5 When can I trust an average rating on Amazon?

Starting with this chapter, in the next four chapters we will walk through a remarkable landscape of intellectual foundations. But sometimes we will also see significant gaps between theory and practice.

5.1 A Short Answer

We continue with the theme of recommendation. Webpage ranking in Chapter 3 turns a graph into a ranked order list of nodes. Movie ranking in Chapter 4 turns a weighted bipartite user-movie graph into a set of ranked order lists of movies, with one list per user. We now examine the aggregation of a vector of rating scores by reviewers of a product or service, and turn that vector into a scalar, one per product. These scalars may in turn be used to rank order a set of similar products. In Chapter 6, we will further study aggregation of many vectors into a single vector.

When you shop on Amazon, likely you will pay attention to the number of stars shown below each product. But you should also care about the number of reviews behind that averaged number of stars. Intuitively, you know that a product with 2 reviews, both 5 stars, may not be better than a competing product with 100 reviews and an average of 4.5 stars, especially if these 100 reviews are all 4 and 5 stars and the reviewers are somewhat trustworthy. We will see how such intuition can be sharpened.

In most online review systems, each review consists of three fields:

1. Rating (a numerical score often on the scale of 1-5 stars). This is the focus of our study.
2. Review (text).
3. Review of review (often a binary up or down vote).

Rarely do people have time to read through all the reviews, so a summary review is needed to aggregate the individual reviews. What is a proper aggregation? That is the subject of this chapter.

Ratings are often not very trustworthy, and yet they are important in so many contexts, from peer reviews in academia to online purchases of every kind. The hope is that the following two approaches can help:

First, we need methods to ensure some level of accuracy, screening out the really bad ones. Unlimited and anonymous reviews have notoriously poor quality, because a competitor may enter many negative reviews, the seller herself may enter many positive reviews, or someone who has never even used the product or service may enter random reviews. So before anything else, we should first check the mechanism used to enter reviews. How strongly are customers encouraged, or even rewarded, to review? Do you need to enter a review of reviews before you are allowed to upload your own review? Sometimes a seemingly minor change in formatting leads to significant differences: Is it a binary review of thumbs up or down, followed by a tally of up vs. down votes? What is the dynamic range of the numerical scale? It has been observed that the scale of 1-10 often returns 7 as the average and then a bimodal distribution around it. A scale of 1-3 gives a very different psychological hint to the reviewers compared to a scale of 1-5, or a scale of 1-10 compared to -5 to 5.

Second, the review population size needs to be large enough to wash out the inaccurate ones. But *how* large is large enough? And can we run the raw ratings through some signal processing to get the most useful aggregation?

These are tough questions with no good answers yet, not even well formulated problem statements. The first question depends on the nature of the product being reviewed. Movies (*e.g.*, on IMDB) are very subjective, whereas electronics (*e.g.*, on Amazon) are much less so, with hotels (*e.g.*, on tripadvisor) and restaurants (*e.g.*, on opentable) somewhere in between. It also depends on the quality of the review, although reputation of the reviewer is a difficult metric to quantify in its own right.

The second question depends on the metric of "usefulness." Each user may have a different metric, and the provider of the service or product may use yet another one. This lack of clarity in what should be optimized is the crux of the ill-definedness of the problem at hand.

With these challenges, it may feel like opinion aggregation is unlikely to work well. But there have been notable exceptions recorded for some special cases. A famous example is Galton's 1906 observation on a farm in Plymouth, UK, where 787 people in a festival there participated in a game of guessing the weight of an ox, each writing down a number *independent* of others. There was also no common bias; everyone could take a good look at the ox. While the estimates by each individual were all over the place, the average was 1197 pounds. It turned out the ox weighed 1198 pounds. Just a simple averaging worked remarkably well. For the task of guessing the weight of an ox, 787 was more than enough to get the right answer (within a margin of error of 0.1%).

But in many other contexts, the story is not quite as simple as Galton's experiment. There were several key factors here that made simple averaging work so well:

- The task is relatively easy; in particular, there is a correct objective answer with a clear numerical meaning.

- The estimates are both unbiased and independent of each other.
- There are enough people participating.

More generally, three factors are important in aggregating individual action:

- *Definition of the task*: Guessing a number is easy. Consensus formation in social choice is hard. Reviewing a product on Amazon is somewhere in between. Maybe we can define "subjectivity" by the size of the review population needed to reach a certain "stabilization number."
- *Independence of reviews*: As we will see, the wisdom of crowds, if there is one to the degree we can identify and quantify, stems not from having many smart individuals in the crowd, but from the independence of each individual's view from the rest. Are Amazon reviews independent of each other? Kind of. Even though you can see the existing reviews before entering your own, usually your rating number will not be significantly affected by the existing ratings. Sometimes, reviews are indeed entered as a reaction to recent reviews posted on the website, either to counter-argue or to reinforce points made there. This influence from the sequential nature of review systems will be partially studied in Chapter 7.
- *Review population*: For a given task and degree of independence, there is correspondingly a minimum number of reviews, a threshold, needed to give a target confidence of trustworthiness to the average. If these ratings pass through some signal processing filters first, then this threshold may be reduced.

