Artificial Intelligence
Homework 5 [14 pts, Due April 23 2024]

[Homework solutions must be submitted through Canvas. Only pdf, word, and txt files are allowed. If you have multiple pictures, please include all pictures in one Word/pdf file. You can always update your submissions before due date, but only the latest version will be graded.]

Question 1 [2 pts]: Table 1 shows probability values of different events. Using the table to calculate following values and show proof:

- The probability that a persona has no cavity [0.25 pt]
- The probability of no toothache [0.25 pt]
- The joint probability of cavity and no toothache [0.25 pt]
- Calculate conditional probability of no cavity, given the patient has toothache [0.25 pt]
- Calculate conditional probability of no cavity, given the patient does not have toothache [0.25 pt]
- Determine whether cavity and toothache are independent or not, why [0.25 pt]
- Given a patient has cavity, determine whether the tooth probe catch is conditionally independent of toothache or not, why [0.25 pt]
- Given a patient does not have cavity, determine whether the tooth probe catch is conditionally independent of toothache or not, why [0.25 pt]

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Table 1

Question 2 [1 pt]: Figure 1 shows a Bayesian network with three nodes. Prove that X and Z are conditionally independent, given Y (show your work).

Figure 1

Question 3: [1 pt]: Figure 2 shows a Bayesian network, and the first letter denotes each named variable, e.g., T denotes tampering. Assume x ⊥ y denotes that x are independent of y, x ⊥ y | z denotes that x and y are conditionally independent, given z. Complete Table 2, and use √ to mark correct answers. [1 pt]

	True	False
T ⊥ F		
A ⊥ S		
T ⊥ S		
$R \perp A$		
$R \perp A \mid L$		
$L \perp S$		
$L \perp S \mid F$		
$S \perp A \mid F$		
$R \perp S \mid L$		

Table 2

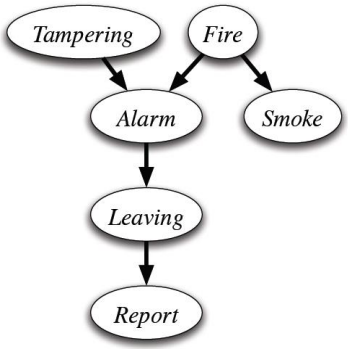


Figure 2

Question 4: [2 pts]: Figure 3 shows a Bayesian network where *r* denotes "rain", *s* denotes "sprinkler", and *w* denotes "wet lawn" (each variable takes binary values 1 or 0). The prior probabilities of rain and sprinkler, and the conditional probabilities values are given as follows:
p(r = 1) = 0.10
p(s = 1) = 0.20
p(w = 1|r = 0, s = 0) = 0.001
p(w = 1|r = 0, s = 1) = 0.97
p(w = 1|r = 1, s = 0) = 0.90
p(w = 1|r = 1, s = 1) = 0.99

- Show joint probability value formula of the whole network, and calculate the joint probability value of P(r=1, s=1, w=1). [0.5 pt]
- Calculate overall probability of lawn is wet, ie., P(w=1) [0.5 pt]
- After observing the law is wet, calculate the probability that the sprinkler was left off (i.e., s=0). [0.5 pt]
- After observing the law is wet, please calculate the probability that there was rain (i.e., r=1). [0.5 pt]

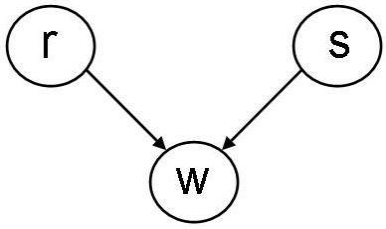


Figure 3

Question 5 [2 pts]: Given the following toy dataset with 16 Instances

- Please manually construct a Naïve Bayes Classifier (list the major steps, including the values of the priori probability [0.5 pt] and the conditional probabilities [1.0 pt]. Use *m*-estimate to calculate the conditional probabilities (*m*=1, and *p* equals to 1 divided by the number of attribute values for each attribute).
- Please use your Naïve Bayes classifier to determine whether a person should play tennis or not, under conditions that "Outlook=Overcast & Temperature=Hot & Humidity =Normal& Wind=Weak". [0.5 pt]

