CS 412: Introduction to Machine Learning
Spring 2017
Homework 2
Due: February 21st (beginning of class)

Problem 1. Independence (15 points)

For each of the following problems, provide your answer and show the steps taken to solve the problem.

(a) For the following distribution, is $A \perp B$ (i.e., A and B are independent)? **(5 points)**

a	b	P(A=a,B=b)
0	0	0.5
0	1	0.0
1	0	0.0
1	1	0.5

(b) For the following distribution, is $A \perp B|C$ (i.e., A and B are conditionally independent given C)? **(5 points)**

a	b	c	P(A=a,B=b,C=c)
0	0	0	0.056
0	0	1	0.120
0	1	0	0.224
0	1	1	0.120
1	0	0	0.024
1	0	1	0.180
1	1	0	0.096
1	1	1	0.180

(c) Consider two binary random variables A and B. If $A \perp B$ (i.e., A and B are independent), and $P(A = 0, B = 0) = 0.18$ and $P(A = 1, B = 1) = 0.28$, what is the probability of $P(A = 0, B = 1)$? **(5 points)**

a	b	P(A=a,B=b)
0	0	0.18
0	1	?
1	0	?
1	1	0.28

Problem 2. Bayesian Parameter Estimation (20 points)

In Homework 1, you showed how to estimate the parameter λ for the exponential distribution: $f_\lambda(x) = \lambda e^{-\lambda x}$. We will now consider Bayesian parameter estimation for this distribution.

(a) Using a prior distribution from the Gamma distribution, $f_{\alpha,\beta}(\lambda) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\lambda\beta}}{\Gamma(\alpha)}$, with parameters α and β, show that the posterior distribution for λ after updating using three datapoints, x_1, x_2, x_3, is also a Gamma distribution and show its new parameter values, α' and β', in terms of α, β, x_1, x_2, and x_3. **(10 points)**

(b) If our prior parameters are $\alpha = 2$ and $\beta = 2$, and our data sample consists of $x_1 = 3.7, x_2 = 4.5, x_3 = 4.8$:

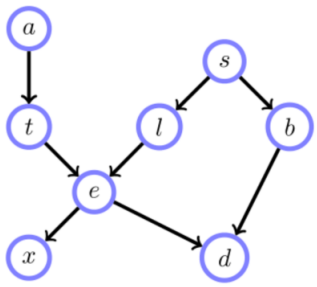
Compute the posterior probability of a new datapoint $x_4 = 3.8$ under the fully Bayesian estimation of λ. You can either leave your answer in terms of the Gamma function, or provide the exact answer. **(5 points)**

Hints: (i) You shouldn't have to solve the complicated integral; (ii) Since the Gamma distribution normalizes to 1, $\int_\lambda \lambda^{\alpha-1} e^{-\lambda\beta} d\lambda = \frac{\Gamma(\alpha)}{\beta^\alpha}$; (iii) The Gamma function is related to the factorial function as $\Gamma(x) = (x-1)!$ for positive integers x.

(c) If we have the same prior and datapoints as in (b), what is the probability of new datapoint $x_4 = 3.8$ using maximum a posteriori estimation of λ? **(5 points)**

Hint: (i) The mode of the Gamma distribution (i.e., the λ that attains its maximal probability) is $\frac{\alpha-1}{\beta}$.

Problem 3. Bayesian Networks (30 points) The following Bayesian network has been proposed to capture the relationships between lung disease diagnosis (tuberculosis, lung cancer, both, or neither) and its relationship to visiting Asia (Lauritzen & Spiegelhalter, 1988). Each is binary-valued and can either take the value *true* or the value *false*.



x = Positive X-ray
d = Dyspnea (Shortness of breath)
e = Either Tuberculosis or Lung Cancer
t = Tuberculosis
l = Lung Cancer
b = Bronchitis
a = Visited Asia
s = Smoker

(a) For any probability distribution corresponding to this Bayesian Network, is each of the following true or false?

(i) tuberculosis $\perp$ bronchitis | positive x-ray, smoker **(3 points)**

(ii) visit to Asia $\perp$ bronchitis | smoker **(3 points)**

(iii) smoker $\perp$ shortness of breath | lunge cancer **(3 points)**

(iv) a $\perp$ s | e, l, d **(3 points)**

(b) Consider the following conditional probability distribution for the Bayesian network.

