Order Matters: Accounting for Anchoring Bias on User Ratings in Recommendation Systems

Greg Borenstein
Massachusetts Institute of Technology
Media Lab
Cambridge, Massachusetts
gregab@mit.edu

ABSTRACT

Many supervised learning systems ask users to make sequences of judgments to produce labels for training. Recommendation systems in particular often prompt the user to rate a series of items in a single session. Most systems assume that such judgments are insensitive to the conditions under which they were elicitied. However there is significant evidence from the psychology of judgment that such decisions are affected by various types of bias. This paper demonstrates evidence of one of these, anchoring bias, in movie ratings provided in training a recommendation system and introduces a method for countering that bias to improve recommendations.

INTRODUCTION

Recommendation systems are rare amongst machine learning systems in that they attempt to learn not from user's impartial descriptions, but from their subjective judgments. Most recommendation systems ask users to provide a series of ratings of items from which the system attempts to predict the user's further preferences. In order to produce these ratings users must perform a series of judgments of their own past experiences and potential future preferences. These judgments are significantly less certain than the kind of objective observations usually required to produce labels for most supervised learning applications.

Most recommendation systems treat these judgments as independent from the conditions in which they were elicitied. However, psychological researchers have shown that judgments made under uncertain conditions are highly sensitive to the specifics of how they are elicited and often subject to systematic errors that lead to various forms of bias.[10]

This study examines one of these forms of bias that is particularly relevant for recommendation systems: anchoring bias.[10, 3, 8] Specifically, this study attempts to demonstrate that user ratings are affected by the order in which items are presented to them to be rated. Rather than considering each item independently, users treat the prior item as an anchor and adjust away from it, producing a rating that is biased towards their rating for the prior item.

To demonstrate the existence of this anchoring bias, we analyze the MovieLens dataset of 100,000 movie ratings[7]. We show a significant difference between the observed sequences of ratings and those that would be expected if the order of presentation had no effect.

In order to begin to mitigate the effect of this bias, we learn a classifier that incorporates the prior rating into its prediction. We demonstrate that this order-aware classifier outperforms an order-ignorant prediction.

In conclusion, we explore how this method for predicting the impact of anchoring could be integrated into a comprehensive recommendation system and present the outline for a future experiment to be conducted to test whether this produces improved recommendations.

BACKGROUND AND HYPOTHESIS

The psychology of judgment and decision-making has provided a framework for understanding how and under what conditions human judgments are subject to bias. In [10], Kahneman and Tversky argue that when "people assess the probability of an uncertain event or the value of an uncertain quantity" they "rely on a limited number of heuristic principles" that "sometimes lead to severe and systematic errors". These conditions have been found to hold in a wide variety of circumstances.[5]

It is well-established that consumer preferences are not constant but constructed at the time of elicitation.[6]. Hence, when providing ratings for recommendation systems users cannot simply access some known and fixed opinion. Instead, they construct their preferences on the fly while using the recommendation system's interface. As such, these ratings fit Kahneman and Tversky's criteria for judgments that are subject to bias.

But what kind of bias is likely to effect ratings in a recommendation system? Nearly all recommendation systems request ratings from users in a context that includes multiple items. Many recommendation systems inlcude an initial

NOTE: This is an unpublished paper. It has not been subjected to peer review. It was written as part of the Interactive Machine Learning class taught by Brad Knox at the MIT Media Lab in the fall of 2013.

training phase where a new user is asked to rate a series of items in in a single session in order to initialize the recommendation algorithm to their tastes (See Figure 3).

In situations like these, where judgments are made in the context of a prior value, the psychological literature leads us to suspect the presence of anchoring bias. Anchoring bias occurs when "people make estimates by starting from an initial value that is adjusted to yield the final answer."[10] Their adjustments are usually insufficient[9] resulting in judgments that are biased towards the prior value.

In the context of producing ratings to train a recommendation system, we hypothesize that the previous item presented acts as an anchor on the subsequent item, pulling the rating the user assign to the later item towards the value they assigned to the previous one. Secondarily, we hypothesize that by incoprorating the previous rating's anchoring effect into each new user-assigned rating before submitting it to a recommendation system we might be able to produce superior recommendations.

RELATED WORK

While most studies of machine learning and recommendation systems elide these psychological and behavioral forces entirely, a few studies have examined the effect of anchoring bias on various components.

Adomavicius, Bockstedt, Curley, and Zhang conducted a series of controlled experiments that demonstrated the anchoring effect of predicted ratings generated by recommendation systems on the eventual ratings given by users after having consumed the items[1]. In [4], Cosley et al demonstrated this same effect on the MovieLens interface specifically, though their study is not framed in terms of anchoring bias specifically.

