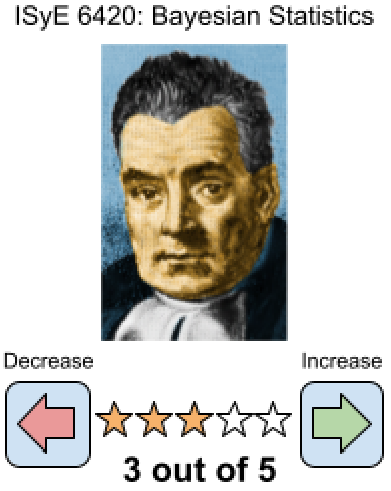
A Bayesian Relative Rating System

ISyE6420 Final Project

Author: Brandon Stewart
ID: rstewart61@gatech.edu

Introduction

Most rating systems allow each user to rate a product or service from one to five stars. Each user must specify exactly how many stars they think the product or service deserves. This paper proposes a different rating system where each user is given only two choices: increase or decrease the current rating. For the current course this may look like:



[1]

Each increase or decrease of the current rating becomes a censored observation. For example, if the user thinks that a rating of three stars is too high, then the censored observation would be [1, 3). When this new observation is combined with previous ones, the overall rating will slightly decrease. These censored observations can be accumulated and combined into an overall rating similar to the average ratings seen ubiquitously today on Amazon, IMDB, OMSCentral, and so on.

We'll see that relative ratings can be modeled as censored observations for a binomial distribution in a way that closely matches the accumulated average of absolute ratings.

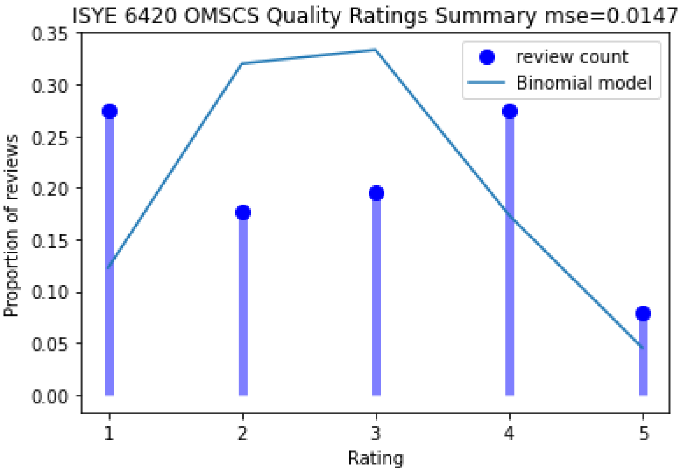
Modeling with the Binomial Distribution

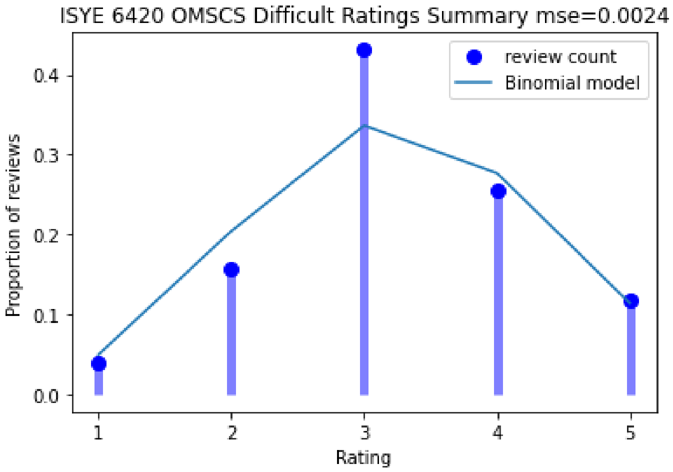
Since ratings can only take five values from 1 to 5, the binomial distribution was used. The binomial distribution is modeled as *B(n, p)* = $\binom{n}{x} p^{x}(1 - p)^{n-x}$ where *n=5* is the number of

[1] Thomas Bayes image provided by Wikimedia (link) under a CC attribution share alike license.

possible ratings, $x \in [1, 5]$ is a particular rating from 1 to 5, and *p* the probability of "success". The parameter to be estimated is *p*, and the expected rating becomes *np* = 5*p*. So a more intuitive explanation for *p* is that it represents the average rating on a scale of [0, 1].

Let's first see how the binomial distribution fits the reviews from ISyE 6420. The ratings for this course[2] were modeled[3] in Pymc3 with a very simple uniform prior: *p ~ U(0, 1)*.





