CS109 Problem Set #2

Andrew Han – handrew@stanford.edu

January 25, 2014

1. Let E and F be events defined on the same sample space S. Prove that $P(EF) \geq P(E) + P(F) - 1$.

$$P(E) + P(F) - P(EF) = P(E \cup F) \leq 1$$
$$P(E) + P(F) - P(EF) \leq 1$$
$$P(EF) \geq P(E) + P(F) - 1$$

Note that the first step to the second step works because E and F are defined on the sample space, so the union of the two will be less than or equal to 1.

2. Say in Silicon Valley, 34% of engineers program in Java and 26% of the engineers who program in Java also program in C++. Furthermore, 39% of engineers program in C++.

   (a) What is the probability that a randomly selected engineer programs in Java and C++?

$$P(JC) = 0.34 * 0.26 = \mathbf{0.0884}$$

   (b) What is the conditional probability that a randomly selected engineer programs in Java given that he/she programs in C++?

$$P(J|C) = P(JC)/P(C)$$
$$0.084/0.39 \approx 0.2158$$

3. Your spam filter encounters a bug 20% of the time in which it always marks an email GOOD. We let p be the probability that an email is GOOD and q be the conditional probability that an email is GOOD given that the filter marked it GOOD.

   (a) q in terms of p:

$$p = P(good)$$
$$q = P(good|passed)$$

Bayes:

$$P(E) = P(E|F) * P(F) + P(E|(F^C)P(F^C)))$$

So

$$p = q * P(passed) + P(good|failed)P(failed)$$

Since $P(good|failed)$ is not possible, then

$$p = q * P(passed)$$

P(passed) $= 0.2 + p * (1 - 0.2)$ because 20% of the time, it will always pass, regardless of whether or not the email was non-spam. The remaining 80% of the time, it will pass only if the email is good.

$$q = p/(0.2 + p * (0.8))$$

(b) Determine whether p or q is greater.

According to the first part, we can theorize that q is greater, since $(0.2 + 0.8p)$ will always be less than 1, so $p/(0.2 + 0.8p)$ will always be greater than p. Intuitively, by the "explaining away" phenomenon, we know that q is greater than p because if the filter marked it good, then 80% of the time it will be because the email is actually good, and in 20% of the time, it will be because the filter was buggy.

4. Say all computers either run operating system W or X. A computer running operating system W is twice as likely to get infected with a virus as a computer running operating system X. If 73% of all computers are running operating system W, what percentage of computers infected with a virus are running operating system W?

$$P(W) = 0.73$$
$$P(V|W^C) = p$$
$$P(V|W) = 2p$$

So we know that

$$P(V) = P(V|W^C) + P(V|W) = 3p$$

So we need to find $P(W|V)$:

$$P(W|V) = P(WV)/P(V)$$

And:

$$P(VW) = P(V|W) * P(W)$$

So.

$$P(W|V) = \frac{P(V|W) * P(W)}{P(V)}$$

So the percentage of computers infected with a virus running Windows is:

$$= (2p * 0.73)/(3p)$$

$$\approx \mathbf{0.48667}$$

5. Two cards are randomly chosen without replacement from an ordinary deck of 52 cards. Let E be the event that both cards are Aces. Let F be the event that the Ace of Spades is one of the chosen cards, and let G be the event that at least one Ace is chosen. Compute:

   (a) P(E—F) =

   (b) P(E—G) =

6. Five servers are located in a computer cluster. After one year, each server independently is still working with probability p, and otherwise fails (with probability 1 p).

   (a) What is the probability that at least 1 server is still working after one year?

   (b) What is the probability that exactly 2 servers are still working after one year?

   (c) What is the probability that at least 2 servers are still working after one year?

7. Consider a hash table with 5 buckets, where the probability of a string getting hashed to bucket i is given by p_i (where $\sum_{i=1}^{5} = 1$). Now, 6 strings are hashed into the hash table.

   (a) Determine the probability that each of the first 4 buckets has at least 1 string hashed to each of them. Explicitly expand your answer in terms of p_is, so that it does not include any summations.

   (b) Assuming $p_1 = 0.1$, $p_2 = 0.25$, $p_3 = 0.3$, $p_4 = 0.25$, $p_5 = 0.1$, explicitly compute your answer to part (a) as a numeric value.