CS760 Spring 2019 Homework 2

Due Mar 7 at 11:59pm

Name: Stewart Kerr

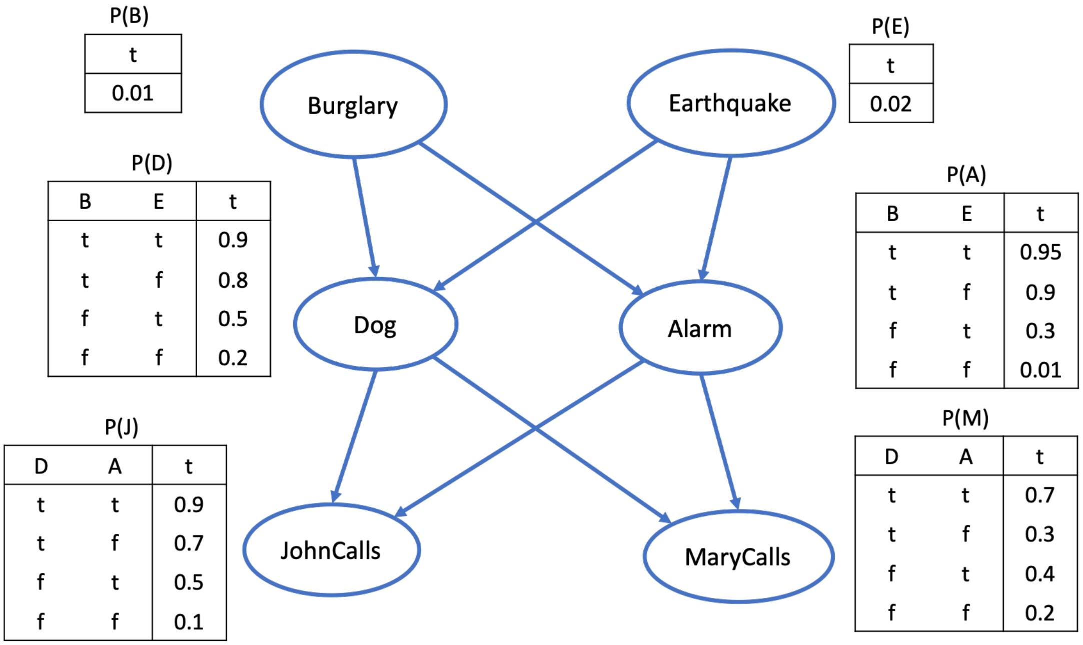
Email: shkerr@wisc.edu

Written Problems

NOTE: For the following written problems, put your answer in `hw2.pdf`. You are required to provide detailed solutions including the intermediate results for each step. Otherwise, you will not get full credit. You can also add figures or tables whenever necessary. If your solutions are handwritten, make sure they are legible.

1. (8 pts) Suppose you have a Bayesian network with 6 binary random variables shown as follows, where t and f stand for *true* and *false* respectively.

   Compute the probability: $P(d|b, \neg a, j, m)$.



P(B)

t
0.01

P(E)

t
0.02

P(D)

B	E	t
t	t	0.9
t	f	0.8
f	t	0.5
f	f	0.2

P(A)

B	E	t
t	t	0.95
t	f	0.9
f	t	0.3
f	f	0.01

P(J)

D	A	t
t	t	0.9
t	f	0.7
f	t	0.5
f	f	0.1

P(M)

D	A	t
t	t	0.7
t	f	0.3
f	t	0.4
f	f	0.2

From the formula of conditional probability, we know that $P(d|b, \neg a, j, m) = \frac{P(d,b,\neg a,j,m)}{P(b,\neg a,j,m)}$.
By the law of total probability, we can account for both earthquake outcomes and rewrite this fraction as

$$P(d|b, \neg a, j, m) = \frac{P(e, d, b, \neg a, j, m) + P(\neg e, d, b, \neg a, j, m)}{P(e, d, b, \neg a, j, m) + P(\neg e, d, b, \neg a, j, m) + P(e, \neg d, b, \neg a, j, m) + P(\neg e, \neg d, b, \neg a, j, m)}$$

