n-1}) P(x_n)] \quad \boxed{\textbf{Marginal independence}} \\
&= -\sum_{x_1} \cdots \sum_{x_n} P(x_1, ..., x_{n-1}) P(x_n) [\log P(x_1, ..., x_{n-1}) + \log P(x_n)] \quad \boxed{\textbf{Logarithms}} \\
&= -\sum_{x_1} \cdots \sum_{x_{n-1}} P(x_1, ..., x_{n-1}) \log P(x_1, ..., x_{n-1}) \sum_{x_n} P(x_n) \quad \boxed{\textbf{Grouping terms}} \\
&\quad + \sum_{x_1} \cdots \sum_{x_{n-1}} P(x_1, ..., x_{n-1}) \left[-\sum_{x_n} P(x_n) \log P(x_n) \right] \\
&= \quad H(X_1, ..., X_{n-1}) + H(X_n) \quad \boxed{\textbf{Distributions are normalized}}
\end{aligned}
$$

Inductively, we conclude:

$$H(X_1, ..., X_n) = H(X_1) + H(X_2) + \cdots + H(X_n)$$

1.8 Kullback-Leibler distance *(9 pts)*

(a) **Inequality** *(2 pts – 1 pt for plot, 1 pt for differentiation*



Let $f(x) = \log x - x + 1$. Then the derivative is $f'(x) = \frac{1}{x} - 1$. The derivative is positive for $x \in (0, 1)$ where $f(x)$ is increasing, zero for $x = 1$ and negative for $x > 1$ where $f(x)$ is decreasing. Thus $f(x)$ has a single maximum at $x = 1$, where $\log x = x - 1$, and $\log x < x - 1$ everywhere else.

(b) **Proof** *(4 pts – 2pts for inequality, 1 pt for each direction of implication)*

To derive the general inequality:

$$
\begin{aligned}
KL(p, q) &= \sum_i p_i \log(\frac{p_i}{q_i}) \\
&= -\sum_i p_i \log(\frac{q_i}{p_i}) \\
&\geq -\sum_i p_i (\frac{q_i}{p_i} - 1) && \text{by inequality from part (a)} \\
&= \sum_i (p_i - q_i) \\
&= \sum_i p_i - \sum_i q_i \\
&= 1 - 1 && \text{because sums are normalized} \\
&= 0
\end{aligned}
$$

Now suppose that $p_i \neq q_i$ for some i. Then from part (a), we have $-\log(\frac{q_i}{p_i}) > -(\frac{q_i}{p_i} - 1)$, and the inequality in the third line of the above derivation is replaced by a strict inequality. In this case it follows that $KL(p, q) > 0$; moreover, from the contrapositive, we have that $KL(p, q) = 0$ implies $p_i = q_i$ for all i.

Finally suppose that $p_i = q_i$ for all i. In this case, the third line of the above derivation is replaced by an equality. It follows then that $KL(p, q) = 0$.

(c) **Refined inequality** *(2 pt)*

$$\begin{aligned} KL(p,q) &= -\sum_i p_i \log(\frac{q_i}{p_i}) \\ &= -\sum_i 2p_i \log(\sqrt{\frac{q_i}{p_i}}) \\ &\geq -\sum_i 2p_i(\sqrt{\frac{q_i}{p_i}} - 1) \qquad \boxed{\textbf{by inequality from part (a)}} \\ &= -\sum_i (2\sqrt{q_i p_i} - 2p_i) \\ &= \sum_i (p_i - 2\sqrt{q_i p_i} + q_i) \\ &= \sum_i (\sqrt{p_i} - \sqrt{q_i})^2. \end{aligned}$$

(d) **Counter example** *(1 pt)* Let $p = (0.5, 0.5)$ and $q = (0.1, 0.9)$. Then:

$$KL(p,q) = (0.5)\log 5 + (0.5)\log\frac{5}{9} \approx \mathbf{0.511}$$
$$KL(q,p) = (0.1)\log\frac{1}{5} + (0.9)\log\frac{9}{5} \approx \mathbf{0.368}$$

1.9 Mutual information *(3 pts)*

(a) **Nonnegativity** *(2 pts)* The quickest proof that $I(X,Y)$ is nonnegative is to recognize that it is equal to the KL distance between the joint distribution $P(X,Y)$ and the product $P(X)P(Y)$ of the marginal distributions. Since the KL distance is always nonnegative, so is the mutual information. In particular:

$$\begin{aligned} I(X,Y) &= \sum_x \sum_y P(x,y) \log\left[\frac{P(x,y)}{P(x)P(y)}\right], \\ &= -\sum_x \sum_y P(x,y) \log\left[\frac{P(x)P(y)}{P(x,y)}\right], \\ &\geq \sum_x \sum_y P(x,y)\left[1 - \frac{P(x)P(y)}{P(x,y)}\right] \\ &= \sum_x \sum_y P(x,y) - \left(\sum_x P(x)\right)\left(\sum_y P(y)\right) \\ &= 1 - (1)(1) \\ &= 0 \end{aligned}$$

(b) **Mutual information and independence** *(1 pt)*
By part (b) of the previous problem, it follows that $I(X,Y) = 0$ if and only if $P(X,Y) = P(X)P(Y)$, which is exactly the condition that X and Y are independent random variables.