Preliminary Written Exam Question for Jeremy Ash

Jacqueline M. Hughes-Oliver

May 29, 2018

Accumulation curves are used to assess the effectiveness of ranking algorithms. Items are ranked according to the algorithm's belief that they possess some desired feature (call it being active), then items are tested according to this *relative rank* or *normalized rank* or *testing order*. Ideally, all early tests will reveal the desired feature (i.e., all items tested early will be active). Accumulation curves plot the cumulative *number of actives* retrieved as a function of the relative rank of the item tested. Alternatively, accumulation curves may be scaled to plot the cumulative *fraction of actives* retrieved as a function of the relative rank of the item tested. **In this exam, you will develop a probabilistic framework for accumulation curves.**

Let S denote the score from a ranking algorithm, where larger values of S suggest stronger belief that the item is active. S is reasonably regarded as a random variable. Activity of an item may also be regarded as a random variable: $X = I(\text{active})$, where $I(\cdot)$ is the indicator function, and $\Pr(X = 1) = \pi$. Given that an item is active, S has cumulative distribution function $F_+(\cdot)$, and given that an item is not active, S has cumulative distribution function $F_-(\cdot)$.

Part I.

Study the probabilistic setting by deriving (formulas, plus graphs where relevant)

- the marginal probability density function of S, and
- the conditional distribution of X given $S = t$

under each of the following scenarios:

1. **binormal**: $F_+(\cdot)$ corresponds to the normal distribution having mean μ_+ and variance σ_+^2, and $F_-(\cdot)$ corresponds to the normal distribution having mean μ_- and variance σ_-^2. Select meaningful values for the parameters to create graphs.
2. **bibeta**: $F_+(\cdot)$ corresponds to the beta distribution with parameters α_+ and β_+, and $F_-(\cdot)$ corresponds to the beta distribution with parameters α_- and β_-. Select meaningful values for the parameters to create graphs.

Part II.

Consider an infinite-sized population of items such that it is only meaningful to view the accumulation curve as plotting the *fraction of actives* retrieved as a function of testing order. Argue for or against the claim below, making sure to investigate the binormal and bibeta scenarios as described above.

Claim: The population accumulation curve is obtained by plotting points $\{\Pr(S > t), \Pr(S > t|X = 1)\}$ as the threshold t (applied to scores from the algorithm) changes over the range of S.

Part III.

Conduct a simulation study to support your investigations in Part II. In other words, sample N items from your population to obtain $\{(S_i, X_i), i = 1, \ldots, N\}$. Use this data to estimate the population accumulation curve. Is the estimate reasonable? Is the quality of the estimate affected by N or π?