

Data mining: lecture 1

Edo liberty

Suppose you are a marine biologist (Although you prefer to pretend to be an architect), and suppose you are tasked with counting the number of individuals in a huge school of tune fish in the middle of the atlantic ocean. How would you go about doing that? One possible approach is called Mark and recapture. Start by catching k fish. Then, mark them somehow and release them. Then catch another group of k fish and count the number of fish that are already marked, Z . You can now guess that the number of fish in the entire school is roughly k^2/Z . Jacques Cousteau would have been proud.

Mark and recapture

Given a set of n elements, sample k without replacement twice. Count the number of identical elements in both groups, Z . Define a random variable $z_{i,j}$ which indicates that element i in the first group is the same as element j in the second. The value of Z is therefore $Z = \sum_{i,j} z_{i,j}$. Lets compute the expectation of Z using linearity of expectation. Note that the $z_{i,j}$ variables are not independent!

$$E[Z] = E\left[\sum_{i,j} z_{i,j}\right] = \sum_{i,j} E[z_{i,j}] = \sum_{i,j} 1/n = k^2/n \quad (1)$$

Lets compute the standard deviation of Z . Recall:

$$\sigma^2[Z] = E[Z - E[Z]]^2 = E[Z^2] - E[Z]^2$$

We need the use the linearity of expectation again to compute $E[Z^2]$:

$$E[Z^2] = E\left[\left(\sum_{i,j} z_{i,j}\right)\left(\sum_{i',j'} z_{i',j'}\right)\right] \quad (2)$$

$$= \sum_{i=i',j=j'} E[z_{i,j}z_{i',j'}] \quad (3)$$

$$+ \sum_{i=i',j \neq j'} E[z_{i,j}z_{i',j'}] + \sum_{i \neq i',j=j'} E[z_{i,j}z_{i',j'}] \quad (4)$$

$$+ \sum_{i \neq i',j \neq j'} E[z_{i,j}z_{i',j'}] \quad (5)$$

$$= \frac{k^2}{n} + 0 + 0 + \frac{k^2(k-1)^2}{n^2} \quad (6)$$

$$\sigma^2[Z] = \frac{k^2}{n} + \frac{k^2(k-1)^2}{n^2} - \left(\frac{k^2}{n}\right)^2 \quad (7)$$

$$\leq \frac{m^2}{n} \quad (8)$$

Now we invoke Chebyshev's inequality.

$$\Pr\left[\left|Z - \frac{k^2}{n}\right| > t\right] \leq \frac{k^2}{nt^2} \quad (9)$$

Choosing $t = 10k/\sqrt{n}$ we get that with probability at least 0.99

$$\left|Z - \frac{k^2}{n}\right| \leq 10k/\sqrt{n} \quad (10)$$

Which gives:

$$n \leq \frac{k^2}{Z} \left(1 + \frac{10\sqrt{n}}{k}\right) \quad (11)$$

$$n \geq \frac{k^2}{Z} \left(1 - \frac{10\sqrt{n}}{k}\right) \quad (12)$$

This gives us the following procedure: First, sample 2 groups of size $k \geq 50\sqrt{n}$ each. Count the number of collision Z . Estimate the size of the set as $n_{alg} = k^2/Z$. We are guaranteed that with probability 0.99 our estimate is within 20% accuracy.

$$\frac{5}{6}n \leq n_{alg} \leq \frac{5}{4}n \quad (13)$$