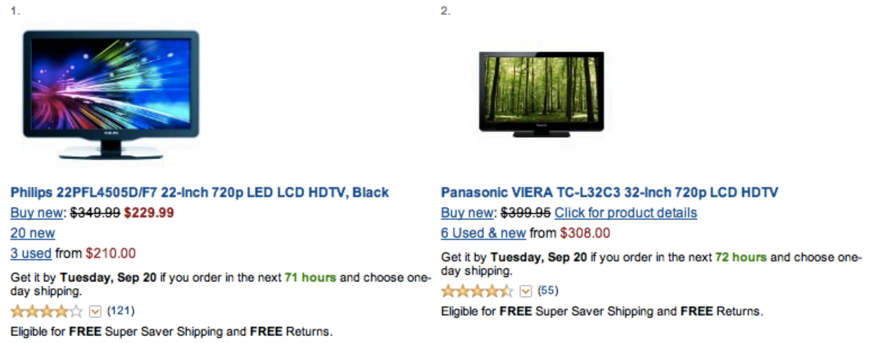
What kind of signal processing do we need? For text reviews, there needs to be natural language tools, *e.g.*, detecting inconsistencies or extreme emotions in a review and discounting it and its associated rating. We in academia face this problem in each decision on a peer-reviewed paper, a funding proposal, a tenure-track position interview, and a tenure or promotion case.

For rating numbers, some kind of weighting is needed, and we will discuss a particularly well-studied one soon. In Chapter 6, we will also discuss voting methods, including majority rule, pairwise comparison, and positional counting. These voting systems require each voter to provide a complete ranking, possibly implicitly, with a numerical rating scale. Therefore, we will have more information, perhaps too much information, as compared to our current problem.

5.2 Challenges of rating aggregation

Back to rating aggregation. Here are several examples illustrating three of the key challenges in deciding when to trust ratings on Amazon.

Example 1. Many online rating systems use a naive averaging method for their product ratings. Moreover, given that different products have different numbers of reviews, it is hard to determine which product has a better quality. For example, in Figure 5.1, Philips 22PFL4504D HDTV has 121 ratings with a mean of 4, while Panasonic VIERA TC-L32C3 HDTV has 55 ratings with a mean of 4.5. So the customer is faced with a tradeoff between choosing a product with a lower average rating and a larger number of reviews versus one with a higher average rating and a smaller number of reviews.



1.

Philips 22PFL4505D/F7 22-Inch 720p LED LCD HDTV, Black
Buy new: ~~$349.99~~ $229.99
20 new
3 used from $210.00

Get it by **Tuesday, Sep 20** if you order in the next 71 hours and choose one-day shipping.

★★★★☆ (121)

Eligible for **FREE** Super Saver Shipping and **FREE** Returns.

2.

Panasonic VIERA TC-L32C3 32-Inch 720p LCD HDTV
Buy new: ~~$399.95~~ Click for product details
6 Used & new from $308.00

Get it by **Tuesday, Sep 20** if you order in the next 72 hours and choose one-day shipping.

★★★★☆ (55)

Eligible for **FREE** Super Saver Shipping and **FREE** Returns.

Figure 5.1 Tradeoff between review population and average rating score. Should a product with fewer reviews but higher average rating be ranked higher than a competing product with more ratings but lower average rating?

Example 2. Consider 2 speaker systems for home theater. Both RCA RT151 and Pyle Home PCB3BK have comparable mean scores around 4. 51.9% of users gave RCA RT151 a rating of 5 stars while 7.69% gave 1 star. On the other hand, 54.2% of users gave 5 stars to Pyle Home PCB3BK while 8.4% gave 1 star. So Pyle Home PCB3BK speaker has not only a higher percentage of people giving it 5 stars than RCA RT151, but also has a higher percentage of people giving it 1 star. There is a larger *variation* in the ratings of Pyle Home PCB3BK than RCA RT151.

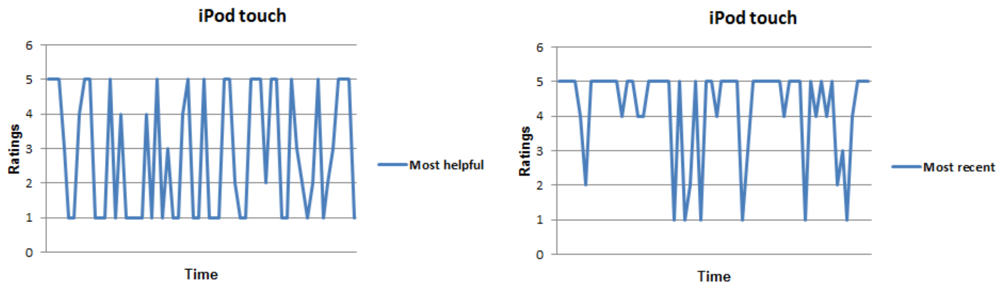


Figure 5.2 How to view the aggregated ratings: should it be based on helpful ratings or on the latest trend? Here the same set of iPod touch ratings on Amazon is used to extract two different subsets of ratings, and their values are quite different.