Slightly further afield, in [2] Cardie uses a model of cognitive bias to guide automatic feature selection for a natural language processing-based machine learning system.

Our Contribution

The contribution of this paper differs from the related work mentioned here in two ways. First, it demonstrates the presence of anchoring bias not at the time of consumption of recommendations, but when the user provides ratings to train the recommendation system. Unlike the anchoring bias described in [1], the version shown here effects the overall quality of the recommendation system's rating predictions rather than the user's enjoyment of a particular item.

Secondly, this study offers not just evidence of a user behavior that might effect the design of recommendation system interfaces, but a statistical approach designed to determine the effect of bias on the user's submitted rating so it can be corrected in order to produce improved recommendations regardless of choice of interface or recommendation algorithm.

EVIDENCE OF ANCHORING BIAS

Method

In order to demonstrate evidence of anchoring bias in an existing body of ratings, we construct a procedure to compare

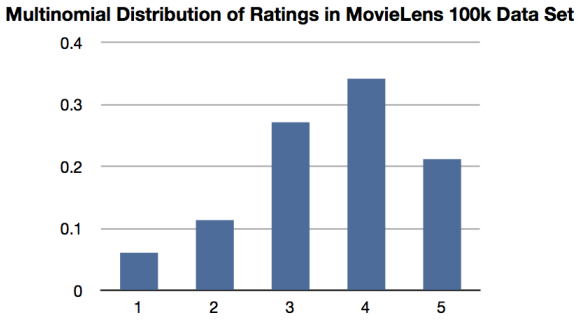

Figure 1. Multinomial distributions of ratings in the MovieLens 100k data set.

the observed frequency of pairs of sequential ratings with the expected probabilities of such pairs if the order had no effect on the rating. Detecting a significant difference between this predicted probability and the observed pair frequency would indicate that the rating pairs were unlikely to have ocurred by chance and hence that the presentation order is having an anchoring effect.

We used the MovieLens 100k rating set which consists of 100,000 ratings ranging from 1-5 assigned by 943 users to 1682 movies with each user rating a minimum of 20 movies.[7]

Predicted Pair Probabilities Under Order Independence

To predict the pair probabilities we'd expect to see in the absence of an anchoring effect we begin by calculating a multinomial distribution for the rating set (See Figure 1). This distribution expresses the probability of seeing any given value when randomly selecting a single rating. Assuming each rating is an independent event, the probability of any sequential pair of ratings is simply the product of their individual probabilities in the multinomial distribution

$$P(r_1, r_2) = P(r_1) * P(r_2)$$

This results in the predicted probability of rating pairs shown in Figure 2. This distribution provides a baseline for comparison with the pair probabilities observed in the data.

Observed Pair Probabilities

In order for anchoring bias to operate, the anchor must be shown immediately before the subsequent judgment is requested. In the MovieLens data set there is a wide distribution of time intervals between adjacent ratings from each user (see Figure 3). To eliminate ratings where this time interval would be so great as to prevent any anchoring effect, we selected only sequential ratings from the same user that were created with less than a 200 second gap. The great majority of ratings were created with gaps beneath this threshold, confirming the suspicion articulated above in the Background section that most ratings are created in extended multi-rating sessions.

Once filtered in this manner, these observed rating pairs produce a frequency distribution that we can compare with the

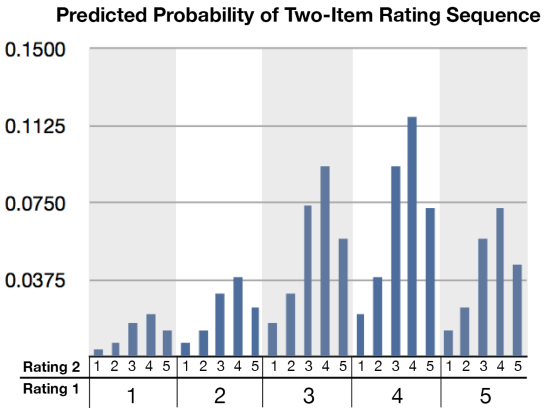

Figure 2. Predicted probability of sequential pairs of ratings assuming order has no effect. Calculated using the multinomial distribution of ratings.

