CSE 202: Design and Analysis of Algorithms Winter 2013

Problem Set 4

Instructor: Kamalika Chaudhuri **Due on:** Thu Mar 7, 2012

Instructions

- For your proofs, you may use any lower bound, algorithm or data structure from the text or in class, and their correctness and analysis, but please cite the result that you use.
- All problems are worth 10 points.
- If you do not prove that your algorithm is correct, we will assume that it is incorrect. If you do not provide an analysis of the running time, we will assume you do not know what the running time is.

Problem 1

Provide brief answers to the following questions. Detailed proofs are not necessary.

1. You are given an array A of n data elements, which come from k clusters. You don't know which element of A belongs to which cluster, but you do know that there are exactly n/k elements in A that belong to each cluster. Suppose you pick a random subset S of m elements from A with replacement. How large does m need to be so that with probability $1/2$, S includes at least one element from every cluster in A?
2. Consider the following variation of the coupon collector's problem. Each cereal box contains one of $2n$ different coupons. The coupons are in pairs – coupons 1 and 2 are a pair, 3 and 4 are a pair, and so on. You get a prize if you get at least one coupon from each pair. Assuming that the coupons in each box are chosen independently and uniformly at random from the $2n$ coupons, what is the expected number of cereal boxes you must buy before you get a prize?

Solution

1. Let A_i be the event that no element from cluster i was chosen. The probability that all clusters are chosen can be given by $1 - \Pr[A_1 \cup \ldots \cup A_k]$. Using the union bound, this probability can be written as:

$$\begin{aligned} 1 - \Pr[A_1 \cup \ldots \cup A_k] &\geq 1 - \sum_{i=1}^{k} \Pr[A_i] \\ &= 1 - k\left(1 - \frac{1}{k}\right)^m. \end{aligned}$$

   This probability is $\geq 1/2$ when $m = \Theta(k \log k)$.

2. This problem is exactly equivalent to the coupon collector's problem we did in class; assume that the relevant pairs of coupons form a label. For example, coupons 1 and 2 both have label 1, coupons 3 and 4 have label 2 and so on. For any i, the probability of drawing a coupon out of the pair that forms label i is $\frac{1}{n}$, and we get a prize when we have collected all the labels. From the result discussed in class, the expected number of boxes collected is $\Theta(n \log n)$.

Problem 2

Here is a very general balls-in-bins problem. Suppose m balls are thrown into n bins, but that the bins are not equally likely to be chosen. Each time a ball is thrown, it goes into bin 1 with probability p_1, bin 2 with probability p_2, and so on. The numbers $p_1, p_2, ..., p_n$ are nonnegative and sum to 1.

1. Let X_i be the number of balls that fall into bin i. What is the probability that X_i is exactly k?
2. What is $\mathbf{E}(X_i)$ and $\mathbf{Var}(X_i)$?
3. Give an upper bound on the probability that there is an empty bin. The larger m is, the smaller this probability. Roughly how large should m be for this probability to be less than $\frac{1}{2}$? (So that with probability at least $\frac{1}{2}$, every bin will contain one or more balls: a generalization of the coupon collector problem.)
4. Give an upper bound on the probability that there is a bin with at least 2 elements. The larger m is, the larger this probability. Roughly how small must m be for this probability to be less than $\frac{1}{2}$? (So that with probability at least $\frac{1}{2}$, every bin has at most one ball: a generalization of the birthday paradox.)
5. Check that your answers to (3) and (4) agree with what we already know for the uniform case, that is, when $p_1 = p_2 = \ldots = p_n$.

Solution

1. The probability that exactly k balls fall into the i-th bin or $\Pr(X_i = k)$ is:

$$\binom{m}{k} p_i^k (1-p_i)^{m-k}$$

   The first term is the number of ways that the k balls can be thrown (out of m tosses), the second term is the probability that exactly k balls land in the same bin, the third term is the probability that all other balls land in other bins