Example 3. In Figure 5.2, we compare the mean rating of the first 60 "most helpful" reviews of iPod3 Touch (32 GB) on Amazon with the mean from the 60 most recent ratings. The ratings are on the y-axis and the times of the ratings (index) are on the x-axis. The mean of the most recent ratings is 1.5 times greater than the mean corresponding to the most helpful reviews. Is this a "real" change or just noise and normal fluctuation? What should the timescale of averaging be?

At the heart of these problems is the challenge of turning *vectors* into *scalars*, which we will meet again in Chapter 6. This can be a "lossy compression" with very different results depending on how we run the process, *e.g.*, just look at the difference between mean and median.

5.3 Beyond basic aggregation of ratings

We may run a time-series analysis to understand the dynamics of rating. In Figure 5.3, the three curves of ratings entered over a period of time give the same average, but "clearly" some of them have not converged to a stable average rating. What kind of *moving window size* should we use to account for cumulative average and variance over time?

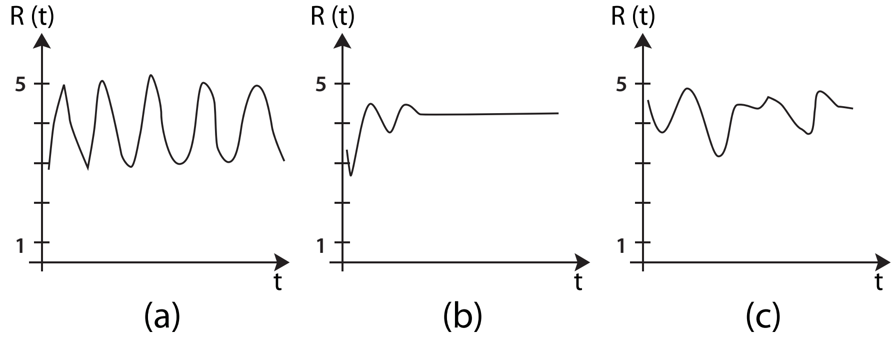


Figure 5.3 Three time series with the same long-term average rating but very different stabilization behavior. The time axis scale is in the order of weeks. (a) shows continued cyclic fluctuations of ratings R over time t. (b) shows a clear convergence. (c) shows promising signs of convergence but it is far from clear that the ratings have converged. Of course, the timescale also matters.

We may consider detecting anomalous ratings and throwing out the highly suspicious ones. If we detect a trend change, that may indicate a change of ownership or generational upgrade. And if such detection is accurate enough, we can significantly discount the ratings before this time. For ratings on the scale of 1-5, the coarse granularity makes this detection more difficult.

We may consider zooming into particular areas of this vector of ratings, *e.g.*, the very satisfied customers and the very dissatisfied ones, although it is often the case that those who care enough to enter ratings are either extremely satisfied or reasonably dissatisfied. There might be a bimodal distribution in the underlying customer satisfaction for certain products, but for many products there is often another bimodal distribution on the biased sampling based on who cared enough to write reviews.

Across all these questions, we can use the cross-validation approach from Chapter 4 to train and test the solution approach. If we can stand back 1 year and predict the general shape of ratings that have unfolded since then, that would be a strong indicator of the utility of our signal processing method.

But these questions do not have well-studied answers yet, so we will now focus instead on some simpler questions as proxies to our real questions: Why does simple averaging sometimes work, and what to do when it does not?

5.4 A Long Answer

5.4.1 Averaging a crowd

We start from a significantly simplified problem. Take the Galton example, and say the number that a crowd of N people wants to guess is x, and each person i in the crowd makes a guess y_i:

$$y_i(x) = x + \epsilon_i(x),$$

i.e., the true value plus some error ϵ_i. The error depends on x but not other j; it is independent of other errors. This error can be positive or negative, but we assume that it averages across different x to be 0; it has no bias. In reality, errors are often neither independent nor unbiased. Sequential estimates based on publicly announced estimates made by others may further exacerbate the dependence and bias. We will see examples of such information cascades in Chapter 7.

We measure error by mean squared error (MSE), just like what we did in Chapter 4. We want to compare the following two quantities:

- The average of individual guesses' errors.
- The error of the averaged guess.

The average of errors and the error of the average are not the same, and we will see how much they differ. Since x is a number that can take on different values with different probabilities, we should talk about the *expected MSE*, where the expectation $\mathbf{E}_x$ is the averaging procedure over the probability distribution of x.

