Assignment 1: Probabilistic Reasoning, Maximum Likelihood, Classification

For due date see `https://courses.cs.sfu.ca`

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student. Any mark in this assignment may be changed on the basis of an oral exam where you need to explain your solution.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment

---

Practice: Joint and Conditional Probabilities (10 marks)

Go to www.aispace.org and start the "belief and decision network" tool. Load the sample file "File/Load Sample Problem/Simple Diagnostic Example". *If you have difficulty with the UBC tool (e.g. Java doesn't run, Simple Diagnostic isn't found), I suggest you download the jar file from the course website.*

We will use this to test some of the basic probability laws. The AIspace tool can do many of these calculations for you, but the purpose of the exercise is to learn the principles behind the calculations. You can use the tool to check your answers, but you should compute them yourself using the proba-

bility calculus together with the conditional probabilities from the Bayesian network. *Compute the following joint probabilities up to 6 significant digits.*

1. Use the product formula of Bayes nets and the conditional probability parameters specified by AIspace to compute the probability that: all nodes are true.
   P(all nodes are True) = 0.00128

2. Use the product formula of Bayes nets and the conditional probability parameters specified by Aispace to compute the probability that: all nodes are true except for Sore Throat, and that Sore Throat is false.
   P(SoreThroat=False, all other nodes True) = 0.00299

3. Show how can you use these two joint probabilities to compute the probability that: all nodes other than Sore Throat are true. (Where the value of Sore Throat is unspecified.)
   P(all nodes other than SoreThroat True)
   = P(SoreThroat=True, all other nodes True) + P(SoreThroat=False, all other nodes True)
   = 0.00128 + 0.00299
   = 0.00427

4. Verify the product formula:
   P(all nodes are true) = P(Sore Throat = true | all other nodes are true) x P(all other nodes are true).
   You may get the first conditional probability by executing a query with the tool.
   P(all nodes are true)
   = P(SoreThroat True | all other nodes are True)×P(all other nodes are True)
   = 0.3 × 0.00427
   = 0.00128

5. Compute the probability that Sore Throat is true and that Fever is true. (Hint: If you use the right formula, you need only 4 conditional probabilities.)
   P(SoreThroat = True, Fever = True) = P(SoreThroat = True, Fever = True | Influenza = True)×P(Influenza = True)

+ P(SoreThroat = True, Fever = True | Influenza = False)×P(Influenza = False)
= P(SoreThroat = True | Influenza = True) × P(Fever = True | Influenza = True)×P(Influenza = True)
+ P(SoreThroat = True | Influenza = False) × P(Fever = True | Influenza = False)×P(Influenza = False)
= 0.3 × 0.9 × 0.05 + 0.001 × 0.05 × 0.95
= 0.0135

You can enter the computed probabilities in the table below.

Probability to be Computed	Your Result
P(all nodes true)	0.00128
P(Sore Throat = False, all other nodes true)	0.00299
P(all nodes other than Sore Throat true)	0.00427
P(all nodes are true) = P(Sore Throat = true | all other nodes are true) x P(all other nodes are true).	0.00128304 = 0.3 × 0.0042768
P(Sore Throat = true, Fever = True)	0.0135475

Theory: Conditional Probabilities (9 marks)

Exercise 13.3 in Russell and Norvig AMAI.

1. True.
   P(a | b, c) = P(b | a, c)
   $\implies \frac{P(a,b,c)}{P(b,c)} = \frac{P(b,a,c)}{P(a,c)}$
   $\implies P(a,c) = P(b,c)$
   $\implies \frac{P(a,c)}{P(c)} = \frac{P(b,c)}{P(c)}$
   $\therefore$ P(a | b, c) = P(b | a, c) $\implies$ P(a | c) = P(b | c)

2. False.
   Counter-Example: Consider two independent coin flips a and b, such that c = b.
   P(a = Head | b = Tail, c = Tail) = P(a = Head) = 0.5

P(b = Tail | c = Tail) = 1
P(b = Tail) = 0.5
P(b = Tail | c = Tail) ≠ P(b = Tail)

3. False.
   Counter-Example: Consider two independent coin flips a and b, such that c = a xor b.
   P(a = Tail | b = Tail) = P(a = Tail) = 0.5
   P(a = Tail | b = Tail, c = Tail) = 1
   P(a = Tail | c = Tail) = 0.5
   P(a = Tail | b = Tail, c = Tail) ≠ P(a = Tail | c = Tail)