2. Let $B_j = 1$ if the ball from the j-th toss lands in bin i and 0 otherwise. Then, $X_i = B_1 + \ldots + B_m$. Now, $\mathbf{E}[B_i] = p_i$, and $\mathbf{Var}(B_i) = p_i(1-p_i)$. Therefore, using linearity of expectation,

$$\mathbf{E}[X_i] = \sum_j \mathbf{E}[B_j] = mp_i$$

   Moreover, as $B_1, \ldots, B_m$ are independent (recall that the balls are tossed independently),

$$\mathbf{Var}(X_i) = \sum_j \mathbf{Var}(B_j) = mp_i(1-p_i)$$

3. Let A_i be the event that bin i is empty. Then, $\Pr(A_i) = (1-p_i)^m$. Now, using the Union Bound, $\Pr(A_1 \cup A_2 \cup \ldots A_n) \leq \sum_i \Pr(A_i) = \sum_i (1-p_i)^m$.
4. Here, let A_i be the event that bin i has at least two balls. Then, $\Pr(A_i) \leq \binom{m}{2} p_i^2$, and using the Union Bound, $\Pr(A_1 \cup A_2 \cup \ldots A_n) \leq \sum_i \Pr(A_i) \leq \binom{m}{2} \sum_i p_i^2$.
5. In the uniform case, $p_i = 1/n$ for all i. Using (3), the probability that there is an empty bin after m tosses is $n \cdot (1-\frac{1}{n})^m < \frac{1}{2}$ when $m = \Theta(n \log n)$. (Note that as $n \to \infty$, $(1-\frac{1}{n})^n \to 1/e$. This coincides with the solution to the coupon collector's problem.

   Using (4), the probability that some bin has at least two balls is $\binom{m}{2} n \cdot \frac{1}{n^2} = \frac{m(m-1)}{2n}$, which coincides with the solution to the birthday paradox in class.

Problem 3

Blood samples from n people are being tested for a disease. It can be costly to test each person separately, so we pool blood from a group of k people and test them. If the test is negative, we know noone in the pool have the disease. If the test is positive, all the k people in the pool have to be tested again individually to find out who has the disease, and thus $k+1$ total tests are required for the k people in the pool. Suppose n is a multiple of k, and we create n/k disjoint pools of k people to test. Moreover, assume that each one of n people has the disease independently with probability p.

1. What is the probability that the test will be positive for a pooled sample of k people?
2. What is the expected number of tests needed?
3. Given n and p, what is the best value of k so that the expected number of tests is minimized?
4. Notice that pooling is not always better than doing n individual tests. Give an inequality to show for what values of p, pooling is better than doing individual tests.

Solution

1. The test will be positive for the pooled sample if any of the k people have the disease. We can write

$$\begin{aligned} \Pr[\text{Test positive}] &= 1 - \Pr[\text{Test negative}] \\ &= 1 - (1-p)^k. \end{aligned}$$

2. Let T be the random variable representing the number of tests needed, and let T_i be the number of tests needed for the i-th pool of k people. We can write $T = \sum_{i=1}^{n/k} T_i$.

   To compute $\mathbf{E}[T_i]$, note that if the test is negative, we only need one test, and if the test is positive, we need $k+1$ tests. We already calculated the probability that the test is positive in part 1; therefore:

$$\begin{aligned} \mathbf{E}[T_i] &= 1 \times (1-p)^k + (k+1) \times (1 - (1-p)^k) \\ &= 1 + k(1 - (1-p)^k). \end{aligned}$$

   Linearity of expectation implies that $\mathbf{E}[T] = \sum_{i=1}^{n/k} \mathbf{E}[T_i]$, or

$$\mathbf{E}[T] = \frac{n}{k} + n(1 - (1-p)^k).$$

3. To find local extrema of $\mathbf{E}[T]$, we can take the first derivative with respect to k:

$$\frac{\partial}{\partial k}\mathbf{E}[T] = -\frac{n}{k^2} - n(1-p)^k \log(1-p).$$

   The roots of $\frac{\partial}{\partial k}\mathbf{E}[T]$ give the local extrema of $\mathbf{E}[T]$. For a root k, to check that it is a *minimum* (and not a maximum), we can take the second derivative:

$$\frac{\partial^2}{\partial k^2}\mathbf{E}[T] = \frac{2n}{k^3} - n(1-p)^k \log^2(1-p).$$

   If k is a root of the first derivative, and the second derivative evaluated at k is positive, then k is a local minimum.

   We must also check the boundary values for k: $k = 1$ and $k = n$. The overall best k is the k which minimizes $\mathbf{E}[T]$ among the local minima, $k = 1$, and $k = n$. Computing $\mathbf{E}[T]$ for such k and comparing will give the best k.

4. Without pooling, n tests are required. Pooling is better than doing individual tests if:

$$\frac{n}{k} + n(1 - (1-p)^k) < n$$

   for the k found in part 3.