We see that the difficulty ratings are a much better match for the binomial distribution than the quality ratings, as evidenced visually and by the calculated MSE provided at the top right of each chart. The concern then is whether the binomial distribution has enough representational power to properly model ratings with bimodal or other complex distributions.

Simulated Relative Ratings

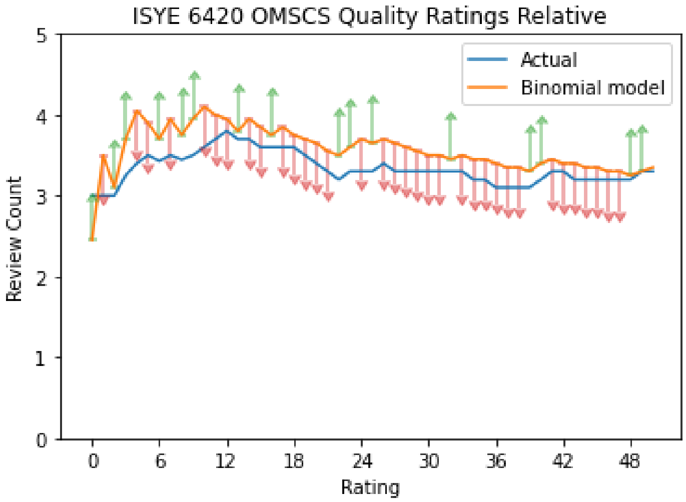
Since the proposed rating system has not actually been implemented, we will simulate it. This has the advantage of providing a ground truth to determine how much error a relative rating system might introduce.

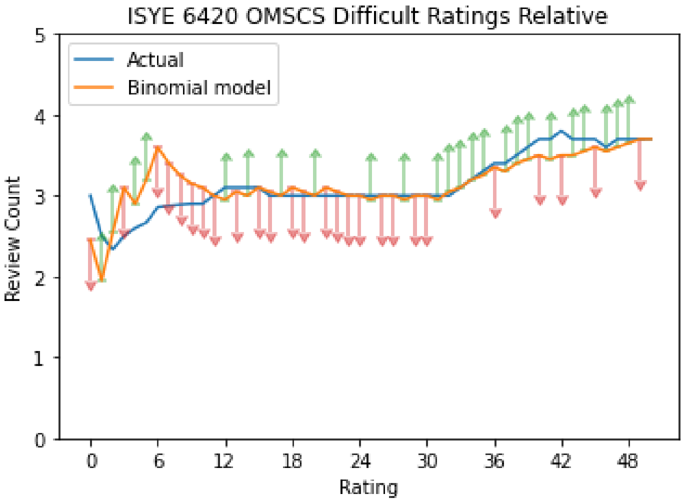
Real rating observations were converted into relative ratings by the following method:

```
1. p=0.5
2. Obs=[]
3. For every real rating r[i]:
      a. R = 5 * p
      b. If r[i] < R, then add censored observation [0, p) to Obs
      c. Else add censored observation (p, 1] to Obs
      d. Simulate binomial with censored observations Obs[4]
      e. Update p from posterior
```

[2] Ratings manually transcribed into numbers from OMSCentral. For example, "strongly liked" becomes 5.
[3] Modeling was done on the range [0, 4] then converted back for plotting.
[4] This simulation with censored observations was done in Stan with the same simple uniform prior: *p ~ U(0, 1)*. Refer to the attached `censored_binomial.stan`.

When we simulate what the relative ratings would have been for ISyE (below), we see a much better agreement between absolute and relative rating systems. Note that the absolute rating is a moving average of the last 10 ratings.


ISYE 6420 OMSCS Quality Ratings Relative


ISYE 6420 OMSCS Difficult Ratings Relative

In each chart, the blue and orange lines represent the actual and simulated relative ratings, respectively. The green and red arrows pointing up or down each represent a censored rating. For an up arrow this represents that the current rating was deemed too low by the last reviewer and so a censored observation of (p, 1] was added. Similar for red arrows.

In the difficulty rating chart, we can see the responsiveness of relative ratings to the increase in course difficulty after 30 reviews. Note that when the *Actual* rating was computed from all absolute reviews and not just the last ten (not shown), the overall rating did not move up as quickly as the relative rating. This was due to the accumulated weight of early ratings, and future analysis can explore why relative ratings seem more responsive to changing trends.