The average of (expected, mean squared) errors (AE), by definition, is

$$E_{AE} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{E}_x \left[\epsilon_i^2(x) \right]. \tag{5.1}$$

On the other hand, the (expected, mean squared) error of the average (EA) is

$$E_{EA} = \mathbf{E}_x \left[\left(\frac{1}{N} \sum_{i=1}^{N} \epsilon_i(x) \right)^2 \right] = \frac{1}{N^2} \mathbf{E}_x \left[\left(\sum_{i=1}^{N} \epsilon_i(x) \right)^2 \right] \tag{5.2}$$

since the error term is now

$$\frac{1}{N} \sum_i y_i - x = \frac{1}{N} \left(\sum_i y_i - Nx \right) = \frac{1}{N} \left(\sum_i (y_i - x) \right) = \frac{1}{N} \left(\sum_i \epsilon \right).$$

It looks like (5.1) and (5.2) are the same, but they are not: sum of squares and square of sum are different. Their difference is a special case of Jensen's inequality on convex quadratic functions. There are many terms in expanding the square in (5.2), some are ϵ_i^2, and others are $\epsilon_i \epsilon_j$ where $i \neq j$. For example, if $N = 2$, we have one cross-term:

$$(\epsilon_1 + \epsilon_2)^2 = \epsilon_1^2 + \epsilon_2^2 + 2\epsilon_1\epsilon_2.$$

These cross terms $\{\epsilon_i \epsilon_j\}$ take on different values depending on whether the estimates $\{y_i\}$ are independent or not. If they are independent, we have

$$\mathbf{E}_x[\epsilon_i(x)\epsilon_j(x)] = 0, \;\; \forall i \neq j.$$

In that case, all the cross terms in expanding the square are zero, and we have

$$E_{EA} = \frac{1}{N} E_{AE}. \tag{5.3}$$

If you take the square root of MSE to get RMSE, the scaling in (5.3) is then $1/\sqrt{N}$.

This may appear to be remarkable. In such a general setting and using such an elementary derivation, we have mathematically crystallized (a type of) the wisdom of crowds in terms of efficiency gain: error is reduced by a factor as large as the size of the crowd if we average the estimates first, provided that the estimates are *independent* of each other. But this result holds for a crowd of 2 as much as it holds for a crowd of 1000. Some people think there must be something beyond this analysis, which is essentially the Law of Large Numbers at work: variance is reduced as the number of estimates go up. There should be some type of the wisdom of crowd that only shows up for a large enough crowd. Furthermore, we have so far assumed there is no systematic bias; averaging will not help reduce any bias that is in everyone's estimate.

The $1/N$ factor is only one dimension of the wisdom of crowds, what we refer to as "multiplexing gain" from independent "channels." We will later see "diversity gain," symbolically summarized as $1 - (1 - p)^N$.

What if the estimates are completely *dependent*? Then the averaged estimate is just the same as each estimate, so the error is exactly the same, and it does not matter how many reviews you have:

$$E_{EA} = E_{AE}. \tag{5.4}$$

In most cases, the estimates are somewhere in between completely independent and completely dependent. We have seen that each person in the crowd can be quite wrong. What is important is that they are wrong in independent ways. In statistical terms, we want their pairwise correlations to be small. Of course, if we could identify who in the crowd have the correct estimates, we should just use their estimates. So the wisdom of crowds we discussed is more about achieving robustness arising out of independent randomization than getting it right by identifying the more trustworthy estimates.

5.4.2 Bayesian estimation

Bayesian analysis can help us quantify the intuition that the number of ratings, N, should matter. Let us first get a feel for the Bayesian view with a simple, illustrative example. We will then go from one product to many products being ranked in the Bayesian way.

Suppose you run an experiment that returns a number, and you run it n times (the same experiment and independent runs). Suppose that s times it returns an outcome of 1. What do you think is the chance that the next experiment, the $(n+1)$th one, will return an outcome of 1 too? Without going into the foundation of probability theory, the answer is the intuitive one:

$$\frac{s}{n}.$$

Now if you know the experiment is actually a flip of a biased coin, what do you think is the chance that the next experiment will be positive? Hold on, is that the same question?

Actually, it is not. Now you have a *prior* knowledge: you know there are two possible outcomes, one with probability p for head, and the other $1 - p$ for tail. That prior knowledge changes the derivation, and this is the essence of the Bayesian reasoning.

We first write down the probability distribution of p given that s out of n flips showed heads. Intuitively, the bigger s is, the more likely the coin is biased towards head, and the larger p is. This is the essence of the Bayesian view: more observations makes the model better.

Now, if p were *fixed*, then the probability of observing s heads and $n - s$ tails follows the Binomial distribution:

$$\left(\begin{array}{c} n \\ s \end{array} \right) p^s (1-p)^{n-s}. \tag{5.5}$$

So the probability distribution of p must be *proportional* to (5.5). This is the key step in Laplace's work that turned Bayes insights into a systematic mathematical language.

This is perhaps less straightforward that it may sound. We are "flipping the table" here. Instead of looking at the probability of observing s out of n heads for a given p, we looking at the probability distribution of p that gave rise to this observation in the first place, since we have the observation but not p.