Theory: Expectations (5 marks)

Any function $f(X)$ of a discrete random variable X defines a random variable. (Define $p(f(X) = y) = \sum_{x:f(x)=y} p(x)$.) Similarly, given a joint probability $p(X_1, \ldots, X_n)$, any function $f(X_1, \ldots, X_n)$ is also a random variable. Show that the expected value of the sum of two random variables is the sum of the expectations. In symbols, show that

Let denote S the sample space underlying a random experiment with elements $s \in S$. Let's define two random variables X_1 and X_2 whose domains are S.
$E[X_1 + X_2]$
$= \sum_{x1,x2 \in S}(x_1 + x_2) \times P(X_1 = x_1, X_2 = x_2)$
$= \sum_{x1 \in S} \sum_{x_2 \in S} x_1 \times P(X_1 = x_1, X_2 = x_2) + \sum_{x1 \in S} \sum_{x_2 \in S} x_2 \times P(X_1 = x_1, X_2 = x_2)$
$= \sum_{x1 \in S} x_1 \sum_{x_2 \in S} \times P(X_1 = x_1, X_2 = x_2) + \sum_{x_2 \in S} x_2 \sum_{x1 \in S} \times P(X_1 = x_1, X_2 = x_2)$
$= \sum_{x1 \in S} x_1 \times P(X_1 = x_1) + \sum_{x_2 \in S} x_2 \times P(X_2 = x_2)$
$= E[X_1] + E[X_2]$

Practice: Decision Tree Learning with ID3 (10 marks)

Figure 1 provides data about whether a customer will wait for a table in a restaurant or not. Assume that ID3 splits first on the Pat attribute (for Patrons). Show the following for the branch $Pat = Full$.

1. The next attribute chosen by ID3. There may be a tie among several attributes; you can list all or just one of them.

2. The expected information gain associated with the next attribute. Compare this with the expected information gain for Hungry.

3. How you calculated the expected information gain.

   Consider:
   $H(S) = -p_+ \times log_2(p_+) - p_- \times log_2(p_-)$
   $Gain(S, A) = H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} H(S_v)$
   H(Pat=Full)=0.91829 [2T, 4F]

   Alt:
   H(Alt=True)=0.97095 [2T, 3F]
   H(Alt=False)=0 [0T, 1F]
   Gain(Pat=Full, Alt)= H(Pat=Full) - $\frac{5}{6}$H(Alt=True) - $\frac{1}{6}$H(Alt=False) = 0.10917

   Bar:
   H(Bar=True)=0.91829 [1T, 2F]
   H(Bar=False)=0.91829 [1T, 2F]
   Gain(Pat=Full, Bar)= H(Pat=Full) - $\frac{1}{2}$H(Bar=True) - $\frac{1}{2}$H(Bar=False) = 0

   Fri:
   H(Fri=True)=1 [2T, 3F]
   H(Fri=False)=0 [0T, 1F]
   Gain(Pat=Full, Fri)= H(Pat=Full) - $\frac{5}{6}$H(Fri=True) - $\frac{1}{6}$H(Fri=False) = 0.1091

   Hun:
   H(Hun=True)=1 [2T, 2F]
   H(Hun=False)=0 [0T, 2F]
   Gain(Pat=Full, Hun)= H(Pat=Full) - $\frac{2}{3}$H(Hun=True) - $\frac{1}{3}$H(Hun=False) = 0.2516

Price:
H(Price=$)=1 [2T, 2F]
H(Price=$$$)=0 [0T, 2F]
Gain(Pat=Full, Price)= H(Pat=Full) - $\frac{2}{3}$H(Price=$) - $\frac{1}{3}$H(Price=$$$)
= 0.2516

Rain:
H(Rain=True)=0 [0T, 1F]
H(Rain=False)=0.97095 [2T, 3F]
Gain(Pat=Full, Rain)= H(Pat=Full) - $\frac{1}{6}$H(Rain=True) - $\frac{5}{6}$H(Rain=False)
= 0.1091

Res:
H(Res=True)=0 [0T, 2F]
H(Res=False)=1 [2T, 2F]
Gain(Pat=Full, Res)= H(Pat=Full) - $\frac{2}{6}$H(Res=True) - $\frac{4}{6}$H(Res=False)
= 0.2516