$$P(a = \text{true}) = 0.03 \qquad P(s = \text{true}) = 0.5$$
$$P(t = \text{true}|a = \text{true}) = 0.1 \qquad P(t = \text{true}|a = \text{false}) = 0.01$$
$$P(l = \text{true}|s = \text{true}) = 0.05 \qquad P(l = \text{true}|s = \text{false}) = 0.01$$
$$P(b = \text{true}|s = \text{true}) = 0.7 \qquad P(b = \text{true}|s = \text{false}) = 0.3$$
$$P(x = \text{true}|e = \text{true}) = 0.98 \qquad P(x = \text{true}|e = \text{false}) = 0.05$$
$$P(d = \text{true}|e = \text{true}, b = \text{true}) = 0.9 \qquad P(d = \text{true}|e = \text{true}, b = \text{false}) = 0.6$$
$$P(d = \text{true}|e = \text{false}, b = \text{true}) = 0.8 \qquad P(d = \text{true}|e = \text{false}, b = \text{false}) = 0.05$$
$$P(e = \text{true}|t, l) = 0 \text{ if t and l are false, 1 otherwise}$$

Using this table, calculate the following probabilities by hand using variable elimination (show your work):

(i) A patient that does not smoke, visited Asia recently. His X-ray indicates a negative lung disease diagnosis, but he has shortness of breath. What is the probability of he has bronchitis? $[P(B = \text{true} \,|A = \text{true}, S = \text{false}, X = \text{false}, D = \text{true})]$ **(9 points)**

(ii) Another patient also recently visited Asia, has negative diagnosis in X-ray test, and also has shortness of breath. However, he does smoke. What is the probability of he has bronchitis? $[P(B = \text{true} \mid A = \text{true}, S = \text{true}, X = \text{false}, D = \text{true})]$ **(9 points)**

Problem 4. Naïve Bayes Classification (Programming) (40 points)
In this problem, we will attempt to identify spam SMS messages using the naïve Bayes model. You will need to download the SMS dataset: http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection.

Consider an SMS message as a (case-insensitive) sequence of words $(X1, \cdots, X_T)$. Ignore all other punctuation. Under the naïve Bayes assumption, the probability of the words in each message factor as:

$$P(\mathbf{x}_{1:T}|y) = \prod_{t=1}^{T} P(x_i|y). \tag{1}$$

When estimated from dataset $\mathcal{D}$ with pseudo-count prior of α, the model parameters are:

$$\hat{P}(x_i|y) = \frac{\text{Count}_{\mathcal{D}}(x_i, y) + \alpha}{\text{Count}_{\mathcal{D}}(y) + N\alpha}, \tag{2}$$

where: $\text{Count}_{\mathcal{D}}(x_i, y)$ and $\text{Count}_{\mathcal{D}}(y)$ are the number of occurances of word x_i in spam/ham messages y (from our sample $\mathcal{D}$); and the number of words for label spam/ham words y (from our sample $\mathcal{D}$) respectively; and N is the total number of words (including words not seen). Lets use $N = 20,000$ and $\alpha = 0.1$ in our experiments.

Note that the classes are very imbalance. The number of spam messages is 747, while the number of ham messages is 4827. If a simple classifier predict that all messages are ham, it will get around 86% accuracy. In this case, accuracy is not a good measurement of the classifier's performance.

Instead of using accuracy, we can use confusion matrix to see the performance of our model. Below is the explanation of confusion matrix:

		True condition	
		Positive	Negative
Predicted Condition	Positive	True positive	False positive
	Negative	False negative	True negative

Other important performance measurements are **precision**, **recall**, and **F-score**, defined as:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{3}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{4}$$

$$\text{F-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

(a) Randomly split the messages into a training set $\mathcal{D}_1$ (80% of messages) and a testing set $\mathcal{D}_2$ (20% of messages). Calculate the testing accuracy, confusion matrix, precision, recall, and F-score of the Naïve Bayes classifier in determining whether a message is spam or ham. Submit your source code.
Note: Let's assume that spam is the positive class. **(20 points)**

(b) How does the change of α effect the classifier performance? Using random split above, evaluate the training and testing accuracy and F-score under different selections of α. The selection of α values are 2^i where $i = -5, \cdots, 0$. Create two plots, the first plot is for the accuracy measure and the second plot is for F-score. In each plot, x-axis represents i, and y-axis represents the performance measure (accuracy/F-score). Each plot contains two line chart, a line chart describing training accuracy/F-score measure, the other line chart is for testing accuracy/F-score. Submit your source code. **(20 points)**

Hints: There are scripts in both Octave (*nbayes.m*) and Julia (*nbayes.jl*) for counting occurances of words from spam/ham messages. A few notes: (i) Octave can make use of Java data structures (e.g., the hashmap); this can be easier / more efficient than using Octave structures. (ii) Julia has its own data structure library which includes many standard data structures e.g. `Dict` for hashmap. (iii) $\log xy = \log x + \log y$.