Recommending Movies to Watch

Q7.1 How did Netflix divide up their dataset for the competition?

The breakdown of the Netflix Prize data is shown in Illustration 22. Two portions of it were released to the public, and two were private:

- *Training set*: Roughly 100 million ratings were made public, which constituted the training set. This is what competitors would use as the input data to tune their algorithms.
- *Probe set*: About 1.4 million additional ratings were made public, and had similar properties to the test and quiz sets described next. Competitors could evaluate the performance of their tuned algorithms on this probe set whenever they wanted.
- *Quiz set*: Another 1.4 million ratings were hidden from the public, and constituted the quiz set. Competitors were allowed to run their algorithms on this and see the RMSE at most once a day. The Netflix Prize's website kept track of these scores on its leaderboard.

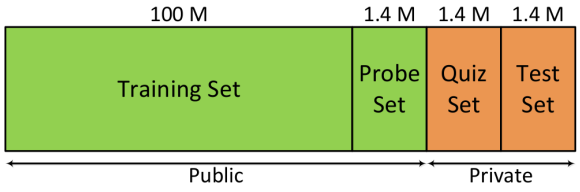

Illustration 22: The Netfix Prize's four data sets. The training set (100 million data points) and probe set (1.4 million) were publicly released, whereas the quiz set and test set were hidden from the public and known only to Netfix.

- *Test set*: Another 1.4 million ratings were also hidden from the competitors, forming the test set. This was the real test of the competition. The RMSE scores on this set would determine the winner.

Why partition the data into multiple sets like this? Netflix did this so that they could be confident that any algorithm which was able to beat CineMatch by 10% would be able to do so in cases besides

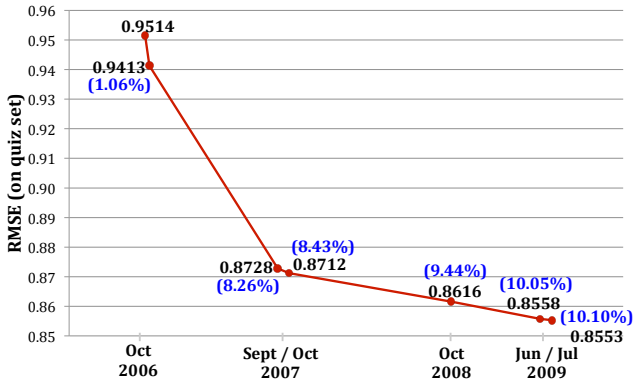

Illustration 23: Key events of the Netflix Prize competition, which lasted for almost three years and came down to a difference of 20 minutes between two leading submissions. Here, we show progress in reducing the RMSE on the *quiz set*. Also included is the percent improvement over that of CineMatch.

this particular dataset. Had they released the test set to the public, many teams would have tried to **reverse-engineer** that specific set of data, adding as many parameters to their algorithm as necessary to get zero error on the test set. But then the algorithm would not work on other data sets. This would be analogous to your teacher giving you the questions that were going to appear on an exam before you take the exam: if your goal is to get a good score on the test, then in your preparation, you could just memorize the answers to these questions, rather than having to build a more general understanding of the material.

So, if Netflix was going to replace CineMatch with the winning predictor, this algorithm needed to show promise of outperforming CineMatch in the most general case possible. In a similar manner, they only allowed the teams to test on the quiz set once a day. The probe set, on the other hand, was released to the public to help them evaluate where they stood on a similar dataset.

Q7.2 How did the Netflix Prize competition play out?

Illustration 23 highlights some of the important milestones in RMSE improvement on the quiz set over this time:

- *Beating CineMatch*: The competition began on October 2, 2006. Within a week of the start of the competition, CineMatch was beaten, with one team obtaining a 1% improvement in RMSE.
- *Progress Prizes*: The competition offered a Progress Prize of $50,000 to the leading team each year, as long as the team had achieved an improvement of at least 1% from the previous year. In September and October of 2007, first place changed hands a few times. A team of researchers called BellKor ended up winning, having reached an RMSE of 0.8712 (*i.e.*, an improvement of 8.43%). In 2008, teams started merging: BellKor and another leading team, BigChaos, combined efforts as BellKor in BigChaos and received the 2008 progress prize, pushing improvement to 9.44%.
- *Reaching 10%*: In June 2009, BellKor's Pragmatic Chaos (merging of BellKor in BigChaos and Pragmatic Theory) became the first team to achieve more than 10% improvement, beating CineMatch by 10.06% on the quiz set. At this point, the competition entered the last call phase for the grand prize. At the end of this period, two teams had beaten CineMatch by more than 10% on the quiz set: BellKor's Pragmatic Chaos had 10.09%, and The Ensemble had 10.10%, slightly better.

So, would The Ensemble win the prize? Not quite. Remember, the winner was determined by comparing the RMSE improvement on the *test set*, not the quiz set. At the end, Netflix found that both teams got the same improvement of 10.06% on the test set. But BellKor's Pragmatic Chaos had submitted their algorithm 20 minutes earlier, and so they were declared the winner of the $1 million grand prize!

Q7.3 How are the cosine similarities in the book determined?

In Illustration 24, where we plot four different cases of the angles between two line segments. If the angle is *very small*, as in (a), then the two line segments are pointing in similar directions, and they exemplify a strong positive correlation. On the other hand, if the angle is *very large*, as in (d), then the two line segments are pointing in opposite directions, and they have a strong negative correlation. Intermediate values of the angle, as in (b) and (c), indicate that the segments are not correlated.