Then, to get each probability in this sum, we just multiply the corresponding entries of the CPT (taking outcomes of dependent variables into account). Thus,

$$P(e, d, b, \neg a, j, m) = P(e) \times P(b) \times P(d|e, b) \times P(\neg a|e, b) \times P(j|d, \neg a) \times P(m, d, \neg a)$$

$$= 0.02 \times 0.01 \times 0.9 \times 0.05 \times 0.7 \times 0.3 = 1.89 \times 10^{-6}$$

$$P(\neg e, d, b, \neg a, j, m) = 0.98 \times 0.01 \times 0.8 \times 0.1 \times 0.7 \times 0.3 = 0.00165$$

$$P(e, \neg d, b, \neg a, j, m) = 0.02 \times 0.01 \times 0.1 \times 0.05 \times 0.1 \times 0.2 = 2 \times 10^{-8}$$

$$P(\neg e, \neg d, b, \neg a, j, m) = 0.98 \times 0.01 \times 0.2 \times 0.1 \times 0.1 \times 0.2 = 3.92 \times 10^{-6}$$

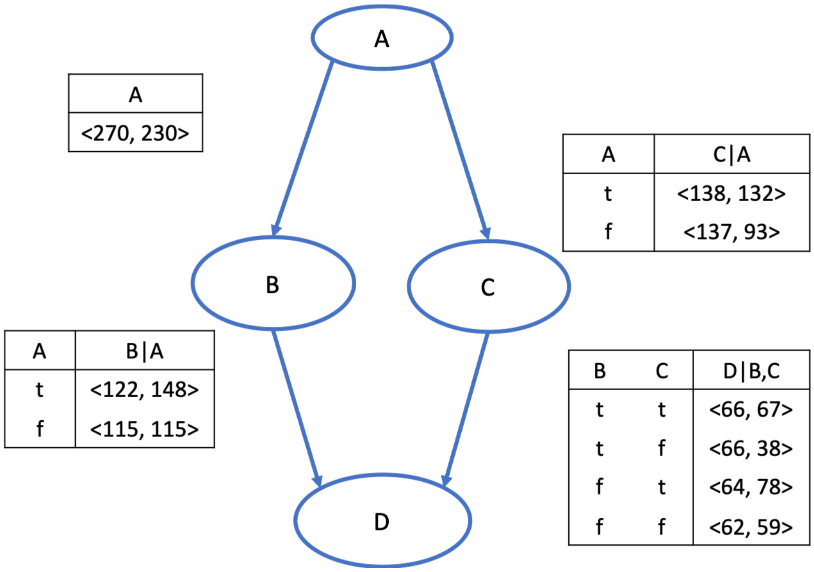
Now, we plug all of those numbers into our equation for $P(d|b, \neg a, j, m)$.

$$P(d|b, \neg a, j, m) = \frac{1.89 \times 10^{-6} + 0.00165}{1.89 \times 10^{-6} + 0.00165 + 2 \times 10^{-8} + 3.92 \times 10^{-6}} = 0.977$$

Intuitively, this makes sense. If both John and Mary call, but the alarm did *not* sound, then that places a higher "burden" on the dog barking to be the reason behind their calling. Thus, the probability that the dog barked will be higher.

2. Given the following Bayesian network and sample counts in each table, where sample counts <$n_{\texttt{true}}$, $n_{\texttt{false}}$>, there are $n_{\texttt{true}}$ samples with true labels and $n_{\texttt{false}}$ samples with false labels for this attribute. For example, `<138, 132>` in table `C|A` says given the condition of $A = true$, there are 138 instances are true and 132 are false with regard to attribute C.

   You need to answer the following two questions.



