CS 70, Spring 2015 — Solutions to Homework 14

Due Monday May 1 at 12 midnight

1. **Virtual Lab**

part 1.1 I would pick 1 because that is the number with the highest occurance probability from all the other numbers.

part 1.2 The very first image looks like it can be the number 5 or 3 depending on how we look at the top of the image.

part 2.1 It is almost impossible to recognize some of the original digits I clicked on. They definitely do not look like handwritten numbers because everything is pixelated. This may be because as we scan the numbers, we might try to pixelate or square each point in order to create a representation.

part 2.2 By creating a lot, we can have a lot of variation and we see where each pixel is primarily created or exists. And as a result, we can be able to see if the image is a specific value.

part 2.3 Determining by the most given pixel location, I would use it to randomly generate a digit by the one that is created the most.

part 2.4 As mentioned, we determine by the line location and convert it into pixel locations. From there, we see where these locations fall in our data. So within our data, we see which one it falls into or is part of the most. This will give us the number that has the highest probability of it being such.

part 2.5 Intuitively, is it because they all connect with each other so each position is not necessarily correct? When we write, it is a connected thing, therefore it should not be considered as an independent pixel. Not only that, it will cause it harder to figure out/find the actual value since it will be scattered all over the place and a lot harder to even find the exact spots.

part 3 It seems like the second block is to be predicted by Naive Bayes by a lot in every single block types. The reason for this might be because of how each number is pixelized. There are a lot of sharp edges, and as a result there should be a more probability that the block with the high percentage from Naive Bayes.

part 4 Similar to how it was calculated originally, we just take in consideration of the pixels around it instead of the one square given earlier. In our case, everything around it needs to match up in order for it to be correct or matched up. As a result the probability is more exact for both probability given.

part 5.1 The newest one looks more like handwriting; there is a more exact and percise line to it than the original we had earlier.

part 5.2 Because now it is more exact. We have more area to work with to see if it maps to or is similar to. Then as we check, we see which one is more similar since there is more area. We can even see this visually by how it prints out each values.

2. **Simpson**

   The paradox is to focus on the numbers of all Caucasians and African Americans for both population. The numbers for the Caucasians is extremely higher than the African American one in terms of ratio. And as a result, the numbers per 100,000 is a lot higher in African American. But when everything is added together and normalized, the total numbers of death in terms of ratio is higher. The numbers split among the two is not true primarily is due to the uneven numbers of Caucasians versus African Americans.

   Technically I am Asian so I don't have data... however whether or not I was Caucasian or African American, I would choose Richmond because it has a lower number of deaths for both, even if per population it is larger.

3. **Polya Urn**

   (a) The probability in this case is after the entire grabbing time. Since every grab is independent, we multiply each instance/grab together. Therefore the denominator should be a multiple of all the total everytime. In this case, we know that it is $\prod_{i=0}^{n-1} b + r + ki$. This is because every time we grab something out, there's an additional k-balls added in. That's why it is the product of all of this. This does not change, whether it is black or red ball. So in this case, the j-th position is independent and does not affect the probability (denominator) in anyway.

   The numerator is also independent because now we're going to think about the numbers that goes into picking one red, and the rest black. Essentially, this mean the chances is only getting red happens once which takes in all of the red possibility which is r. Then everything else, we multiply all of the numbers of existing black balls, which is $\prod_{i \neq j}^{n} b + ik$, at the end, we multiply r in. Therefore none of these are affected by j-th position. Essentially as a whole, it is the same since it is the product of each other.

   (b) $Pr[B_1, R_2, B_3, B_4, R_5] = \frac{b}{b+r} * \frac{r}{b+k+r} * \frac{b+k}{b+2k+r} * \frac{b+2k}{b+3k+r} * \frac{r+k}{b+4k+r} = \frac{r(r+k)*b(b+k)(b+2k)}{(b+r)(b+k+r)(b+2k+r)(b+3k+r)(b+4k+r)}$

   $Pr[R_1, R_2, B_3, B_4, B_5] = \frac{r}{b+r} * \frac{r+k}{b+k+r} * \frac{b}{b+2k+r} * \frac{b+k}{b+3k+r} * \frac{b+2k}{b+4k+r} = \frac{r(r+k)*b(b+k)(b+2k)}{(b+r)(b+k+r)(b+2k+r)(b+3k+r)(b+4k+r)}$

   Therefore these two are the same.

   (c) The proof for this branches off of part one and two. As stated above, the denominator is always independent and will take on the form of $\prod_{i=0}^{n-1} b + r + ki$. Now we should notice the pattern from number 2. We see that the numerator would be a product of all the reds over the time. Then the rest would be the total of blacks. So the number of reds is: $\prod_{i=0}^{l-1} r + ik$. Then the number of black will be: $\prod_{i=l}^{n-1} b + ik$. Now the product of those two will give us the numerator. Both of these work is because we know that it is independent, from part one and two. Therefore we know that it is the multiple of each independent probability. All of these represent the independent values of specific numerator and denominator. Which ultimately gives us the answer as: $\frac{(\prod_{i=l}^{n-1} b+ik)(\prod_{i=0}^{l-1} r+ik)}{\prod_{i=0}^{n-1} b+r+ki}$

   (d)

   (e)

4. **Surveys: the limit of trust**

(a) Induction is on the number of edges we have.

Base Case: $n = 0$. Since there are no edges, both of their sums for the edge's degree should be zero.

Inductive Hypothesis: Say that our case works for n-edges (the sum of all edges of the vertex for L and R are the same).

Inductive Step: Now for $n + 1$ edges, we can remove one of the edge so that it fits our hypothesis. In this case, we will put the edge back onto the graph and it should by geometry form the same degree on both the R and L sides. Then in this case, we add the same amount of degree to both the summation of degree for both the R and L side. And as a result, the equality remains the same.

(b) In this case, L will be all of the males who are sexually active, and R will be all of the females who are sexually active. Then our vertices will connect those who have sexual relations with each other – in this case, it is our edges. So essentially the degree of each vertex represents how many people each vertex (person) has had sexual relations with. However the studies seem to be inaccurate because even though there are an even number of females nad males, the average is way off. From our first proof, we stated that they have to be equal – or adds up to be the same. Which is not the case if our average is not equivalent to each other. This means that within our study, there is a room of error or incorrect responses.