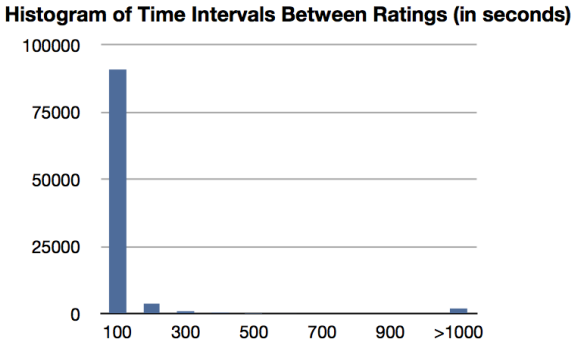

Figure 3. The MovieLens data shows a wide distribution of time intervals between adjacent ratings from the same user. To look for anchoring effects we filter out ratings separated by more than 200 seconds.

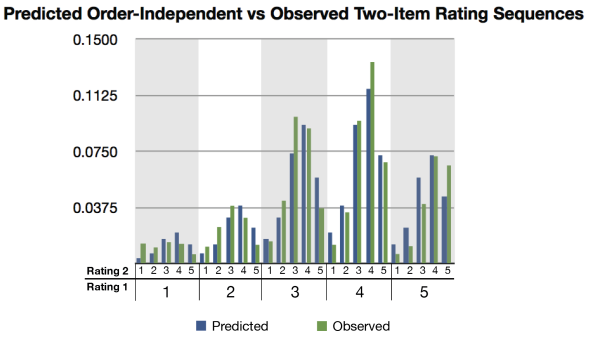

Figure 4. Observed sequential pairs of ratings compared to the pair frequencies predicted if there was no ordering effect. Calculated using the multinomial distribution of ratings.

pair probabilities predicted above in order to demonstrate evidence of anchoring bias (see Figure 4).

Evaluation

When we compare the predicted order-independent distribtuion to the observed result, we see significant evidence of an anchoring effect when the prior rating is a 1, 2, or 5.

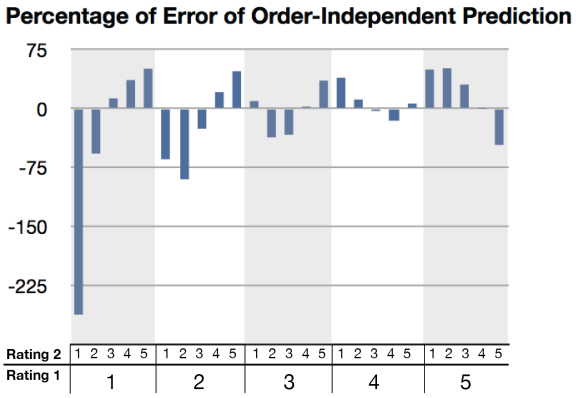

Figure 5. Percentage error of the order-independent distribution for each sequential pair of ratings. The order-independent distribution significantly under predicts the occurrence of ratings of 1 and 2 following a 1 and 2 and of 5s following a 5.

Figure 5, which shows the percentage error of the order-independent prediction on a per-pair basis, demonstrates the effect most clearly. When the anchoring rating is a 1, the order-independent distribution under-predicts the proportion of 1-star ratings that follows by 261.5% and when the anchoring rating is 2, it under-predicts 2-star ratings by 89.7%. In both of these conditions, this distribution conversely over-predicts the proportion of 4- and 5-star ratings.

With a prior rating of 3 or 4 stars, the order-independent prediction fairs much better, producing lower rates of error and distributing them less systematically. While it does under-predict following ratings of 3 and 4 stars respectively in these cases, it does so by a much smaller amount (33.19% and 15.35%), which could be explained by reversion to the mean particularly as 4- and 3-star ratings are the two most common ratings.

When the prior rating is 5 stars, we again see the clear pattern of anchoring emerge. The order-independent distribution over-predicts low ratings (1-and 2-star ratings by 49.65% and 51.24% respectively). Simultaneously it under-predicts 5-star ratings by a similar amount (46.1%). As in the cases of prior ratings of 1 and 2 stars we see a distribution of error in the order-ignorant prediction that demonstrates not reversion to the mean, but an anchoring towards the prior rating.

ADJUSTING FOR ANCHORING BIAS

Method

Now that we'e seen evidence for the existence of anchoring bias, how would we go about adjusting for it? We can approach the problem in machine learning terms as follows. First learn a baseline classifier that predicts ratings directly based on the observed distribution of ratings without reference to the prior rating. Then, learn a second classifier which is a function of both the current and prior ratings. Can this second classifier outperform the first one in predicting further ratings? If so, then that provides further evidence of anchoring bias and suggests that such a classifier could act as an adjustment to user ratings that would correct for anchoring bias

and improve the results of any subsequent recommendation calculation based on those ratings.