Type:
H(Type=Thai)=1 [1T, 1F]
H(Type=French)=0 [0T, 1F]
H(Type=Burger)=1 [1T, 1F]
H(Type=Italian)=0 [0T, 1F]
Gain(Pat=Full, Type)= H(Pat=Full) - $\frac{2}{6}$H(Type=Thai) - $\frac{1}{6}$H(Type=French)
-$\frac{2}{6}$H(Type=Burger) -$\frac{1}{6}$H(Type=Italian) = 0.2516

Est:
H(Est=10-30)=0 [1T, 1F]
H(Est=30-60)=1 [1T, 1F]
H(Est=>60)=1 [0T, 2F]
Gain(Pat=Full, Est)= H(Pat=Full) - $\frac{2}{6}$H(Est=10-30) - $\frac{2}{6}$H(Est=30-60)
- $\frac{2}{6}$H(Est=>60) = 0.2516

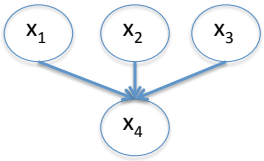
The highest information gain possible is 0.2516 and the tie is between Hungry, Price, Reservation, Type and Estimated Wait Time.

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Figure 1: Data Set for Question 5, Decision Tree Learning

Theory: Maximum Likelihood Parameter Estimation for Bayesian Networks (8 marks)

Consider learning the parameters for the Bayesian network shown in the figure.



Suppose we have a training set $(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N)$, where each $\boldsymbol{x}^i = (x_1^i, \ldots, x_4^i)$ is a vector containing values for all 4 random variables in the network.

1. Write down the likelihood and the log-likelihood of the training data given a parameter setting of the Bayes net. Please use the following notation.

   (a) θ_{ijk} = the conditional probability of node i taking on value k given that the parents of i are in state j. In our example, $i = 1, \ldots, 4$, and $k = 1, \ldots, L$. If $i = 1, 2, 3$, then $j = 0$, that is j is just a dummy index since the first three nodes have no parents. If $i = 4$,

then $j = 1, \ldots, L^3$.

(b) n_{ijk} = the number of training cases where node i takes on value k and the parents of i are in state j.

$$
\begin{aligned}
&\text{P(X} \mid \theta\text{)} \\
&= \textstyle\prod_{n=1:N} P(x_1^n, x_2^n, x_3^n, x_4^n|\theta) \\
&= \textstyle\prod_{n=1:N} P(x_4^n|x_1^n, x_2^n, x_3^n, \theta) \times P(x_3^n|\theta) \times P(x_2^n|\theta) \times P(x_1^n|\theta) \\
&= (\textstyle\prod_{k=1:L} \theta_{10k}^{n_{10k}} \times \theta_{20k}^{n_{20k}} \times \theta_{30k}^{n_{30k}}) \times (\prod_{k=1:L} \prod_{j=1:L^3} \theta_{4jk}^{n_{4jk}}) \\
&= (\textstyle\prod_{i=1:3} \prod_{k=1:L} \theta_{i0k}^{n_{i0k}}) \times (\prod_{k=1:L} \prod_{j=1:L^3} \theta_{4jk}^{n_{4jk}})
\end{aligned}
$$

$$
\begin{aligned}
&\text{log(P(X} \mid \theta\text{))} \\
&= log((\textstyle\prod_{i=1:3} \prod_{k=1:L} \theta_{i0k}^{n_{i0k}}) \times (\prod_{k=1:L} \prod_{j=1:L^3} \theta_{4jk}^{n_{4jk}})) \\
&= (\textstyle\sum_{i=1:3} \sum_{k=1:L} log(\theta_{i0k}^{n_{i0k}})) + (\sum_{k=1:L} \sum_{j=1:L^3} log(\theta_{4jk}^{n_{4jk}})) \\
&= (\textstyle\sum_{i=1:3} \sum_{k=1:L} n_{i0k} \times log(\theta_{i0k})) + (\sum_{k=1:L} \sum_{j=1:L^3} n_{4jk} \times log(\theta_{4jk}))
\end{aligned}
$$

