INFO 4220: Homework 1

Due by: Thursday, Feb 7 2019, 12 noon

General instructions:

1. (a) Don't panic. (b) This homework is Mostly Harmless.
2. Finish and turn in the homework, on CMS, *well in advance of the deadline.*
3. If (Not 2.), *i.e.*, if you fail to turn in your homework in time:
   (a) See the course homework policy (Lecture 1; CMS handout).
   (b) Realize that the policy does *not* say "email instructor asking for extension". In fact, it suggests that you should perhaps *not* do this, unless you have a Most Excellent Reason (these are *very* hard to come by, so you're statistically *very* unlikely to have one).
   (c) Let realization in 3(b) sink in.
   (d) Make peace with your situation (refering to 1(a) as often as necessary).
   (e) Have a great rest-of-day. You might even want to whistle a jolly tune!

---

1. (*Abstract modeling; basic probability.*) A lot of commerce on the Internet is constructed around ratings and reviews, which allow users to convey their experience with sellers, or websites, or what-have-you to potential future users. In this problem, we're going to try to understand what ratings might tell us, using basic probability.

   As a specific example, there are many websites which allow patients to rate their experience with doctors. Suppose doctors are either good or bad. If the doctor is actually bad, there is a 96% chance that a patient will have a negative experience and rate the doctor as bad. However, even when the doctor is good, there is a 2% chance that the patient will have a negative experience and rate the doctor as bad.

(a) (5pts) Suppose the doctors on a website are known to be mostly good: Only 5% of all registered practitioners on the website are bad. If a patient rates a doctor as bad, what is the probability the doctor is actually bad? Now suppose a doctor gets 2 bad ratings instead of 1—what is the probability the doctor is actually bad?

(b) (5pts) Suppose the website is instead full of second-rate doctors: 60% of all registered practitioners on the website are bad. Now what is the probability the doctor is actually bad if a single patient rates a doctor as bad?

Note: Did you find it surprising how different the answers were, depending on the *prior*? A well-known fallacy related to priors in Bayesian updating is the *base rate fallacy*, where people, roughly speaking, overweight the data and underweight the prior in forming their estimates of posterior probabilities—a fun tidbit for you to look up if you're curious!

2. (*Logical reasoning: Converses and contrapositives.*) Technology is fantastically useful, but is also being linked with a number of health issues. One specific effect that is widely discussed is the effect of technology on sleep. Now we're interested in logical reasoning—so we're just going to play around with what *inferences* you can make, if you take some claims of this kind as given.
Note: Here, and throughout this course, to say a statement is true means it holds for all instances (here, individuals) it addresses. However, saying it is false means it *need not* hold for all individuals—not that it must be violated for every individual! (In other words, saying a statement is false does not mean it needs to be false for every instance; it could indeed be accurate for some instances: For a statement to be false, it only needs to be false for at least one instance. However, to say a statement is true, it must be true for all instances.)

(a) (*15 pts.*) Consider the specific claim: "Individuals who use devices with screens emitting blue light around bedtime have trouble with sleep". Suppose you trust this claim, *i.e.*, believe it is true, and have no other (authoritative) information about the subject. Which of the three statements below can you also infer to be true, given this (*i.e.*, the truth of this claim)—and *only* this? You must explain your answer as to why or why not for each statement *separately*, using the language of converses and contrapositives.

i. An individual who has sleep problems must be using blue-light devices around bedtime.

ii. An individual who does not have sleep problems must not be using blue-light devices around bedtime.

iii. An individual who does not use blue-light devices around bedtime will not have sleep problems.

(b) (*20 pts.*) Suppose that a curious student, Carla, wants to investigate for herself whether this claim—that using blue-light devices right before bedtime disrupts sleep—is true or false. Carla has four good friends, who all own such devices, and all of whom she can ask pesky questions. She knows that David is an avid user of such devices at all times, including right before bed, while Jessica does not use devices much at all, least of all late in the day. She also knows that Kim has no sleep issues, while Ethan struggles to fall asleep, counting vast herds of sheep each night. Which of these friends could possibly help Carla reach a conclusion about whether this simpler claim is accurate or not?

Specify what conclusion can be drawn, by asking which individual(s) and what question he needs to ask each individual, as well as what the answer(s) needs to be to draw that conclusion; also specify which individual(s)' answers cannot be useful towards drawing any conclusion.

3. (*Pareto-efficiency, in real life.*) The notion of Pareto optimality, or Pareto efficiency, is a very general concept that is used to compare alternatives when alternatives have *multiple attributes* (in this case, two) that determine their 'goodness'. You'll study Pareto-efficiency in class in the context of comparing *matchings*, but here we'll investigate how it might show up in less obviously academic settings as well.

Suppose you are deciding on which mobile chat application to use. You have a number of choices, and the two factors that most influence your choice are (i) the number of your friends who are using that app—the more, the better, and (ii) usability, or ease of use—how well-designed the app interface is, how quickly it loads on your phone, and so on; say that for every app you're considering, you've assigned a score to summarize its usability, with higher scores meaning a more usable app than a lower score.

We'll say that

I. App X *Pareto-dominates* App Y if and only if App X (i) is at least as good as App Y on *both* of the two dimensions (number of friends, and usability), **and** (ii) is strictly better than App Y on at least one of the two dimensions, and

II. An app is *Pareto-efficient* or *Pareto-optimal* if no other app Pareto-dominates it.

Suppose you are considering 7 different apps, with the scores on each of the two attributes you care about listed in the table below.

Let's investigate the 7 choices in the language of Pareto-dominance. Be sure to read each statement *carefully* before answering, and also to provide brief explanations for your yes/no answers to (a), (b) and (c).

App	Number of friends	Usability score
A	64	1
B	53	9
C	282	8
D	395	5
E	84	6
F	152	4
G	279	3

(a) (*10 pts.*) Does App G Pareto-dominate B? Does B Pareto-dominate G?

(b) (*10 pts.*) List all the apps (if any) that are Pareto-dominated by G. Is G Pareto-optimal?

(c) (*10 pts.*) Is App E Pareto-optimal? Does it Pareto-dominate any app?

(d) (*5 pts.*) Are there any Pareto-optimal choices amongst these 7 apps? List all the apps you think are Pareto-optimal, if any.

4. *(Matchings in bipartite graphs.)* Recall that a constricted set in a bipartite graph is a set of nodes S such that $|N(S)| < |S|$; for simplicity, we will only consider constricted sets S that are a subset of the RHS nodes (*i.e.*, the set of students). Consider a bipartite graph with n nodes on each side that contains no perfect matching.

   For each question below, you must supply a proof if you answer true (or yes), and a counterexample if you say it is false (or no).

   (a) (*10pts.*) True or false: Suppose such a bipartite graph (*i.e.*, with no perfect matching) has more than one maximum matching. Then, there must be more than one constricted set.

   (b) Consider bipartite graphs where the maximum size of a matching in the graph is strictly smaller than $n-1$, *i.e.*, every maximum matching leaves at least two students (and a corresponding number of rooms, of course) unmatched.

   i. (*5 pts.*) Is each student who is unmatched in a maximum matching sure to belong to some constricted set of RHS nodes?

   ii. (*5 pts.*) Consider any maximum matching in such a graph. Does an ABFS (Alternating Breadth First Search) starting from an unmatched RHS node on the RHS always yield a constricted set containing **all** unmatched RHS nodes (*i.e.*, for all possible starting points, in all possible maximum matchings, in all such graphs)?