To evaluate these classifiers, we select the ratings that participated in continuous sequences of user ratings (as described above under Observed Pair Probabilities). We then separate these ratings into training and testing sets (with 80% of the data used for training and 20% for testing). We then train each of these classifiers on the training set and compare their results on the testing set in order to evaluate them.

Learning an Order-Ignorant Classifier

We can learn a baseline order-ignorant classifier by selecting the constant prediction that produces the least mean squared error across the entire data set. We select this prediction using:

$$\underset{r}{\operatorname{argmin}} \sum_{i=1}^{5} P(R=i) * (r-i)^2$$

where *r* is each potential rating prediction (1-5) and *P(R=i)* is the probability of seeing a rating of *i* (calculated, as before, based on the multinomial distribution). This product represents each prediction's total mean squared error for each rating choice it enocunters multiplied by how frequently that choice appears in the data. The summation of this product across all possible rating choices will produce the total squared error for that candidate prediction (and dividing by the number of samples will produce the mean squared error). We'll select the prediction with the lowest such error as our constant prediction.

We see the outcome of this calculation in Figure 6. The constant prediction of 3 stars barely outperforms that of 4. As we saw in Figure 1, 4-star ratings are the most common. In this case, that frequency of ocurrence almost manages to supercede a 3-star prediction's advantage of being in the center of the range of possible ratings. We select 3 stars as our constant prediction.

Learning an Order-Aware Classifier

Unlike the constant order-ignorant classifier, our order-aware classifier will be a function of the prior rating. Other than this change, we can approach the problem in a similar manner to how we learned our order-ignorant classifier, simply replacing the probability of of the current rating with its conditional probability *given the prior rating*.

$$f(R_{n-1}) = \underset{r}{\operatorname{argmin}} \sum_{i=1}^{5} P(R_n = i | R_{n-1}) * (r-i)^2$$

where *r* is each potential rating prediction and $P(R_n{=}i|R_{n-1})$ is the conditional probability of seeing a rating of *i* given the observed prior rating.

To train this classifier, we first construct this conditional probability distribution by performing the same filtering of user sequences as described above in the Observed Pair Probabilities section, but this time on our training data set. We can use that distribution to calcuate the mean squared error for each subsequent rating given its prior and thence select the prediction for each prior with the lowest error. As shown in Figure 7, the results of this process are that we'll predict a rating of 3 stars for prior ratings of 1, 2, and 3 and a rating of 4 stars for prior ratings of 4 and 5.

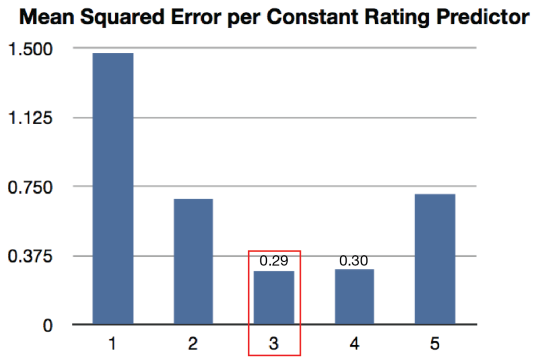

Figure 6. Mean squared error for each constant prediction candidate. We select a constant prediction of 3 stars, which just outperforms a prediction of 4 stars.

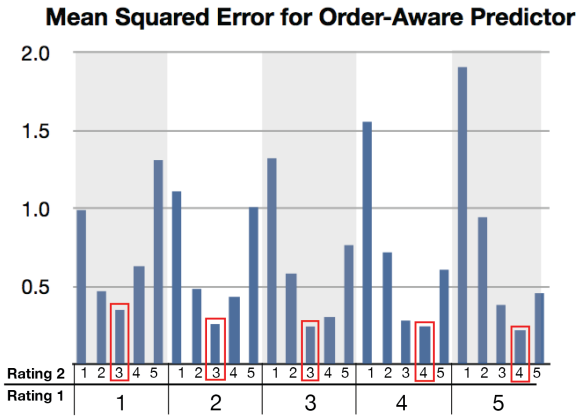

Figure 7. Mean squared error for order-aware predictor. Unlike the order-independent condition, this predictor selects a different prediction depending on the prior rating.

Evaluation

We can now compare the performance of these two classifiers using the test data we set aside at the beginning of the training process.