2. Show that with binary nodes ($L = 2$) the maximum likelihood parameter values $\hat{\boldsymbol{\theta}}_{ijk}$ are the conditional frequencies observed in the data:

$$\hat{\theta}_{ijk} = \frac{n_{ijk}}{\sum_{k'} n_{ijk'}}$$

Setting L=2:

$$
\begin{aligned}
&(\textstyle\sum_{i=1:3} \sum_{k=1:2} n_{i0k} \times log(\theta_{i0k})) + (\sum_{k=1:2} \sum_{j=1:8} n_{4jk} \times log(\theta_{4jk})) \\
&= (\textstyle\sum_{i=1:3} n_{i01} \times log(\theta_{i01}) + n_{i02} \times log(1 - \theta_{i01})) \\
&+(\textstyle\sum_{j=1:8} n_{4j1} \times log(\theta_{4j1}) + n_{4j2} \times log(1 - \theta_{4j1}))
\end{aligned}
$$

Derive and equal to zero:

$$
\begin{aligned}
&\tfrac{\partial log(P(X|\theta))}{\partial \theta_{ij1}} = 0 \\
&\implies \tfrac{\partial (n_{ij1} \times log(\theta_{ij1}) + n_{ij2} \times log(1-\theta_{ij1}))}{\partial \theta_{ij1}} = 0 \\
&\implies \tfrac{n_{ij1}}{\theta_{ij1}} - \tfrac{n_{ij2}}{1-\theta_{ij1}} = 0 \\
&\implies \theta_{ij1} \times (n_{ij1} + n_{ij2}) = n_{ij1} \\
&\implies \theta_{ij1} = \tfrac{n_{ij1}}{\sum_{k'} n_{ijk'}} \\
&\implies \theta_{ijk} = \tfrac{n_{ijk}}{\sum_{k'} n_{ijk'}}
\end{aligned}
$$

The maximum likelihood result holds generally for a Bayesian network with discrete variables. I have given you a specific structure with binary variables only for the sake of concreteness.

Theory: Maximum Likelihood Parameter Estimation for a Gaussian Distribution (12 marks)

Consider a Gaussian or Normal Distribution with parameters mean μ and variance σ^2 and probability density function $f(x; \mu, \sigma^2)$. Suppose we have a training set $\boldsymbol{x} = (x^1, \ldots, x^N)$ of observed values for random variable X. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the maximum likelihood estimates of the distribution parameters estimated from the training set.

1. Write down the log-likelihood function $\mathcal{L}(\boldsymbol{x}; \mu, \sigma^2)$.
   Considering:
   $f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
   $P(D|\mu, \sigma^2) = \prod_{i=1:N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^i-\mu)^2}{2\sigma^2}}$
   $log(P(D|\mu, \sigma^2)) = \sum_{i=1:N} log(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^i-\mu)^2}{2\sigma^2}})$
   $= N \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \sum_{i=1:N} (x^i - \mu)^2$

2. Show that the maximum likelihood estimate for the distribution mean is the sample mean: $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x^i \equiv \overline{x}$.

   $\frac{\partial(P(D|\mu,\sigma^2))}{\partial\mu} = 0$
   $\implies \frac{\partial(N \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \sum_{i=1:N} (x^i-\mu)^2)}{\partial\mu} = 0$
   $\implies \sum_{i=1:N} (x^i - \mu) = 0$
   $\implies \sum_{i=1:N} x^i = N\mu$
   $\implies \mu = \frac{\sum_{i=1:N} x^i}{N} \equiv \bar{x}$

3. Show that the maximum likelihood estimate for the distribution variance is the sample variance: $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x^i - \overline{x})^2$ assuming that $\mu = \overline{x}$.
   $\frac{\partial(P(D|\mu,\sigma))}{\partial\sigma} = 0$
   $\implies \frac{\partial(N \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \sum_{i=1:N} (x^i-\mu)^2)}{\partial\sigma} = 0$
   $\implies -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1:N} (x^i - \mu)^2 = 0$
   $\implies \frac{N}{\sigma} = \frac{1}{\sigma^3} \sum_{i=1:N} (x^i - \mu)^2$
   $\implies \sigma^2 = \frac{1}{N} \sum_{i=1:N} (x^i - \mu)^2$
   $\implies \sigma^2 = \frac{1}{N} \sum_{i=1:N} (x^i - \bar{x})^2$