EECS 492: Introduction to AI
Homework 5 (100 pts)

General Information

Due: 11:59 PM, November 22, 2016
Notes:

- Put your answers in a pdf (with your name in the pdf) and submit to the corresponding assignment on Gradescope.
- Late homework will be penalized 10% per day (each day starts at 11:59 PM on the due day).
- Homework turned in after **three** late days **will not be accepted**.

Exercise 1

[Probability – 20 pts]

1. [5 pts] Answer the following questions assuming that there are *n* persons in a room and 365 days in a year. Show all work. You may use a computer or calculator to get the numerical probabilities, but show the expression you typed in.

    a. What is the probability that at least two persons have the same birthday?
    b. Calculate this probability for $n = 50$.
    c. How large need *n* be for this probability to be greater than 0.5?

2. [5 pts] A manufacturer produces batches of 100 hard drives. In a given batch, there are 20 hard drives that are defective. Quality control selects two hard drives to test at random, without replacement, from the batch.

    a. What is the probability that the first hard drive selected is defective?
    b. What is the probability that the second hard drive selected is defective given that the first hard drive was defective?
    c. What is the probability that both hard drives are defective?

3. [5 pts] Assume there are three variables: A, B, C. Each variable is binary (e.g., A=a or A=~a). Use the following table to calculate P(A,B) for all values of A and B (list these values). What is P(A)?

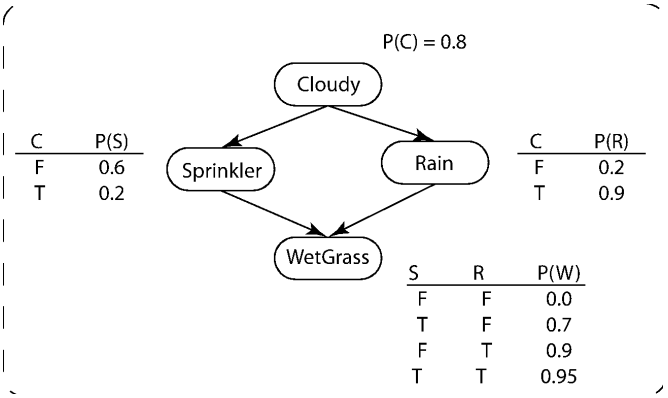
Probability(A,B,C)			
A	B	P(A,B,c)	P(A,B,~c)
T	T	0.25	0.15
T	F	0.05	0.1

F	T	0.05	0.2
F	F	0.1	0.1

4. [5 pts] Tomorrow is Game Day. Scientists have demonstrated a strange phenomenon relating solar flares to the outcome of the game. When the number of flares per day is over a specific threshold, team A tends to win (very strange!). Luckily, for team B, this threshold is only exceeded about 6 days a year. We will call this event, 'SFOT' (solar flares over threshold). Unluckily for team B, scientists believe that tomorrow we will see an SFOT. Scientists are pretty good at predicting this. When there is an SFOT, scientists correctly predict an SFOT 95% of the time. When there is no SFOT in reality, scientists still predict an SFOT 15% of the time. If scientists predict there will be an SFOT tomorrow, what is the probability there will be an actual SFOT tomorrow during the big game?

Exercise 2

[Calculating the probabilities of a Bayesian Network – 34 pts]



Let's use this Bayesian Network to calculate the probability of a set of events. In this assignment we will use shorthand to describe the various events. In this problem, w represents "WetGrass is true," while ~w represents "WetGrass is false." Remember, W (capital letter) is the variable! Show all work for full credit.

1. [2 pts] P(w, s, ~r, c)
2. [10 pts] P(r|~w)
3. [10 pts] P(~s|w)
4. [10 pts] P(~s|w, r)
5. [2 pts] What do you notice about P(~s|w, r) vs. P(~s|w)? What is the reason for this difference?

Exercise 3

[Variable Elimination in Bayesian Networks – 18 pts]
In class we covered variable elimination. We discussed how every ordering choice results in a valid algorithm (you get the same probabilities). However, different orderings result in different factor generations. Use variable elimination to solve the following conditional probability given the burglary problem Refer to Figure 14.2 in text, also in slides.

$$P(J \mid b) = \alpha P(b) \sum_{e} P(e) \sum_{a} P(a \mid b, e) P(J \mid a) \sum_{m} P(m \mid a)$$

1. [2 pt] In the above expression, what quantity does α represent?
2. [4 pts] Write the factorized expression for P(J | b) after eliminating variable M. Why can we leave out $\sum_{m} P(m \mid a)$?
3. [4 pts] Write the conditional probability table for variable A after eliminating variable E.
4. [4 pts] Write the conditional probability table for J after eliminating A.
5. [4 pts] Eliminate variable B and write the probability table for J. What do these probabilities represent?

Exercise 4

[Building a Bayesian Network – 28 pts]
In this problem we will build a Bayesian Network that seeks to understand runners and joggers. Runner's World conducted an informal poll and there were 10,000 respondents. They found that 2,000 people considered themselves runners (as opposed to the 8,000 people who considered themselves joggers). Interestingly, they found whether or not someone was a runner affected their preference for running on roads or trails. Specifically, they found that:

- 1080 respondents were runners and liked to run on both the road and trails
- 720 respondents were runners and liked to run on roads, but not on trails
- 120 respondents were runners and liked to run on trails, but not on roads
- 960 respondents were joggers and liked to run on both the road and the trails
- 1440 respondents were joggers and liked to run on the roads, but not on the trails
- 2240 respondents were joggers and liked to run on the trails, but not on the roads

Runner's World is also interested in the Tough Mudder races (http://toughmudder.com/). They wanted to know who was running these things. They found that the chance of running a Tough Mudder race was dependent on whether people like to run on roads and/or trails. They found that the respondents could be further broken down:

- 1,680 had run a Tough Mudder and liked to run on both the road and trails
- 420 had run a Tough Mudder liked to run on roads, but not on trails
- 800 had run a Tough Mudder liked to run on trails, but not on roads
- 420 had run a Tough Mudder liked to run on neither roads nor trails

There is also another kind of race that people enjoy, ultramarathons (http://en.wikipedia.org/wiki/Ultramarathon). Runner's World believes that you can determine if people will enjoy ultramarathons based on whether they enjoy trail running and whether they recently lost a bet (friends can be cruel). Luckily, only 500 people reported losing a bet that required them to start running ultramarathons. Specifically the chance of running an ultramarathon is:

- 50% – if people like to run on trails and lost a bet.
- 30% – if people like running on trails and did not lose a bet.
- 5% – if people do not like running on trails and lost a bet.
- 1% – if people do not like running on trails and did not lose a bet.

Runner's World would like to get some more insight. It would like you to do the following:

1. [2 pts] Draw a bayesian network based on this data and the independence and conditional independence relationships described. When constructing the network, think about the probabilities that can (and cannot) be inferred from the data above. If a conditional probability between two variables cannot be calculated given the data, then there is no link between the two variables.
2. [2 pts] Write out the factorized expression for the full joint distribution, P(*N, R, T, M, U, B*), using the conditional independence relationships described by the network.
3. [12 pts] Show the conditional probability tables for each node.
4. [12 pts] Calculate the marginal probabilities for each node.

For consistency, use the following naming conventions (N is true if a person is a runner and false if a person is a jogger):

- N = Runner/Jogger
- R = RunOnRoad
- T = RunOnTrail
- M = ToughMudder
- U = Ultramarathon
- B = LostBet