ID	Outlook	Temperature	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Mild	Normal	Weak	No
14	Rain	Hot	High	Strong	Yes
15	Rain	Mild	High	Strong	No
16	Overcast	Cool	Normal	Weak	No

Question 6 [2 pts]: Figure 4 shows a Bayesian network which is similar to Figure 3, but w_1 denotes your lawn, and w_2 denotes neighbor's lawn (each variable takes binary values 1 or 0). In this case, rain will cause both yours and your neighbor's lawn being wet, whereas your sprinkler will only cause your lawn to be wet. The prior probabilities and conditional probability values are given as follows:

$p(r = 1) = 0.10$
$p(s = 1) = 0.20$
$p(w_1 = 1|r = 0, s = 0) = 0.001$
$p(w_1 = 1|r = 0, s = 1) = 0.97$
$p(w_1 = 1|r = 1, s = 0) = 0.90$
$p(w_1 = 1|r = 1, s = 1) = 0.99$

$p(w_2 = 1|r = 1) = 0.90$
$p(w_2 = 1|r = 0) = 0.1$

- Show joint probability value formula of the whole network, and calculate the joint probability value of P(r=1, s=1, w_1=1, w_2=1). [0.5 pt]
- Calculate overall probability of yours and your neighbor's lawn are wet, i.e., P(w_1=1, w_2=1) [0.5 pt]
- After observing that yours and your neighbors' lawn are both wet, calculate the probability that the sprinkler was left on (i.e., s=1). [0.5 pt]
- After observing that yours and your neighbors' lawn are both wet, calculate the probability that there was rain (i.e., r=1). [0.5 pt]

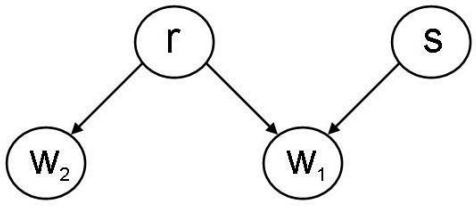


Figure 4

For all programming tasks, please submit the Notebook as an html notebook file or a pdf notebook file for grading (your submission must show scripts/code and the running results of the script). The code running results must reflect/answer the task requirements.

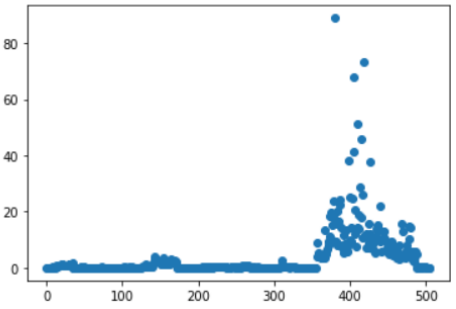
For each subtask, please use task description (requirement) as comments, and report your coding and results in following format:

```
# Report all samples with respect to the Crim index on a plot (the x-axis shows the index of the sample, and the y-axis
# shows the crim index of the sample).

y = boston['Crim']
x=np.arange(y.shape[0]) # generate x index

plt.scatter(x, y, marker='o')
```

```
: <matplotlib.collections.PathCollection at 0x1c0d3ceebc8>
```



Question 7 [4 pts]: [**NB for Text Classification [Notebook, html]**] notebook shows how to design naïve bayes classification for email spam classification.

- Report class conditional probabilities of the top-10 most frequent words, with respect to each class (either as tables or plots) [0.5 pt]
- For each class, report the top-10 words with the highest class conditional probabilities (either as tables or plots) [0.5 pt]
- Implement a *k*-fold cross-validation framework and report 5-fold cross validation results for email spam classification [1 pt]
- Implement an algorithm which randomly samples *x*% of data as training set, and the rest data are the test set. Repeat the algorithm for 5 times and repeat the mean accuracy and F1 scores:
    - Vary *x* values from 5%, 10%, 30%, 50%, to 80% and report the mean accuracy and F1 scores (using 5 time repetitions) [1.5 pt]
    - Compare performance with respect to different *x* values, and explain how does NB perform with increasing size of training set [0.5 pt]