CS 70 Discrete Mathematics and Probability Theory
Spring 2015 Vazirani HW 14

Reminder: Deadline for HW13 self-grade is Thursday, April 30 at noon.

Reading: Note 21, 22, and 23.

Due Midnight Friday, May 1

1. **Virtual Lab**

   Follow the virtual lab for this homework (provided on the course website). Include screenshot(s) of the lab and provide answers to the questions posed there in your homework submission.

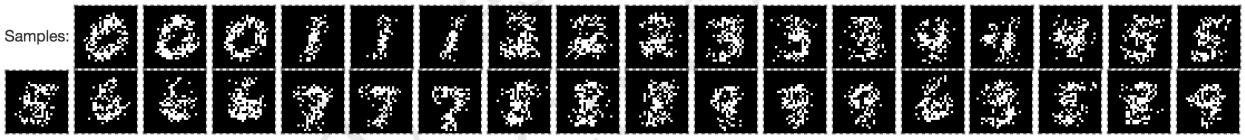
   Answer: MNIST Database

   Questions

   1. We should pick 1, because it is the most frequent digit based on the given prior distribution.
   2. No, the samples are clear for a human being.

   Naive Bayes

   Questions

   1. In general it is difficult to recognize the samples, although it is easier for simpler digits like 0 or 1. The generated samples do not look like handwritten digits because these samples treat each pixel independently and thus do not retain the coherence of handwriting.
   2. Here for each digit we are given the the conditional probability for each pixel in the image of that digit. To generate a sample from a digit, for each pixel we independently flip a biased coin with bias equal to the conditional probability of that pixel. Specifically, for each pixel, we use the tool to generate a random number x in $[0,1]$ and compare it to the conditional probability p of that pixel. If $x < p$, then we output 1 for that pixel (the pixel is on/white); otherwise, we output 0 (the pixel is off/black). This procedure works because $\Pr[\text{pixel is on}] = \Pr[x < p] = p$, as desired.
   3. Suppose $z_0, z_1, \ldots, z_9$ are the probabilities for each digit in the posterior distribution. Since they sum up to 1, we can partition $[0,1]$ into ten intervals $A_0 = [0, z_0], A_1 = [z_0, z_0 + z_1], A_2 = [z_0 + z_1, z_0 + z_1 + z_2], \ldots, A_9 = [z_0 + \cdots + z_8, 1]$.

      Now we generate a random number in $[0,1]$ and see which interval A_i it falls into. The probability that it falls into A_i is exactly z_i, so we have effectively sampled a digit i from the posterior distribution. Finally we generate a sample as in the previous question, using the grayscale image for digit i.

4. We simply choose the digit x that maximizes the posterior probability. Since the posterior probability of x is proportional to the product of the prior probability of digit x and the conditional probability of that image being generated given that it is an image of x, it is equivalent to choosing x that maximizes this product.
5. In handwriting, if one pixel is white, it is likely that its neighboring pixels are also white. This is because handwriting is contiguous and also because the pen has some stroke width. An illustrating example of this is that in handwriting we almost never see an isolated white pixel. In contrast, we can see that samples generated by Naive Bayes tend to have a large number of isolated white pixels.

Weaknesses

Questions

1. The block type in which all four pixels are white is underestimated by Naive Bayes. While this block type appears very frequently in the actual data, the probability that all four pixels become white in Naive Bayes is relatively small because it treats each pixel independently.

   In contrast, the block types in which two pixels on a diagonal are white are terribly overestimated by Naive Bayes. These block types almost never occur in the actual data, but Naive Bayes predicts them with considerable probability because of the independence assumption between pixels. The block types in which only one of the four pixels is white also tend to be overestimated by Naive Bayes for the same reason (although not as dramatically).

2×2 blocks

Questions

1. The 28×28 image can be viewed as a 14×14 grid of 2×2 blocks ($14 \times 14 = 196$ blocks total). We will denote these blocks by B_{ij} where $1 \leq i, j \leq 14$. Now, for each digit x and each B_{ij}, we do the following:
   - Going through the images for digit x, we count the frequency of each of the 16 block types for B_{ij}.
   - Estimate the probability of each block type k to be proportional to the frequency:

$$\Pr[B_{ij} = k \mid \text{digit } x] = \frac{\text{freq. of } k \text{ in digit } x}{\text{number of images for digit } x} = p_{ij,x}(k). \tag{1}$$

   Now suppose we are given an image and we want to compute the conditional probability of this image being generated given that it is an image of digit x. Suppose the given image has block type b_{ij} (one of 16 possible values) for block B_{ij}, for each $1 \leq i, j \leq 14$. Assuming that each 2×2 block behaves independently, we can compute the desired conditional probability as:

$$\Pr[\forall i, j : B_{ij} = b_{ij} \mid \text{digit } x] = \prod_{i,j} \Pr[B_{ij} = b_{ij} \mid \text{digit } x] = \prod_{i,j} p_{ij,x}(b_{ij}) \tag{2}$$

   where we plug in our estimate (1) of $p_{ij,x}(k) = \Pr[B_{ij} = k \mid \text{digit } x]$ in the product above.