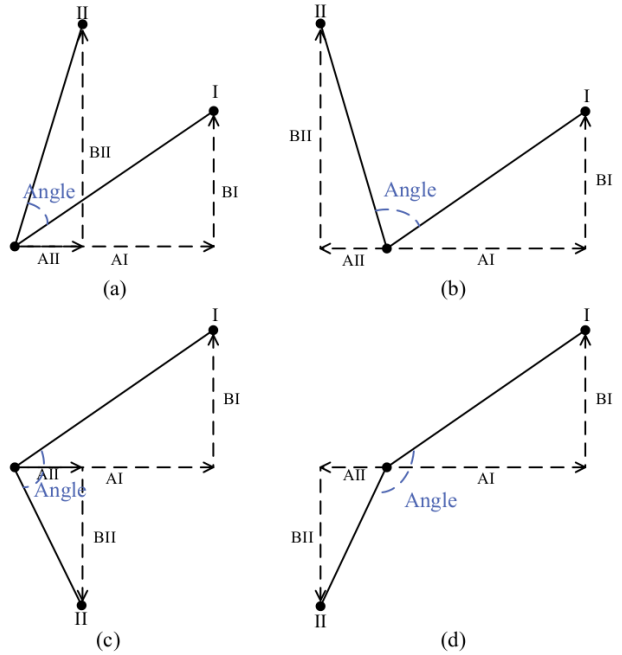

Illustration 24: Four possibilities of the angle between two line-segments. For our purposes, if we abstract the segments to two Movies I and II, then AI becomes an indication of the rating User A gave to Movie I, BI of B to I, AII of A to II, and BII of B to II. The smaller the angle between the two segments, the more *similar* the movies are, and the larger, the more *dissimilar*. In this way, we are interested in cases where the angle between the movies is close to 0° (as in (a)) or 180° (as in (d)).

Now, let each line segment represent one of two Movies, I and II. AI is an indication of the rating that a User A gave Movie I, BI of a rating that a User B gave Movie I, AII of A to II, and BII of B to II. We can directly apply the correlation logic of the previous paragraph here. In (a), User A rated both I and II positively, as did User B: this shows a strong similarity, or positive correlation, between the movies. And in (d), User A rated Movie I positively and II negatively, and User B also rated them opposite to one another: this shows a strong dissimilarity, or negative correlation, among the movies.

We can quantify the similarity / dissimilarity between movies by taking

	I	II	III	IV	V
A	0.37	?	-1.00	?	-0.43
B	-0.30	0.17	0.25	-	-0.10
C	-	0.67	?	?	0.40
D	-0.30	-	0.25	-0.50	-0.10
E	-0.63	?	-	0.17	0.57
F	-0.05	-0.58	0.50	0.75	-

Illustration 25: Table of baseline-prediction errors.

the *cosine* of the angle between the segments. The cosine is given as

$$\frac{\mathrm{AI} \times \mathrm{AII} + \mathrm{BI} \times \mathrm{BII} + \cdots}{\sqrt{\mathrm{AI}^2 + \mathrm{BI}^2 + \cdots} \times \sqrt{\mathrm{AII}^2 + \mathrm{BII}^2} + \cdots}$$

In the numerator, each term is the product of how a user rated each movie. If he rated them both "good" or both "bad," the term will be positive, otherwise it will be negative. Then, these products are summed. For strong positive correlation, we need at least mostly positive terms, and for strong negative correlation, we need at least mostly negative terms. The denominator simply normalizes the sum by the length of the movie "segments," thereby restricting this to lie between -1 and $+1$. This equation is easily extended to more than two users: just add more terms in the numerator and denominator, as indicated.

You may be wondering how some of the values of AI, BI, AII, BII, ... can be negative if Netflix requires a rating scale of 1 to 5. Well, we don't compute similarity based on the *absolute* ratings, but rather based on the *error* between the baseline predictor from the book and the actual values. For each user / movie pair, we subtract the baseline from the raw ratings. This is given in Illustration 25.

Take, for instance, Movies I and II. For I, we have entries in the training data for Users A, B, D, E, and F, and for II, we have it for B, C, and F. We only make use of those that rated *both* Movies: B and F. Applying

the cosine similarity equation, we have

$$\begin{aligned}
\frac{\mathrm{BI} \times \mathrm{BII} + \mathrm{FI} \times \mathrm{FII}}{\sqrt{\mathrm{BI}^2 + \mathrm{FI}^2} \times \sqrt{\mathrm{BII}^2 + \mathrm{FII}^2}} &= \frac{-0.30 \times 0.17 + -0.05 \times -0.58}{\sqrt{0.30^2 + 0.05^2} \times \sqrt{0.17^2 + 0.58^2}} \\
&= \frac{-0.0220}{0.3041 \times 0.6044} \\
&= -0.11.
\end{aligned}$$

Let's take another example: Movies III and V. This time, we have three Users – A, B, and D – that rated both movies:

$$\frac{\mathrm{AIII} \times \mathrm{AV} + \mathrm{BIII} \times \mathrm{BV} + \mathrm{DIII} \times \mathrm{DV}}{\sqrt{\mathrm{AIII}^2 + \mathrm{BIII}^2 + \mathrm{DIII}^2} \times \sqrt{\mathrm{AV}^2 + \mathrm{BV}^2 + \mathrm{DV}^2}}$$

$$\begin{aligned}
&= \frac{-1.00 \times -0.43 + 0.25 \times -0.10 + 0.25 \times -0.10}{\sqrt{1.00^2 + 0.25^2 + 0.25^2} \times \sqrt{0.43^2 + 0.10^2 + 0.10^2}} \\
&= 0.79.
\end{aligned}$$