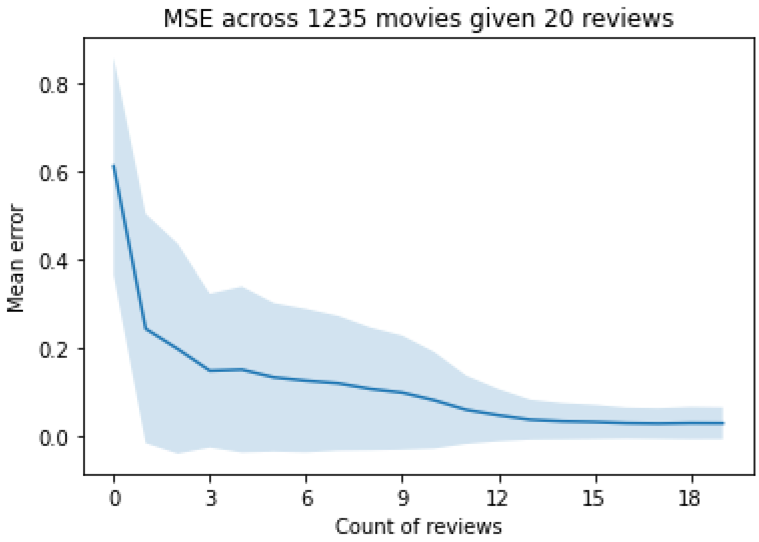
Verification with 1235 Movie Reviews

Based on the few examples so far, it seems the relative rating system yields a final result that closely matches absolute rating systems. However, a more systematic appraisal is warranted. So 1235 movies[5] with at least 20 reviews were given simulated relative ratings. The chart below shows the average MSE across all movies by the review count. Since the order of absolute reviews is not important to their overall score, the ordering of reviews for each movie on the chart is deterministic but arbitrary.

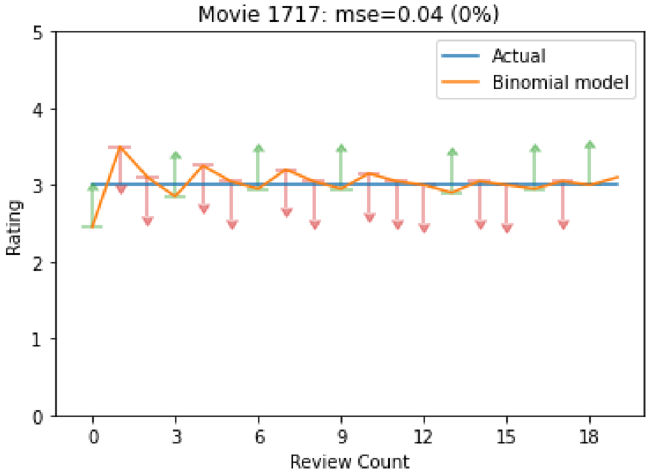
We can see in the chart below that as the number of reviews increases, MSE and it's variance decreases. More specifically after twelve reviews, the average MSE is within 0.05 stars of the ground truth absolute rating. The standard deviation for twelve or more reviews is +/- 0.06. This

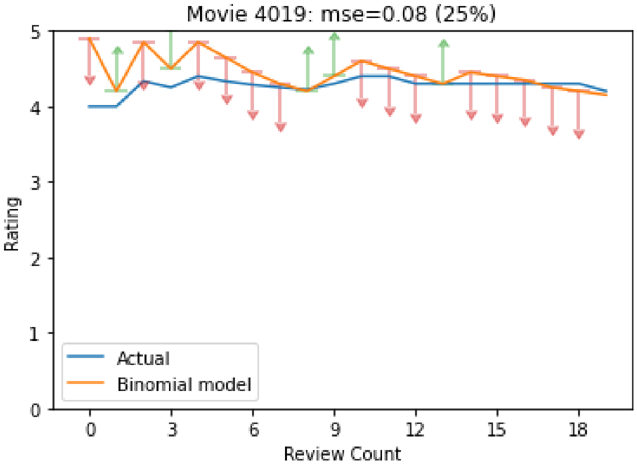
[5] Movie reviews from “small” MovieLens dataset: https://grouplens.org/datasets/movielens/