   After computing the conditional probability, we can compute the posterior distribution of the digits given the observed image via Bayes' rule:

$$\Pr[\text{digit } x \mid \forall i, j : B_{ij} = b_{ij}] = \frac{\Pr[\text{digit } x] \cdot \Pr[\forall i, j : B_{ij} = b_{ij} \mid \text{digit } x]}{\Pr[\forall i, j : B_{ij} = b_{ij}]}.$$

In the numerator we plug in the prior estimate of digit x from the dataset (this is the same as in the usual Naive Bayes) and the conditional probability from (2). We can compute the denominator via the total probability rule:

$$\Pr[\forall i, j : B_{ij} = b_{ij}] = \sum_{y=0}^{9} \Pr[\text{digit } y] \cdot \Pr[\forall i, j : B_{ij} = b_{ij} \mid \text{digit } y].$$

Questions

1. Although still not very clear, samples generated by 2×2 blocks are less fuzzy than samples generated by Naive Bayes and have more contiguous blocks, and thus resemble handwriting more.
2. In Naive Bayes we were making an inaccurate assumption that each pixel behaves independently. By contrast, in the 2×2 block model we are making a slightly less inaccurate assumption that each 2×2 block behaves independently, so we are taking into account some of the correlations between adjacent pixels, and thus we would expect the error rate to go down a little bit.

2. **Simpson**

Data from the actual death rates in 1910 from tuberculosis is reproduced in the table below:

	Population		Deaths		Death Rate per 100,000	
	NY	**Richmond**	**NY**	**Richmond**	**NY**	**Richmond**
Caucasians	4,675,174	80,895	8,365	131	179	162
African Americans	91,709	46,733	513	155	559	332
Totals	4,766,883	127,268	8,878	286	186	225

Note that it shows that the death rate for both African Americans and Caucasians was lower in Richmond than in New York. Yet, the death rate for the total combined population of African Americans and Caucasians was higher in Richmond than in New York. Can you explain this paradox? If you had been given the choice, back in 1910, of living in either city, which would you have chosen based only on risk of tuberculosis death?

Answer: Observe that African Americans have a significantly higher death rate than Caucasians. Richmond has a much higher share of African Americans than New York, so when we combine the totals, Richmond is more affected by the high death rate among African Americans, and hence has a higher death rate. We should choose to live in Richmond since it has a lower death rate for every population.

3. **Polya Urn**

An urn starts with r red balls and b black balls. In each step, we pick a random ball from the urn, put it back, and add k additional balls of the same color. We wish to answer the question: at the n-th such step, what is probability that the randomly drawn ball is red? The following parts will guide you to an answer. Let R_j (B_j) be the event that the ball drawn in the j-th step is red (respectively, black).

(a) Suppose we want to compute the probability that the ball drawn in the j-th step is red while the balls drawn in all other $n-1$ steps are black, i.e., $\Pr[B_1,\ldots,B_{j-1},R_j,B_{j+1},\ldots,B_n]$. Show that this probability is independent of j, for any $1 \le j \le n$.

Answer: We can compute this probability as

$$\begin{aligned}\Pr[B_1,\ldots,B_{j-1},R_j,B_{j+1},\ldots,B_n] &= \frac{b}{b+r} \times \frac{b+k}{b+r+k} \times \frac{b+2k}{b+r+2k} \times \cdots \times \frac{b+(j-2)k}{b+r+(j-2)k} \\ &\quad \times \frac{r}{b+r+(j-1)k} \times \frac{b+(j-1)k}{b+r+jk} \times \cdots \times \frac{b+(n-2)k}{b+r+(n-1)k} \\ &= \frac{r \times \prod_{i=0}^{n-2}(b+ik)}{\prod_{i=0}^{n-1}(b+r+ik)},\end{aligned}$$

which is independent of j.

The point is that the denominator is clearly independent of j, and we can multiply the pieces in the numerator in any order, in which case we see that the product of the numerators is also independent of j.

(b) Actually, there is a more basic symmetry. We claim that the probability of drawing any sequence of red and black balls only depends on the number of red balls that we get, not on their order. For example, the probability of the sequence B_1,R_2,B_3,B_4,R_5 is the same as the probability of R_1,R_2,B_3,B_4,B_5 because they both have 2 red balls and 3 black balls. Verify this by computing $\Pr[B_1,R_2,B_3,B_4,R_5]$ and $\Pr[R_1,R_2,B_3,B_4,B_5]$ and checking that they are the same.

Answer: We can compute

$$\begin{aligned}\Pr[B_1,R_2,B_3,B_4,R_5] &= \frac{b}{b+r} \times \frac{r}{b+r+k} \times \frac{b+k}{b+r+2k} \times \frac{b+2k}{b+r+3k} \times \frac{r+k}{b+r+4k}, \\ \Pr[R_1,R_2,B_3,B_4,B_5] &= \frac{r}{b+r} \times \frac{r+k}{b+r+k} \times \frac{b}{b+r+2k} \times \frac{b+k}{b+r+3k} \times \frac{b+2k}{b+r+4k}.\end{aligned}$$

We see that these two quantities are equal because they have the same denominators, and the same numerators (in a different order), and we can multiply in any order.

