Homework on Probability

Benjamin Cosman, Patrick Lin and Mahesh Viswanathan

Fall 2020

Problem 1. Consider a simplified version of the game of *Blackjack*. In this game, a standard 52 card deck is used.[1] A set of two cards is chosen uniformly at random, and the player wins if the total points from the two cards is exactly 21, under the following scoring system: the point-value of each of the *number cards* (i.e., the cards labeled 2 through 10) is just the number of the card; the Jack, Queen, and King are each worth 10 points, and the Ace is worth either 1 or 11 points (player's choice). What is the probability of the player winning? (You do not have to simplify your answer.)

[1] See `https://en.wikipedia.org/wiki/Standard_52-card_deck` if you need to remind yourself what the cards in a 52 card deck are.

Problem 2. Prove the following conditional versions of various Probability rules:

a) $\sum_{x \in S} \Pr[x|B] = 1$ (S is the entire sample space).

b) For disjoint $E_1, \ldots, E_n$, $\Pr\left[\bigcup_{i=1}^{n} E_i | B\right] = \sum_{i=1}^{n} \Pr[E_i|B]$.

c) $\Pr[A \setminus B|C] = \Pr[A|C] - \Pr[A \cap B|C]$.

d) $\Pr\left[\bigcup_{i=1}^{n} E_i | B\right] \leq \sum_{i=1}^{n} \Pr[E_i|B]$.

e) If $A \subseteq B$ then $\Pr[A|C] \leq \Pr[B|C]$.

Problem 3. An important task when designing networks is to ensure that it is robust to random network failures. For example, if a small failure randomly occurs, ideally the probability that the network becomes disconnected should be low.[2] While K_n is obviously the most robust against such failures, building connections is expensive, so we want to get away with significantly fewer edges. In each network and failure scenario below, what is the probability that the given failure disconnects the network?

[2] Recall that a graph is connected if there exists a walk between every pair of vertices, and disconnected otherwise.

a) The network is shaped like C_n, where $n > 3$. A set of *two edges*, chosen uniformly at random, simultaneously fails.

b) The network is shaped like W_n, where $n > 3$. A set of *two edges*, chosen uniformly at random, simultaneously fails.

c) The network is shaped like W_n, where $n > 3$. A set of *three edges*, chosen uniformly at random, simultaneously fails.

d) The network is shaped like C_n, where $n > 3$. A set of *two vertices*, chosen uniformly at random, simultaneously fails.[3]

e) The network is shaped like W_n, where $n > 3$. A set of *two vertices*, chosen uniformly at random, simultaneously fails.

f) The network is shaped like W_n, where $n > 3$. A set of *three vertices*, chosen uniformly at random, simultaneously fails.

[3] Here and below: when a vertex fails, any links to that vertex become unusable as well: effectively the vertex as well as all edges incident to it are removed from the graph.

Problem 4. A classic application of Bayes' rule is that of a (Naïve) Bayesian Spam Filter for eliminating spam from email.[4] In the setup, we identify a set G of Good$^{\text{TM}}$ emails (aka not-Spam) and a set B of Bad$^{\text{TM}}$ emails (aka Spam). For a given word w, let $\#_G(w)$ and $\#_B(w)$ be the number of in Good$^{\text{TM}}$ and Bad$^{\text{TM}}$ emails, respectively, that containt he word w. We will set $p_G(w) = \frac{\#_G(w)}{|G|}$ and $p_B = \frac{\#_B(w)}{|B|}$, the (empirical) probabilities of seeing w in a Good$^{\text{TM}}$ or Bad$^{\text{TM}}$ email, respectively.

[4] Modern Spam Filters are based on much more sophisticated methods, but it is still nice to see how the concepts we have learned apply to a rudimentary version of such a system.

Given a new email, the goal of the Bayesian Spam Filter is to decide, based on the words appearing in the email, whether the email is Spam or not. The methodology is as follows: Let S be the event that the email is Spam, and W be the event that the email contains word w.

a) Assume that $\Pr[S] = \Pr[\bar{S}] = \frac{1}{2}$. Show that under this assumption, $\Pr[S|W] = \frac{\Pr[W|S]}{\Pr[W|\bar{S}]+\Pr[W|S]}$.

Absent any information, $\Pr[S] = \Pr[\bar{S}] = \frac{1}{2}$ is a fairly standard *prior* to assume. Since $p_G(w)$ and $p_B(w)$ are empirical estimates of $\Pr[W|\bar{S}]$ and $\Pr[W|S]$, respectively, we can substitute these values in to an estimate $p_S(w)$ of $P[S|W]$ by $p_S(w) = \frac{p_B(w)}{p_G(w)+p_B(w)}$. We then decide on a threshold θ and declare that if $p_S(w) \geq \theta$, any email containing w is marked is Bad$^{\text{TM}}$.

b) The University hires a few undergraduate students to classify sample emails sent to University email addresses as either Good$^{\text{TM}}$ or Bad$^{\text{TM}}$, and then trained a Bayesian Spam Filter based on their classification. Later, the administration found that the word `MASSMAIL` appeared in 25 of 200 emails classified as Bad$^{\text{TM}}$, and only 1 out of 100 emails classified as Good$^{\text{TM}}$. Obviously, it would not be good if `MASSMAIL`s sent from the University got sent to students' Spam folders. For what values of θ would such emails be kept from being flagged as Bad$^{\text{TM}}$?