Once the above realization is internalized in your brain, the rest is easy. The probability distribution of p is proportional to (5.5), but we need to divide it by a normalization constant so that it is between 0 and 1. Knowing $p \in [0, 1]$, the normalization constant is simply:

$$\int_0^1 \left(\begin{array}{c} n \\ s \end{array} \right) p^s (1-p)^{n-s} dp.$$

Using beta function to get the above integral, we have:

$$f(p) = \frac{\left(\begin{array}{c} n \\ s \end{array} \right) p^s (1-p)^{n-s}}{\int_0^1 \left(\begin{array}{c} n \\ s \end{array} \right) p^s (1-p)^{n-s} dp} = \frac{(n+1)!}{s!(n-s)!} p^s (1-p)^{n-s}.$$

Finally, since *conditional* probability of seeing a head given p is just p, the *unconditional* probability of seeing a head is simply

$$\int_0^1 p f(p) dp,$$

an integral that evaluates to

$$\frac{s+1}{n+2}.$$

A remarkable and remarkably simple answer, this is called the **rule of succession** in probability theory. It is perhaps somewhat unexpected. The intermediate step to understanding p's distribution is not directly visible in the final answer, but that was in the core of the innerworking of Bayesian analysis. It is similar in spirit to the latent factor model in Chapter 4, and to other hidden-factor models like hidden Markov models used in many applications, from voice recognition to portfolio optimization.

Why is it *not* s/n? One intuitive explanation is that if you know the outcome must be success or failure, it is as if you have already seen 2 experiments “for free”, 1 success and 1 failure. If you incorporate the prior knowledge in this way, then the same intuition on the case without prior knowledge indeed gives you $(s+1)/(n+2)$.

5.4.3 Bayesian ranking

So why is Bayesian analysis related to ratings on Amazon? Because ratings' population size matters. Back to our motivating question: should a product with only 2 reviews, even though both are 5 stars, be placed higher than a competing product with 100 reviews that averages 4.5 stars? Intuitively, this would be wrong. We should somehow weight the raw rating scores with the population sizes, just like we weighted a node's importance by the in-degree in Chapter

3. Knowing how many reviews there are gives us a prior knowledge, just like knowing a coin shows up heads 100 times out of 103 flips is a very different observation than knowing it shows up heads 3 times out of 13 flips.

More generally, we can think of a "sliding ruler" between the average rating of all the products, R, and the averaged rating of brand i, r_i. The more reviews there are for brand i relative to the total number of reviews for all the brands, the more trustworthy r_i is relative to R. The resulting Bayesian rating for brand i is

$$\tilde{r}_i = \frac{NR + n_i r_i}{N + n_i}. \tag{5.6}$$

We may also want to put an upper bound on N, for otherwise as time goes by and N monotonically increases, the dynamic range of the above ratio can only shrink.

Quite a few websites adopt Bayesian ranking. The Internet Movie DataBase (IMDB)'s top 250 movies ranking follows (5.6) exactly. So the prior knowledge used is the average of all the movie ratings.

Beer Advocate's top beer ranking *e.g.*, `http://beeradvocate.com/lists/popular` uses the following formula:

$$\frac{N_{min} R + n_i r_i}{N_{min} + n_i},$$

where N_{min} is the minimum number of reviews needed for a beer to be listed there. Perhaps a number in between N and N_{min} would have been a better choice, striking a tradeoff between following the Bayesian adjustment exactly and avoiding the saturation effect (when some beers get a disproportionately large numbers of reviews).

All of the above suffer from a drawback in their assuming that there is a single, "true" value of a product's ranking, as the mean of some Gaussian distribution. But some products simply create bipolar reactions: some love it and some hate it. The idea of Bayesian ranking can be extended to a multinomial model and the corresponding Dirichlet prior.

Of course, this methodology only applies to adjusting the ratings of each brand within a comparable family of products, so that proper ranking can be achieved based on $\{\tilde{r}_i\}$. It cannot adjust ratings without this backdrop of a whole family of products that provides the *scale* of relative trustworthiness of ratings. It is good for *ranking*, but not for *rating* refinement, and it does not take into account time series analysis.

5.5 Examples

5.5.1 Bayesian ranking changes order

Consider Table 5.5.1, a compilation of ratings and review populations for MacBooks. The items are listed in descending order of their average rating. Following

(5.6), the Bayesian rankings for each of the five items can be computed. First, we compute the product NR as follows:

$$NR = \sum_i n_i r_i = 10{\times}4.920{+}15{\times}4.667{+}228{\times}4.535{+}150{\times}4.310{+}124{\times}4.298 = 2332.752$$

Then, with $N = \sum_i n_i = 527$, we apply (5.6) to each of the items:

$$\bar{r}_1 = \frac{2332.752 + 10 \times 4.920}{527 + 10} = 4.436,$$

$$\bar{r}_2 = \frac{2332.752 + 15 \times 4.667}{527 + 15} = 4.433,$$

$$\bar{r}_3 = \frac{2332.752 + 228 \times 4.535}{527 + 228} = 4.459,$$

$$\bar{r}_4 = \frac{2332.752 + 150 \times 4.310}{527 + 150} = 4.401,$$

$$\bar{r}_5 = \frac{2332.752 + 124 \times 4.298}{527 + 124} = 4.402.$$

These calculations and the Bayesian-adjusted rankings are shown in Table 5.5.1. All of the MacBook's ranking positions change after the adjustment is applied, because Bayesian adjustment takes into account the number of reviews as well as the average rating for each item. The third MacBook (MB402LL/A) rises to the top because the first and second were rated by far less people, and as a result, the ratings of both of these items drop significantly.

