Assignment 4

CMSC 471 (03/01) — Artificial Intelligence

Item	Summary
Assigned	Saturday April 24th
Due	Tuesday May 4th, 11:59 PM Baltimore time
Topic	Probability & Machine Learning (part 2)
Points	75

In this assignment you will gain experience with probability and some machine learning inference techniques.

You are to *complete* this assignment on your own: that is, the code and writeup you submit must be entirely your own. However, you may discuss the assignment at a high level with other students or on the discussion board. Note at the top of your assignment who you discussed this with or what resources you used (beyond course staff, any course materials, or public Piazza discussions).

The following table gives the overall point breakdown for this assignment.

Question	1	2	3
Points	25	25	25

What To Turn In You must turn in two items:

1. A writeup in PDF format that answer the questions.
2. A single `zip` or `tar.gz` file containing all code, environment files (if applicable) and execution instructions necessary to replicate your output.

As part of your submission, be sure to include specific instructions on how to build (compile) your code. Answers to the following questions should be long-form. Provide any necessary analyses and discussion of your results.

How To Submit Submit the assignment on the submission site:

https://www.csee.umbc.edu/courses/undergraduate/471/spring21/01_03/submit.

Be sure to select "`Assignment 4`."

Questions

1. (**25 points**) Assume we have three variables, X, Y, and Z, where X is defined over three values (a, b, c), Y is defined over two values (y, n), and Z is defined over four values (1, 2, 3, 4).

   Let $p(Z|X,Y)$ be defined as

		$Z=1$	$Z=2$	$Z=3$	$Z=4$
$X=a$	$Y=y$	0.41	0.11	0.41	0.07
	$Y=n$	0.12	0.48	0.14	0.26
$X=b$	$Y=y$	0.22	0.38	0.25	0.15
	$Y=n$	0.23	0.15	0.37	0.25
$X=c$	$Y=y$	0.38	0.26	0.11	0.25
	$Y=n$	0.18	0.25	0.38	0.19

   and define the prior probabilities $p(X=a)=0.4, p(X=b)=0.35$, and $p(Y=y)=0.6$.

   (a) If we know two of the values of the prior definition of X, how can we compute the remaining one?

   (b) Let's say we want to compute $p(Z|X,Y)$. How many different distributions do we need to compute?

   (c) Compute $p(Z=3)$, using the CPTs of $p(Z|X,Y)$ and priors $p(X)$ and $p(Y)$ above. Show your work.

   (d) Compute $p(X=a|Z=3)$, using the CPTs of $p(Z|X,Y)$ and priors $p(X)$ and $p(Y)$ above. Show your work.

   (e) Are X and Y independent? Show your work/justification.

2. (**25 points**) The following spreadsheet is a mini-demo of Naive Bayes as applied to a text classification task:

   `https://www.csee.umbc.edu/courses/undergraduate/471/spring21/01_03/materials/a4/nb-spreadsheet/`.

   (You should be able to make a copy of it if you want to make edits, but you do not need to edit it to answer this question.)

   This spreadsheet demos predicting whether or not a tweet will be retweeted, based solely on how many times any given word appears in that tweet. Your task in this question is to replicate the training and evaluation shown in this spreadsheet using the sklearn.naive_bayes.MultinomialNB class. Turn in your code, including instructions on how to run your code. For this question, there is no starter code.

   Allowed Online Resources: In addition to anything discussed or shown during lecture, for this question, you may reference any page on the scikit-learn.org website. You may *not* use external websites or references (such as StackOverflow). **Cite** whatever you reference.

   Answer the following for "Fixed Corpora BOW." These data are contained in the "Fixed Corpora BOW," but for your ease, the per-instance word counts are replicated in cells H9-N18 of the Training tab.

(a) How many classes are there?

(b) Draw the Bayesian network associated with this model.

(c) In instantiating the MultinomialNB class, what values do you use for the constructor arguments alpha, fit_prior and class_prior? Explain why those are the values you use.

(d) Use your code to replicate the values provided in the three probability tables in the Training tab (roughly, cells A4-E26). Provide a printout or screenshot of your code replicating these.

(e) Explain why we need to compute columns H, I, K, and L in the Testing tab.

(f) Use your code to replicate the six different log probability values in the Testing tab (roughly, cells H1-M11). Use the natural logarithm. Provide a printout or screenshot of your code replicating these.

