The Secretary Problem

Sebastian Liu

Problem

The Problem: Suppose $N = \{1, \ldots, n\}$ is a set of rankable individuals who are seeking to be hired by you for a secretary position. (By rankable, we mean that each candidate has a value given to him or her based on his or her skill, and that these values form a totally ordered set.) The candidates are interviewed in a random order. You know how each candidate has done with respect to the other candidates you have already interviewed, but you do not know how he or she fares against the candidates you have not interviewed yet. After each interview, you must immediately make a decision to reject or hire the candidate. Once a candidate is rejected, you cannot go back to hire him or her. Once a candidate is hired, the interview process stops. How do you maximize your chances of hiring the best candidate?

Two Naive Approaches

Perhaps an initial thought might be to have a baseline set of criteria and reject every candidate not meeting this set of criteria. We could then hire the first candidate who meets all the criteria, or hire the last candidate should all the previous candidates fail. With such a strategy, however, it is difficult to see how we are optimizing for the best candidate.

Another line of thought is that we should choose a random number between 1 and n, and simply hire the corresponding candidate. After all, the candidates are interviewed in a random order, so the probability that any one of them is the best is equal to exactly $\frac{1}{n}$. This strategy gives us a $\frac{1}{n}$ chance of picking the best candidate.

An Optimal Solution

As it turns out, (and I will spend most of the remainder of this paper proving this fact) the optimal strategy is to automatically reject the first $\frac{n}{e}$ candidates, keeping track of the best candidate in that set, and then hiring the first candidate better than that one in the remaining set of candidates (or hiring the last candidate if the remaining candidates are all worse). The intuition here is that we are 'sampling' the set of $\frac{n}{e}$ candidates. This information is then used to help us choose a candidate to hire in the remaining set of candidates. But why $\frac{n}{e}$?

Suppose there is some optimal stopping point, k for which you reject all candidates before k and are selecting between $k+1$ and n. Our set of candidates can be written as $N = \{1, \ldots, k-1, k, k+1, \ldots, n\}$. We need to show that $k = \frac{n}{e}$.

Let us define success as choosing the best candidate in N. Then the probability of success is given by the following:

$$\sum_{i=0}^{n} P(i \text{ is the best} \cap \text{ choose}_i)$$

(where i is the candidate in position i). The above is simply:

$$\sum_{i=0}^{n} P(i \text{ is the best }) \times P(\text{choose}_i \mid i \text{ is the best})$$

The probability that the candidate in position i is the best is straightforward: $P(i \text{ is the best }) = \frac{1}{n}$. A justification of this is left as an exercise to the reader. We need only to find $P(\text{choose}_i \mid i \text{ is the best})$ for each i.

Since we are partitioning our set into two disjoint sets (those before and including k and those after k), the probability of success is simply:

$$\sum_{i=1}^{k} \frac{1}{n} \cdot P(\text{choose}_i \mid i \text{ is the best}) \; + \sum_{i=k+1}^{n} \frac{1}{n} \cdot P(\text{choose}_i \mid i \text{ is the best})$$

We first consider $P(\text{choose}_i \mid i \text{ is the best})$ for $1 \leq i \leq k$. Because our strategy mandates that we skip the first k candidates, $P(\text{choose}_i \mid i \text{ is the best}) = 0$ for $1 \leq i \leq k$. Therefore we have our probability of success as:

$$0 \; + \sum_{i=k+1}^{n} \frac{1}{n} \cdot P(\text{choose}_i \mid i \text{ is the best}) = \frac{1}{n}\left(\sum_{i=k+1}^{n} P(\text{choose}_i \mid i \text{ is the best})\right)$$

So what is $P(\text{choose}_i \mid i \text{ is the best})$ for $k+1 \leq i \leq n$? Well consider the case that the best candidate is in position $k+1$. Since we have skipped the first k and we are assuming that $k+1$ is the best candidate, our probability of choosing this candidate is 1.

Now what if the best candidate is in position $k+2$? On the one hand, if the best candidate is in position $k+2$ and the candidate in position $k+1$ is not better than all the previous candidates, then we would be guaranteed to pick candidate $k+2$. On the other hand, if the best candidate is in position $k+2$ but the candidate in position $k+1$ is better than all the previous candidates, then we would pick candidate $k+1$. What we really care about then, is the probability that candidate $k+1$ is better than all the previous candidates (that candidate $k+1$ is the best candidate of the set $\{1, \ldots, k+1\}$). That probability is $\frac{1}{k+1}$. A justification of this is left to the reader. Since we are picking candidate $k+2$ if candidate $k+1$ is *not* the best candidate in the set $\{1, \ldots, k+1\}$, the probability we pick candidate $k+2$ given that he or she is the best candidate is the complement of the probability that candidate $k+1$ is the best, which is $1 - \frac{1}{k+1} = \frac{k}{k+1}$. Therefore, the probability that we pick candidate $k+2$ given that he or she is the best candidate is $\frac{k}{k+2}$. We leave the reader to compute the probability that we pick candidate $k+3$ given that he or she is the best candidate, $k+4$ given that he or she is the best candidate, and so on.

We now have the following:

$$\frac{1}{n}\left(\sum_{i=k+1}^{n} P(\text{choose}_i \mid i \text{ is the best})\right) = \frac{1}{n}\left(1 + \frac{k}{k+1} + \frac{k}{k+2} + \cdots + \frac{k}{n-1}\right)$$

We can factor out a k term to obtain:

$$\frac{k}{n}\left(\frac{1}{k} + \frac{1}{k+1} + \frac{1}{k+2} + \cdots + \frac{1}{n-1}\right)$$

Moreover, we can approximate this:

$$\frac{k}{n}\left(\frac{1}{k}+\frac{1}{k+1}+\frac{1}{k+2}+\cdots+\frac{1}{n-1}\right)\approx\frac{k}{n}\left(\int_k^n\frac{1}{x}\,dx\right)$$

The remaining it a straightforward computation:

$$\frac{k}{n}\cdot\int_k^n\frac{1}{x}\,dx=\frac{k}{n}\cdot\ln(x)\Big|_k^n=\frac{k}{n}\left(\ln(n)-\ln(k)\right)=\frac{k}{n}\cdot\ln\left(\frac{n}{k}\right)=-\frac{k}{n}\cdot\ln\left(\frac{k}{n}\right)$$

Since we want to find the optimal number k, we take the derivative (we substitute $y=\frac{k}{n}$ for ease of readability):

$$\frac{d}{dy}\left(-y\cdot\ln(y)\right)=-y\cdot\frac{1}{y}+\ln(y)\cdot(-1)=-\ln(y)-1$$

We set $-\ln(y)-1$ to find the optimal y value and substitute $\frac{k}{n}$ back in for y:

$$\begin{aligned}-\ln(y)-1&=0\\ \ln(y)&=-1\\ e^{\ln(y)}&=e^{-1}\\ y&=e^{-1}\\ \frac{k}{n}&=e^{-1}\\ k&=\frac{n}{e}\end{aligned}$$

Thus we reach our conclusion that the optimal number of candidates to reject is $\frac{n}{e}$ as desired. Furthermore, the probability of choosing the best candidate given this strategy is $\frac{1}{e}$. A justification of this claim is left as an exercise to the reader.