MacBook	Total Ratings	Average Rating	Rank	Bayes Rating	Bayes Rank
MB991LL/A	10	4.920	1	4.436	2
MB403LL/A	15	4.667	2	4.433	3
MB402LL/A	228	4.535	3	4.459	1
MC204LL/A	150	4.310	4	4.401	5
MB061LL/A	124	4.298	5	4.402	4

Table 5.1 An example where average ratings and Bayesian-adjusted ratings lead to entirely different rankings of the items. For instance, though the first listed MacBook (MB991LL/A) has the highest average, this average is based on a small number of ratings (10), which lowers its Bayes ranking by two places.

5.5.2 Bayesian ranking quantifies subjectivity

Sometimes Bayesian adjustment does not alter the ranked order of a set of comparable products, but we can still look at the "distance" between the original average rating and the Bayesian adjusted average rating across these products, *e.g.*, using l-2 norm of the difference between these two vectors.

For example, the adjusted rating for a set of digital cameras and women's shoes is computed in Table 5.2 and Table 5.3, respectively. This distance, normalized by the number of products in the product category, is 0.041 for digital cameras and 0.049 for shoes. This difference of almost 20% is a quantified indicator about the higher subjectivity and stronger dependence on review population size for fashion goods compared to electronic goods.

Digital Camera	Number of reviews, n_i	Mean Rating, r_i	Bayesian Rating, $\tilde{r}_i$
Canon Powershot	392	4.301	4.133
Nikon S8000	163	3.852	4.008
Polaroid 10011P	168	3.627	3.965

Table 5.2 Bayesian adjustment of ratings for 3 digital cameras, with a total of 723 ratings. The L-2 distance between the vector of mean ratings and that of Bayesian ratings is 0.023.

Women's Shoes	Number of reviews, n_i	Mean Rating, r_i	Bayesian Rating, $\tilde{r}_i$
Easy Spirit Traveltime	150	3.967	4.182
UGG Classic Footwear	148	4.655	4.289
BearPaw Shearling Boots	201	4.134	4.204
Skechers Shape-Ups	186	4.344	4.245
Tamarac Slippers	120	3.967	4.189

Table 5.3 Bayesian adjustment of ratings for 5 women's fashion shows, with a total of 805 ratings. The L-2 distance between the vector of mean ratings and that of Bayesian ratings is 0.041, about twice the difference in the case of digital camera example.

5.5.3 What does Amazon do

On Amazon, each individual product rating shows only the raw scores (the averaged number of stars), but when it comes to ranking similar products by "average customer review," it actually follows some secret formula that combines raw score with three additional elements:

- Bayesian adjustment by review population

- Recency of the reviews
- Reputation score of the reviewer (or quality of review, as reflected in review of review). See for example `www.amazon.com/review/top-reviewers-classic` for the hall of fame of Amazon reviewers.

The exact formula is not known outside of Amazon. In fact, even the reviewer reputation scores, which leads to a ranking of Amazon reviewers, follows some formula that apparently has been changed three times in the past years and remains a secret. Obviously, how high a reviewer is ranked depends on how many yes/useful votes (say, x) and no/not-useful votes (say, y) are received by each review she writes. If x is larger than a threshold, either x itself or the fraction $x/(x+y)$ can be used to assess this review's quality. And the reviewer reputation changes as some moving window average of these review quality measures change over time. The effect of fan vote or loyalty vote, where some people always vote yes on a particular reviewer's reviews, is then somehow subtracted. We will see in Chapter 6 that Wikipedia committee elections also follow some formula that turns binary votes into a ranking.

Let us consider the list of the top 20 LCD HDTVs of size 30 to 34 inches in April 2012, "top" according to average customer reviews. It can be obtained from Amazon by the following sequence of filters: Electronics > Television & Video > Televisions > LCD > 30 to 34 inches.

There are actually three rank ordered lists:

- The first is the ranking by Amazon, which orders the list in Table 5.5.3.
- Then there are the numerical scores of "average customer review", which, interestingly enough, does not lead to the actual ranking provided by Amazon.
- There are also the averaged rating scores, which lead to yet another rank order.

First, we look at the difference between the Amazon ordering and how the HDTVs would have been ranked had they been sorted based only on the average customer ratings. The two lists are as follows:

- 1, 2, 7, 12, 14, 3, 4, 5, 8, 9, 15, 16, 20, 6, 10, 11, 18, 13, 17, 19.
- 1, 7, 12, 14, 2, 5, 4, 8, 3, 9, 15, 16, 20, 18, 6, 10, 11, 13, 17, 19.