A
<270, 230>

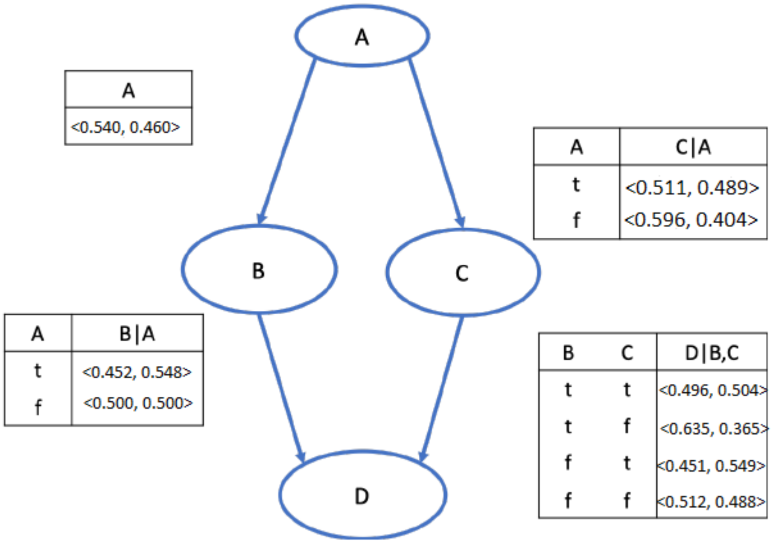
A	C|A
t	<138, 132>
f	<137, 93>

A	B|A
t	<122, 148>
f	<115, 115>

B	C	D|B,C
t	t	<66, 67>
t	f	<66, 38>
f	t	<64, 78>
f	f	<62, 59>

(a) (2 pts) Construct the conditional probability tables (CPTs) based on the above sample count tables, using maximum likelihood estimation. You need to both show the true probability $P_{\texttt{true}}$ and false probability P_{false} for each case, and organize them in the format of <$P_{\texttt{true}}$, $P_{\texttt{false}}$>. For example, for the case `Y|X`$_1$`,X`$_2$, your answer will look like `<P(Y|X`$_1$`,X`$_2$`), P(¬Y|X`$_1$`,X`$_2$`)>`. Keep **at least 3 digits of precision.** (You may reuse the same structure as the above tables, just plugging in the conditional probabilities in the place of sample counts. For more information, please refer to the lecture notes `BNs-1.pdf`)

See the figure below for the CPT tables.



A
<0.540, 0.460>

A	C|A
t	<0.511, 0.489>
f	<0.596, 0.404>

A	B|A
t	<0.452, 0.548>
f	<0.500, 0.500>

B	C	D|B,C
t	t	<0.496, 0.504>
t	f	<0.635, 0.365>
f	t	<0.451, 0.549>
f	f	<0.512, 0.488>

(b) (10 pts) Show the result of one cycle of the EM algorithm to update the CPTs you derived in step (a), using 10 another instances with `A=true`, `B=false`, `C=?`, and `D=true` ('?' means missing value). Keep **at least 2 digits of precision**.

The EM algorithm has two steps -
E) Using the current CPT, calculate expected values of the missing data.
M) Update the model using the imputed values from E which maximize probability of the data

Expectation step Given the CPT tables constructed above, we have

$$P(c|a, \neg b, d) = \frac{P(a, \neg b, c, d)}{P(a, \neg b, c, d) + P(a, \neg b, \neg c, d)} = \frac{0.540 \times 0.548 \times 0.511 \times 0.451}{(0.540 \times 0.548 \times 0.511 \times 0.451) + (0.540 \times 0.548 \times 0.489 \times 0.512)}$$