(g) Assuming you're looking to evaluate how well you can identify retweeted tweets, compute accuracy, precision, recall, and F1 on the test data. You may use the sklearn library to do this.

3. (**25 points**) Consider a Bayesian network defined over 6 variables: Y_1, Y_2, Y_3 and X_1, X_2, X_3. This network is shown in Fig. 1, which shows that three of the variables (X_1, X_2, X_3) have observed values, with the other three (Y_1, Y_2, Y_3) being unobserved. Your **goal in this question is use variable elimination** to compute

$$p(Y_3 = \text{END} | X_1 = \texttt{a}, X_2 = \texttt{b}, X_3 = \#).$$

All of these values are described below.

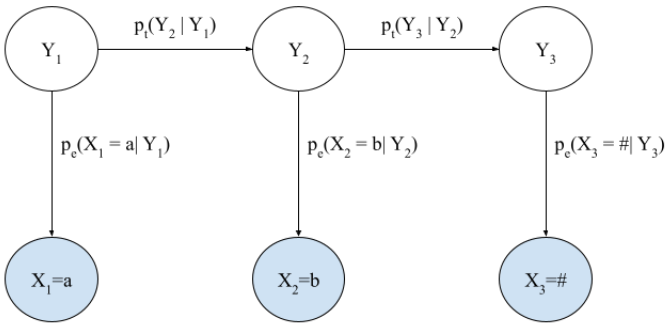

Figure 1: Bayesian network for question 3

Assume that the Y_i variables are all defined over three different values (Z_1, Z_2, and END), where the *same* conditional distribution $p_t(Y_i|Y_{i-1})$ is used to define the distribution of Y_i, given its parent Y_{i-1} (where there's a prior distribution over Y_1). Further assume that the X_i variables are all defined over three different values (a, b, and #), where the *same* conditional distribution $p_e(X_i|Y_i)$ is used to define the distribution of X_i given its parent Y_i. These

$Y_i =$	$X_i =$ a	b	#
Z_1	.7	.3	0
Z_2	.2	.8	0
END	0	0	1

(a) The definition of p_e, for $1 \leq i \leq 3$.

$Y_{i-1} =$	$Y_i =$ Z_1	Z_2	END
Z_1	.15	.8	.05
Z_2	.6	.35	.05
END	0	0	1

(b) The definition of p_t, for $i = 2, 3$.

$Y_1 =$ Z_1	Z_2	END
.7	.2	.1

(c) The prior distribution for Y_1.

Table 1: The distributions needed for question 3.

distributions are labeled (for ease) on the arcs of the figure, and are provided below in Table 1 (in both sub-tables, each row is a different distribution):

In order to compute $p(Y_3 = \text{END}|X_1 = \texttt{a}, X_2 = \texttt{b}, X_3 = \#)$, you'll need to eliminate both Y_1 and Y_2. To make the computations a bit easier, we'll start with Y_1 and then eliminiate to Y_2. Assume that the factors have been defined as given in Table 2:

Factor	Definition
$e_1(Y_1)$	$p_e(X_1 = \texttt{a}\|Y_1)$
$e_2(Y_2)$	$p_e(X_2 = \texttt{b}\|Y_2)$
$e_3(Y_3)$	$p_e(X_3 = \#\|Y_3)$
$t_1(Y_1)$	$p(Y_1)$
$t_2(Y_1, Y_2)$	$p_t(Y_2\|Y_1)$
$t_3(Y_2, Y_3)$	$p_t(Y_3\|Y_2)$

Table 2: Initial factor definitions.

(a) Eliminate Y_1. Call the resulting factor f.
   (i) Identify the factors that need to be considered to eliminate Y_1.
   (ii) Identify the variable(s) that f will be defined over.
   (iii) Write the formula for f (in terms of the factors you've identified).
   (iv) Compute f for each value (or values) of the variable(s) it is defined over.

(b) Eliminate Y_2. Call the resulting factor g.
   (i) Identify the factors that need to be considered to eliminate Y_2.
   (ii) Identify the variable(s) that g will be defined over.
   (iii) Write the formula for g (in terms of the factors you've identified).
   (iv) Compute g for each value (or values) of the variable(s) it is defined over.

(c) Answer the original question: compute the distribution $p(Y_3 = \text{END}|X_1 = \texttt{a}, X_2 = \texttt{b}, X_3 = \#)$.