Clearly, the average customer review ranking is closer to the actual ranking. It follows the general trend of the actual ranking, with a few outliers: 7, 12, 14, 20, and 13 all seem to be strangely out of order. Let us try to reverse-engineer what other factors might have contributed to the actual ranking:

1. *Bayesian adjustment*: The population size of the ratings matter. The raw rating scores must be weighted with the population size in some way.
2. *Recency of the reviews*: Perhaps some of the reviewers rated their HDTVs as soon as they purchased them, and gave them high ratings because the products worked initially. But, especially with electronics, sometimes faulty

HDTV	Total reviews	5 star	4 star	3 star	2 star	1 star	Avg. review	Avg. rating
1	47	37	8	1	1	0	4.7	4.723
2	117	89	19	0	3	6	4.6	4.556
3	315	215	61	19	9	11	4.5	4.460
4	180	116	47	9	2	6	4.5	4.472
5	53	36	12	3	1	1	4.5	4.528
6	111	71	19	6	6	9	4.2	4.234
7	22	16	4	2	0	0	4.6	4.636
8	56	43	5	3	1	4	4.5	4.464
9	130	89	22	8	4	7	4.4	4.400
10	155	96	26	11	9	13	4.2	4.181
11	231	135	48	17	15	16	4.2	4.173
12	8	5	3	0	0	0	4.6	4.625
13	116	55	35	9	5	12	4.0	4.000
14	249	175	60	3	3	8	4.6	4.570
15	8	5	1	2	0	0	4.4	4.375
16	34	20	8	4	0	2	4.3	4.294
17	47	20	14	6	5	2	4.0	3.957
18	44	20	20	1	1	2	4.2	4.250
19	56	24	17	4	5	6	3.9	3.857
20	7	3	3	1	0	0	4.3	4.286

Table 5.4 List of the top twenty 30 to 34 inch LCD HDTVs on Amazon when sorted by average customer review.

components cause equipment to stop working over time. As a result, recent reviews should be considered more credible.

3. *Quality of the reviewers or reviews*: (a) Reputation score of the reviewer: Reviewers with higher reputations should be given more "say" in the average customer review of a product. (b) Quality of review: The quality of a review can be measured in terms of its length or associated keywords in the text. (c) Review of review: Higher review scores indicate that customers found the review "helpful" and accurate. (d) Timing of reviews: Review spamming from competing products can be detected based upon review timing.

Bayesian adjustment will be performed using equation (5.6). Here, R is the averaged rating of all the products (here assumed to be the top 20), and N is either the total number of reviews or, as in the Beer Advocate's website, is the *minimum* number of reviews necessary for a product to be listed, or possibly some mid-range number. Some number in-between these extremes is most likely, as lots of reviews are entered on Amazon, and the Bayesian adjustment will saturate as N keeps increasing and become simply R. From the above table, we can compute $R = \sum_i n_i r_i / \sum_i n_i = 4.36$. Now, what to choose for N? We compare the Bayesian adjustment rankings for $N_{min} = 7$, which is the lowest number of reviews for any product (20), $N_{max} = 249$, which is the highest number of reviews for any product (14), $N_{avg} = 100$, which is the average of the reviews across the products, and $N_{sum} = 1986$, the total number of reviews entered for the top 20 products. The results are as follows:

- N_{min}: 1, 7, 14, 2, 5, 12, 4, 3, 8, 9, 15, 20, 16, 18, 6, 10, 11, 13, 17, 19
- N_{max}: 1, 14, 2, 3, 4, 5, 7, 8, 9, 12, 15, 20, 16, 18, 6, 17, 10, 11, 19, 13
- N_{avg}: 14, 1, 2, 3, 4, 5, 7, 8, 9, 12, 15, 20, 16, 18, 6, 10, 11, 17, 19, 13
- N_{sum}: 14, 18, 2, 1, 3, 4, 7, 8, 9, 20, 6, 13, 10, 17, 5, 11, 12, 19, 15, 16

Clearly, having N too large or too small is undesirable: Both result in rankings that are far out of order. Both N_{max} and N_{avg} give better results, at least in terms of grouping clusters together. Still, there are some outliers in each case: 14, 15, 20, 6, 10, and 11 are considerably out of order.

We notice that products 12, 15, and 20 all have a very small number of ratings, specifically 8, 8, and 7, respectively. But even so, why would they be placed so far apart in the top 20? To answer this, we take into account the review of reviews: In the case of product 12, for instance, the "most helpful" review had 26 people find it helpful, whereas in the cases of products 15 and 20 it was 6 and 3, respectively. In addition, we can look at the recency of the reviews: The "most helpful" review for product 12 was made on November of 2011. Product 15's "saving grace" is that its corresponding review was more recent, made in December of 2011, which would push it closer to product 12 in the rankings. On the other hand, Amazon may have deemed that product 20's review in July of 2011 was too outdated. Finally, product 12 had an extremely high quality of review in its descriptive listing the pros and cons in each case. Amazon probably trusts this integrity.

On the other hand, why would Amazon decide to rank an item such as 6 so high, given that the Bayesian adjustment places it around rank 15? Well, when we look at item 6, we see that its "most helpful" review had 139 out of 144 people find it helpful, and similar percentages exist for all reviews below it. Further, the reviewers all have high ratings, one of which is an Amazon "top reviewer".

The final point of discussion is why product 14 is ranked so low in the top 20, but has one of the first three positions on each of the Bayesian adjustments. This can be explained by a few factors:

- The most helpful review was from 2010, extremely outdated.
- The 8 reviewers who gave it 1 star all said that the TV had stopped working after a month, many of whom were high up in the "helpful" rankings. These reviewers dramatically increased the spread of the ratings and opinions for this product.