(c) Generalize your argument in part (b) to prove that the probability of any sequence of n red and black balls (e.g., $R_1,B_2,R_3,R_4,B_5,\ldots,B_n$) only depends on the number of R_j terms and not on their order.

Hint: If there are ℓ red balls, show that the probability is equal to $\Pr[R_1,\ldots,R_\ell,B_{\ell+1},\ldots,B_n]$.

Answer: By the same argument in part (b), we see that changing the order of the sequence of red and black balls does not change the denominator of the probability. The numerator contains the same terms, but they appear in a different order. Since we can multiply in any order, this does not change the probability.

We have that

$$\Pr[R_1,\ldots,R_\ell,B_{\ell+1},\ldots,B_n] = \frac{\prod_{i=0}^{\ell-1}(r+ik) \times \prod_{i=0}^{n-\ell-1}(b+ik)}{\prod_{i=0}^{n-1}(b+r+ik)}.$$

If the colors came in a different order, the terms in the numerator would correspondingly be ordered differently, but that makes no difference to the resulting product. This is what we wanted to show.

(d) Now compute the probability of getting a red ball at the n-th step, i.e., $\Pr[R_n]$.
Hint: Use part (c) to show that $\Pr[R_n] = \Pr[R_1]$.

Answer: We apply the total probability rule to compute $\Pr[R_n]$. This is a sum of the probabilities for all possible sequences of $(n-1)$ red and black balls, followed by a red n-th ball. In each term, we can swap the first and last ball, to get a sequence of red first ball followed by $(n-1)$ red and black balls, which by the total probability rule is $\Pr[R_1]$.
Now, we can easily compute

$$\Pr[R_n] = \Pr[R_1] = \frac{r}{b+r}.$$

(e) What is the expected number of red balls in the urn after the n-th step?

Answer: Let random variable X denote the number of red balls in the first n steps. Then $X = X_1 + \cdots + X_n$ where each X_i is the indicator for the event R_i that we draw a red ball at the i-th step.
From part (d), we see that $\mathrm{E}[X_i] = \Pr[R_i] = \frac{r}{b+r}$. By linearity of expectation, we therefore see that $\mathrm{E}[X] = \sum_{i=1}^{n} \mathrm{E}[X_i] = \frac{nr}{b+r}$.
At each step, we add red balls only when we draw a red ball. Each time that we draw a red ball, we replace it and add k more red balls. Therefore, the expected number of red balls after the n-th step is $kE[X] + r = \frac{knr}{b+r} + r$.

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

A graph is called bipartite if its vertices are divided into two disjoint sets L and R such that every edge $\{u, v\}$ has one endpoint in L and one endpoint in R.

(a) Prove by induction that $\sum_{v \in L} d_v = \sum_{v \in R} d_v$, where d_v is the degree of vertex v.
Clearly state what variable the induction is on.

Answer: We use induction on the number of edges m in a bipartite graph.
Base case: If $m = 0$, obviously $\sum_{v \in L} d_v = \sum_{v \in R} d_v = 0$.
Inductive hypothesis: The equality $\sum_{v \in L} d_v = \sum_{v \in R} d_v$ holds for bipartite graphs with m edges.
Inductive step: Consider a bipartite graph with $m+1$ edges. Delete one edge and a bipartite graph still remains bipartite (since all the remaining edges still go from one set to another). By our inductive hypothesis, in this smaller bipartite graph $\sum_{v \in L} d_v = \sum_{v \in R} d_v$. Adding the deleted edge back to the graph adds one to both $\sum_{v \in L} d_v$ and $\sum_{v \in R} d_v$. Therefore, $\sum_{v \in L} d_v = \sum_{v \in R} d_v$ still holds.

(b) A number of surveys, including one by the Federal government published a decade ago, show that men have had on average more sexual partners than women — for example, one in particular claims that in the U.S. men had on average 6.5 sexual partners whereas women had on average only 4.2. In 2007, the New York Times published an article debunking these studies, quoting prominent mathematicians, including Prof. Gale from our Mathematics department at Berkeley, saying that these claims are mathematically impossible.
In this question, you will model the situation using bipartite graphs and apply the claim from part (a) to pinpoint a problem with these studies. The surveys in question refer only to heterosexual

couples and for simplicity assume that the number of men in the US is equal to the number of women. Model the claims of the study as claims about average degrees in a certain bipartite graph. What are the vertices and edges in the graph? And what are the precise claims about the average degrees? What can you conclude about the studies?

Answer: Suppose we model this situation using the following bipartite graph; the vertices in L will represent men and and the vertices in R will represent women. There will be an edge for each pair that were partners. Then, the degree of a given vertex is the number of partners that person had. Assuming that the number of men is equal to the number of women ($|L| = |R| = n$), $\frac{1}{n}\sum_{v \in L} d_v$ is the average number of partners for men and $\frac{1}{n}\sum_{v \in R} d_v$ is the average number of partners for women. But as we argued in part (a), these two quantities must be equal. This means that some of the people that participated in the surveys did not respond accurately.