Specifically, we can conduct a paired t-test to determine if the two classifiers produce significantly different results and, if they do, to determine if the order-aware classifier performs

p-value	t-statistic	df	sd
4.5959e-24	-10.1327	17814	1.6022

Table 1. Results of paired t-test between squared error for order-ignorant and order-aware condition. The negative t-stat indicates the reduction of error in the order-aware condition.

significantly better than the order-ignorant one. To do this, we run both of our classifiers against each sample in the test set and record the squared error for each condition. We treat these as the pairs for our paired t-test, which produces the results seen in Table 1.

A p-value of *4.5959e-24* means we can safely reject the null hypothesis and conclude that the two classifiers produce significantly different predictions. Further, since this t-test was run comparing the squared error of the order-aware classifier to that of the order-ignorant one, the negative sign on the t-statistic indicates that the order-aware classifier reduced errors relative to the order-independent one.

DISCUSSION AND FUTURE WORK

We conducted analysis of a large body of rating data to demonstrate statistical evidence of anchoring bias affecting user ratings in the MovieLens recommendation system. We found evidence that a prior rating of 1, 2, and 5 stars shifted the distribution of subsequent ratings away from that likely to be seen by chance without an anchoring effect.

Further, we approached the problem from a machine learning perspective and demonstrated that a simple classifier that is aware of prior ratings significantly outperforms a similar classifier that is not.

Together, these results lay the groundwork for incorporating anchoring bias correction into real world recommendation systems. To that end, we here describe proposed future work designed to test the value of that integration.

Proposed Experiment to Test Recommendation Improvements

The chief goal of measuring the effect of anchoring bias on user ratings is to correct for that bias in order to produce superior recommendations using existing recommendation algorithms. We have shown the essence of that idea here in our comparison of order-ignorant and order-aware classifiers. The next step in this research is to integrate our anchoring effect corrections into a full-scale recommendation system and to test if such an addition produces recommendations that users prefer.

Such an investigation could be conducted by using the conditional probability-based order-aware classifier demonstrated here to calculate an adjustment factor that would be applied to each rating before submitting it to an existing recommendation system. The predictions based on these anchor-adjusted ratings could then be compared to an identical recommendation system trained on the raw ratings to determine which recommendations users prefer. In order to avoid further anchoring effects and to detect the relatively small changes in rating prediction caused by anchoring correction, users could be shown pairs of items for which each classifier made close predictions and asked to select which they prefer. The classifier whose predictions were overturned less frequently would be considered to be prefered.

This improvement in rating predictions would be small as measured in fractional ratings, but could lead to signficantly altered user behavior. Recommendations are often presented in the form of a list ranked by predicted rating where only the n-items predicted to have the highest rating are actually seen by the user. Accounting for anchoring bias may cause different items to make this cutoff and hence experiene a large difference in user attention.

CONCLUSION

The application of the psychology of judgment and decision-making to recommendation systems and interactive machine learning generally has barely begun. We hope that this paper, in addition to its specific contribution to improving recommendation systems by accounting for anchoring bias, may open up a wider discussion about methods for incorporating the long list of known biases into many learning systems based on user judgments.

REFERENCES

1. Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. Recommender systems, consumer preferences, and anchoring effects. In *Decisions@RecSys11* (2011).
2. Cardie, C. A Cognitive Bias Approach to Feature Selection and Weighting for Case-Based Learners. *Machine Learning 41* (2000), 85–116.
3. Chapman, G., and Johnson, E. Incorporating the Irrelevant: Anchors in Judgments of Belief and Value. In *The psychology of intuitive judgment: Heuristics and biases*, T. Gilovich, D. Griffin, and D. Kahneman, Eds. Cambridge University Press, New York, 2000.
4. Cosley, D., Lam, S., Albert, I., Konstan, J., and Riedl, J. Is seeing believing? how recommender interfaces affect users opinions. In *CHI* (2003).
5. Epleya, N., and Gilovich, T. Anchoring unbound. *Journal of Consumer Psychology 20* (2010), 20–24.
6. Lichtenstein, S., and Slovic, P., Eds. *The construction of preference*. Cambridge University Press, 2006.
7. The MovieLens 100k data set. `http://grouplens.org/datasets/movielens/`.
8. Mussweiler, T., and Strack, F. Numeric Judgments under Uncertainty: The Role of Knowledge in Anchoring. *Journal of Experimental Social Psychology 36* (2000), 495518.
9. Slovic, P., and Lichtenstein, S. *Organizational Behavior and Human Performance 6* (1971).
10. Tversky, A., and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science 4157* (1974), 1124–1131.