To summarize, the following set of guidelines is inferred from this (small) sample on how Amazon comes up with its rankings:

1. An initial Bayesian ranking is made, with N chosen to be somewhere around N_{max} or N_{avg}.

2. Products that have small numbers of reviews or low recency of their most helpful reviews are ranked separately amongst themselves, and re-distributed in the top 20 at lower locations (*e.g.*, products 12, 15, and 20).
3. Products that have very high quality, positive reviews from top reviewers are bumped up in the rankings (*e.g.*, product 6).
4. Products that could cause a potential risk to their sales due to the possibility of faulty electronics (*e.g.*, product 14) are severely demoted in the rankings.

5.6 Advanced Material

As we just saw, review of review can be quite helpful: higher scores means more confidence in the review, and naturally leads to a heavier weight for that review. If a review does not have any reviews, we may take some number between the lowest score for a review and the average score for all reviews. More generally, we may take the Bayesian likelihood approach to determine the trustworthiness of a review based on the observation of the reviews it receives.

While a systematic study of the above approach for rating analytics is still ongoing, this general idea of weighting individual estimates based on each estimate's effectiveness has been studied in a slightly different context of statistical learning theory, called **boosting**. If we view each estimator as a person, boosting shows how they can collaborate, through sequential decision making, to make the resulting estimate much better than any individual one can be.

5.6.1 Adaptive boosting

Ada Boost, short for adaptive boosting, captures the idea that by *sequentially* training estimators, we can make the average estimator more accurate. It is like an experience many students have while reviewing class material before an exam. We tend to review those points that we already know well (since that makes us feel better), while the right approach is exactly the opposite: to focus on those points that we do not know very well yet.

As in Figure 5.4, consider N estimators $y_i(\mathbf{x})$ that each map an input $\mathbf{x}$ into an estimate, *e.g.*, the number of stars in an aggregate rating. The final aggregate rating is a weighted sum: $\sum_i \alpha_i y_i$, where $\{\alpha_i\}$ are scalar weights and $\{y_i\}$ are functions that map vector $\mathbf{x}$ to a scalar. The question is how to select the right weights α_i. Of course, those y_i that are more accurate deserve a larger weight α_i. But how much larger?

Let us divide the training data into M sets indexed by j: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$. For each training set, there is a right answer $t_j,\ j = 1, 2, \ldots, M$, known to us since we are training the estimators. So now we have N estimators and M data sets. As each estimator y_i gets trained by the data sets, some data sets are well handled while others are less so. We should *adapt* accordingly, and give challenging data

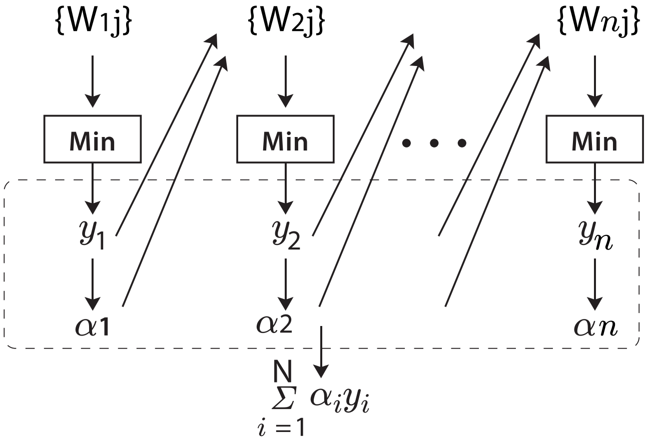


Figure 5.4 The schematic of Ada Boosting. There are N estimators, indexed by i: $\{y_i\}$, and M training data sets, indexed by j. Training is done to minimize weighted errors, where the weights w_{ij} are sequentially chosen from i to $i + 1$ according to how well each training data set is learned so far. The final estimator is a weighted sum of individual estimators, where the weights $\{\alpha_i\}$ are also set according to the error of each estimator y_i.

sets, those leading to poor performance thus far, more weight w in the next estimator's parameter training.

So both the training weights $\{w_{ij}\}$ and the estimator combining weights $\{\alpha_i\}$ are determined by the performance of the estimators on the training sets.

We start by initializing the training weights w_{1j} for estimator 1 to be even across the data sets:

$$w_{1j} = \frac{1}{M}, \quad j = 1, 2, \ldots, M.$$

Then *sequentially* for each i, we train estimator y_i by minimizing:

$$\sum_{j=1}^{M} w_{ij} 1_{y_i(\mathbf{x}_j) \neq t_j},$$

where 1 is an indicator function, which returns 1 if the subscript is true (the estimator is wrong) and 0 otherwise (the estimator is correct).

After this minimization, we get the resulting estimator leading to an error indicator function abbreviated as 1_{ij}: the optimized error indicator of estimator i being wrong on data set j.

The (normalized and weighted) sum of these error terms becomes:

$$\epsilon_i = \frac{\sum_j w_{ij} 1_{ij}}{\sum_j w_{ij}}.$$

Let the estimator combining weight for estimator i be the (natural) log scaled