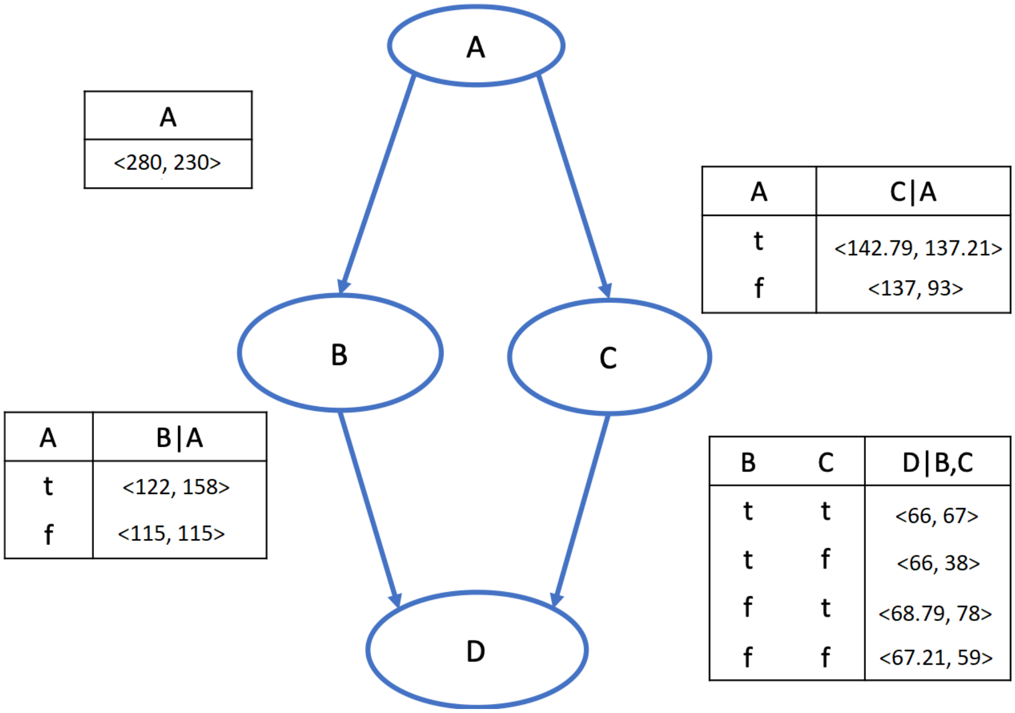
$$= 0.479$$

and

$$P(\neg c|a, \neg b, d) = 1 - P(c|a, \neg b, d) = 1 - 0.479 = 0.521$$

Thus, the expected counts for c and $\neg c$ using another 10 instances are $10 \times 0.479 = 4.79$ and $10 \times 0.521 = 5.21$ respectively.

Maximization step In the maximization step, we add the new data (with expected counts) to the previous counts and then recalculate the CPT table. After adding the new data, the counts table is:



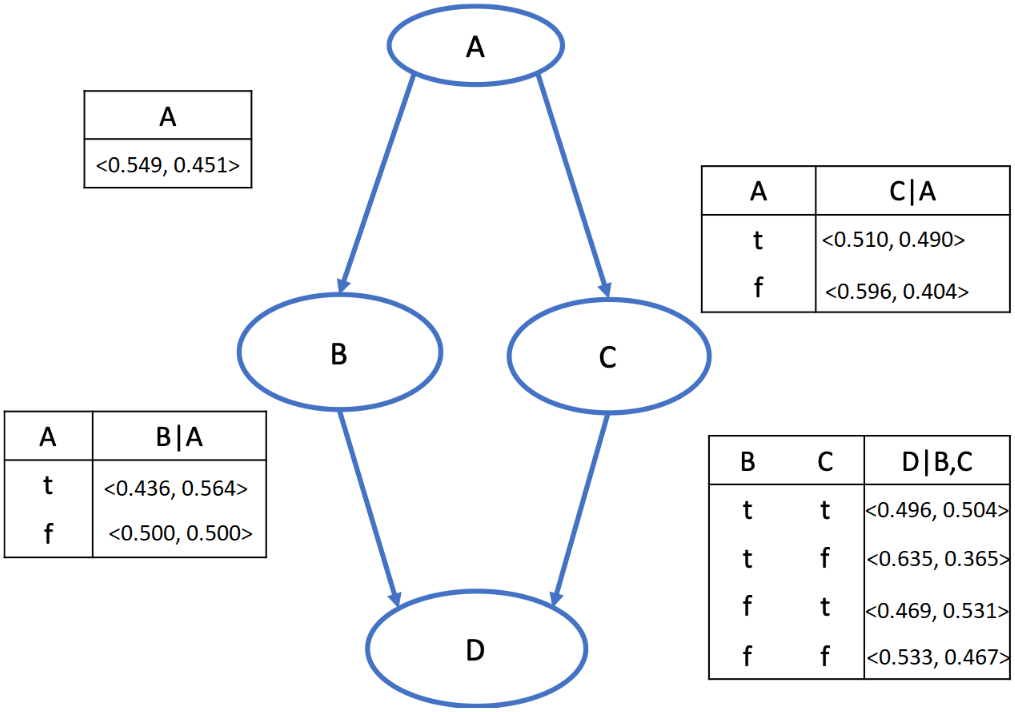
A
<280, 230>

A	C|A
t	<142.79, 137.21>
f	<137, 93>

A	B|A
t	<122, 158>
f	<115, 115>

B	C	D|B,C
t	t	<66, 67>
t	f	<66, 38>
f	t	<68.79, 78>
f	f	<67.21, 59>

Thus, the new CPT table is:



A
<0.549, 0.451>

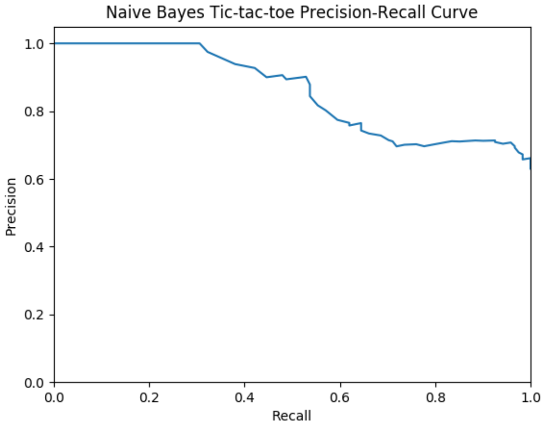
A	C|A
t	<0.510, 0.490>
f	<0.596, 0.404>

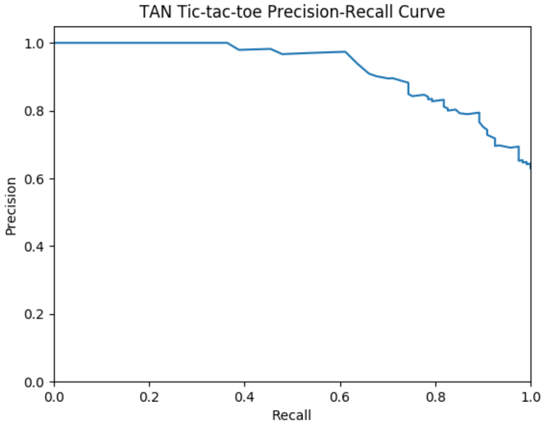
A	B|A
t	<0.436, 0.564>
f	<0.500, 0.500>

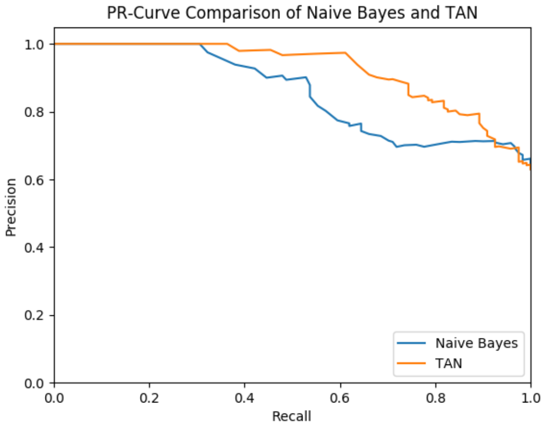
B	C	D|B,C
t	t	<0.496, 0.504>
t	f	<0.635, 0.365>
f	t	<0.469, 0.531>
f	f	<0.533, 0.467>

3. (15 pts) Plot a precision/recall curve for both methods (NB and TAN) and answer the following question:

   (a) Compare the two curves, and make a comment about which method (NB or TAN) seems to have more predictive power. Explain why you think that (i.e. what features of the precision/recall curve lead you to this conclusion?)
   I will include a plot of each method and then a combined plot of both methods. Note that I used a step method with 100 steps between a decision threshold of 0.00 and 1.00 to make my PR curves. This yields a PR-curve with slightly less resolution but it was quicker for me to construct.