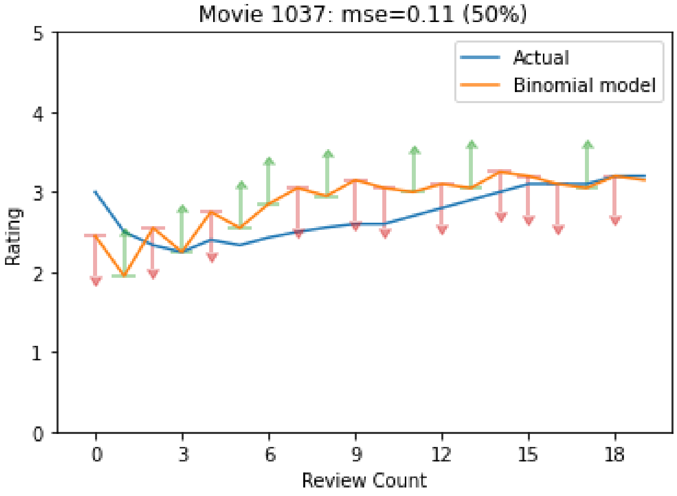
demonstrates a close correspondence between the ubiquitous absolute rating systems and the proposed relative rating system.

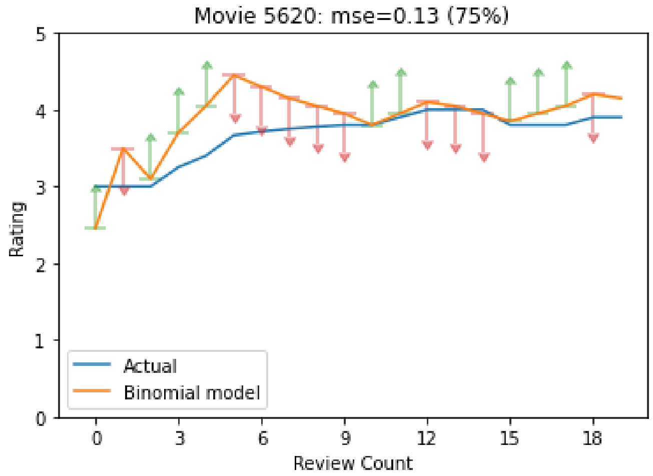


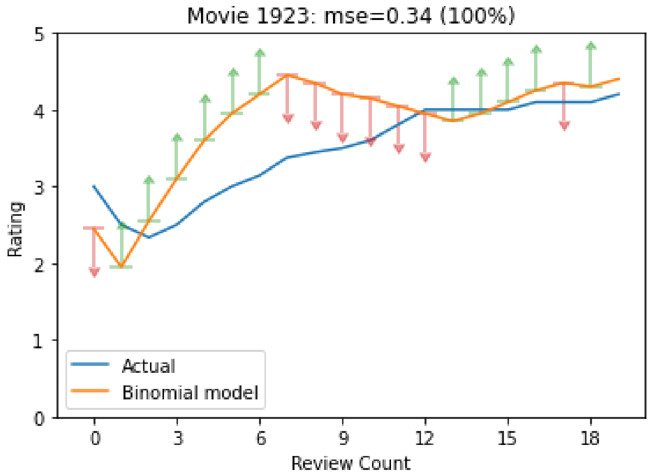
It’s helpful to pick out a few movies and visualize their absolute and relative rating trends. Five movies were selected based on the percentile of their MSE. The percentiles chosen were 0%, 25%, 50%, 75%, and 100% with respective MSEs of 0.04, 0.08, 0.11, 0.13, and 0.34:











Movies with lower MSE seem to have more consistent ratings--that is early and later ratings are largely the same. But when there is a run of high ratings, followed by a run of low ratings (or many low followed by many high ratings), then the relative rating system is more responsive to trends in recent reviews than the absolute rating system. This leads to a higher computed MSE between relative and absolute ratings for these movies. This may not be problematic however. Such responsiveness to more recent reviews could be seen as a feature rather than a deficiency of the proposed relative rating system.

Conclusion

In the proposed relative rating system, each user's rating can be seen as a reaction to the prior aggregated rating rather than an independent absolute rating. This has the benefit of simplifying the user's number of choices from five to two. Furthermore, implementation of relative ratings may encourage more measured ratings than absolute ratings which typically skew to the extremes[6].

We've shown how such a relative rating system can be implemented using Bayesian estimation of a Binomial distribution with each relative rating being considered a censored observation. Comparison to a large existing rating dataset shows that absolute and relative rating systems will likely converge to within a small delta within twelve ratings.

[6] For example, on Amazon, over 50% of reviews are 5 stars, and under 14% of reviews are 2 or 3 stars: https://www.researchgate.net/figure/Illustration-of-star-rating-distribution-for-Amazon-website-unbalanced-dataset-of-mobile_fig1_334185442