Naive Bayes Tic-tac-toe Precision-Recall Curve


TAN Tic-tac-toe Precision-Recall Curve


PR-Curve Comparison of Naive Bayes and TAN

If we focus on the third plot, which is a comparison of the two methods, we see that aside from decision thresholds that have high recall - (that is a low probability decision threshold) - the TAN method has better precision for a given recall. In other words, the TAN PR curve is generally higher than the NB PR curve. This means that if TAN and NB predict the same number of positive classes in a test data set, then the percent of those positive predictions that are correct is generally higher for TAN than NB. **For this reason, I will say that TAN has a higher predictive power for this data set.**

4. (15 pts) Using the given data set named tic-tac-toe.json, use 10-fold cross validation to obtain 10 accuracy measures for each method. You'll notice that tic-tac-toe.json is simply a concatenation of its given train and test files. Use these accuracies to conduct your paired t test and discover whether you accept or reject the alternative hypothesis (that the classifiers truly differ in accuracy). Specifically, calculate the accuracy deltas for each cross validation fold and report the following values/answers:

   (a) Calculate the sample mean

   For this paired t-test, we are interested in examining the difference between the naive Bayes method and the tree-augmented naive Bayes method. Specifically, we are interested in the following hypothesis:

$$H_0 : \bar{D}_{TAN-NB} = 0$$

$$H_A : \bar{D}_{TAN-NB} \neq 0$$

   Where $\bar{D}$ represents the average difference between the TAN and NB methods using the same training and test sets.

   To calculate the average difference, I have divided the data set tic-tac-toe into 10 folds. I will cycle through the 10 folds choosing one fold to be the test set and the remaining folds to be the training set. With each iteration, I will use both NB and TAN to construct the Bayes network and then use that network to predict classes for the observations in the training set. For each iteration, the accuracy (that is number of correct predictions divided by the total number of predictions) for NB and TAN will be compared. The accuracies I observed using this method are summarised in the table below.

Test Fold Number	NB	TAN	Difference (TAN - NB)
1	0.719	0.802	0.083
2	0.854	0.833	-0.02
3	0.854	0.896	0.042
4	0.844	0.781	-0.063
5	0.708	0.906	0.198
6	0.917	0.906	-0.010
7	0.583	0.792	0.208
8	0.281	0.573	0.292
9	0.463	0.632	0.168
10	0.316	0.589	0.274

   Thus, the average difference (TAN - NB) is:

$$\frac{\sum_{i=1,\ldots,10} \text{Difference}_i}{10} = 0.117$$

   (b) Calculate the t statistic

   The t-test statistic is defined as $t = \frac{\bar{x}}{SE_x}$ where $SE_x = \frac{SD_x}{\sqrt{n}}$

   We used 10-fold cross validation, thus our number of differences is n = 10. Then, we can use Numpy to calculate the standard deviation, SD = 0.121. Therefore, the standard error (SE) is calculated to be:

$$SE = \frac{0.121}{\sqrt{n}} = 0.0383$$

   Thus,

$$t = \frac{\bar{x}}{SE_x} = \frac{0.117}{0.383} = 3.06$$

(c) Determine the corresponding p-value for a two-tailed t-test by looking up t in a t-table with n-1 degrees of freedom. Use a threshold of p = 0.05 when determining if it is significant or not. Rather than look up the t-statistic in a t-table, I will use scipy.stat's t module to directly calculate the p-value. For this test, we have n-1 = 10-1 = 9 degrees of freedom. Thus, the p-value for the two-tailed paired t-test is:

$$p = 2 \times P(t_9 \geq 3.06) = 0.0137$$

This p-value is lower than our threshold of $\alpha = 0.05$, so we will reject the null hypothesis that there is no difference between the two methods. There is strong evidence to suggest that the alternative hypothesis - that the two methods give different prediction accuracies - is true.