Log-likelihood ratios in recommendation systems

Data Science, Adobe Target.

This document reviews the the use of log-likelihood ratios as a similarity metric in item-item recommendation algorithms. We discuss its origin in likelihood ratio tests in statistical literature, and provide a precise definition when employed in recommendation systems.

Likelihood ratio tests are a class of statistical test used to compare the goodness of fit for two statistical models. The test statistic is the *likelihood ratio*, which is essentially the a ratio of the likelihoods of the data under each model. The hypothesis test which. motivates such a test statistic typically proceeds as follows.

Consider two hypotheses for the distributions of data, each parametrized by a different distributional parameter θ:

$$H_0 : \quad \theta \in \Theta_0, \tag{1}$$

$$H_1 : \quad \theta \in \Theta_1 \tag{2}$$

where Θ_1 is typically disjoint from Θ_0 (e.g. if Θ is the full space of parameters θ, $\Theta_1 \equiv \Theta \backslash \Theta_0$).

If we assume the data is IID, consisting of samples $\boldsymbol{x} = x_1, x_2, \ldots x_n$ and comes from some unknown probability density function, f, the likelihood of the data is

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = f(x_1, x_2, \ldots x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta). \tag{3}$$

The likelihood ratio is then the quantity

$$\lambda = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta; \boldsymbol{x})}{\max_{\theta \in \Theta_1} \mathcal{L}(\theta; \boldsymbol{x})} \tag{4}$$

Often it is much simpler to compute maxima of the log-likelihood functions, $\ell(\theta; \boldsymbol{x}) = \log \mathcal{L}(\theta; \boldsymbol{x})$, and so we compute that *log-likelihood ratio* $-2 \log \lambda$. We compute this statistic, and then reject the null hypothesis if the likelihood ratio is less than some critical value c (or equivalently, the log-likelihood ratio is greater than some critical value $-2 \log c$.

This may seem awfully formal, but to take a concrete example which we will develop further when considering recommender systems, we can imagine that we have two datasets comprising a series of Bernoulli (binary) trials. We wish to ask whether the two datasets come from the same, or different distributions. In such a scenario, the number of successes in each dataset will be distributed according to a binomial distribution, and our null hypothesis is that both datasets are described by a binomial distribution with the same probability of success p_0. The alternative hypothesis is that each dataset is generated by separate Bernoulli processes with different probabilities of success p_1 and p_2. To complete the example, note that Θ is the set of all possible values for the probability of success for both distributions, Θ_0 will be subset of this where both distributions have an identical probability of success p_0, while Θ_1 is the complement of that set (*i.e.* both distribution have different sets of parameters).

Application to recommendation systems

In recommender systems we often need to recommend items based on some notion of similarity, e.g. *people who viewed/bought item A, also viewed/bought item B.* For concreteness, let us focus on the question of "viewed A, bought B", but this discussion applies equally to any other combination of viewed/purchased.

A simple (and naive) metric for calculating item similarities is to use their *raw* cooccurrence scores. In this case of "people who viewed A, bought B," this cooccurrence is essentially the unnormalized probability

$$\begin{aligned} p(\text{bought } B|\text{viewed } A) &= \frac{\textit{number of people who bought B \textbf{and} viewed A}}{\textit{number of people who viewed A}} \\ &= \frac{\textit{cooccurrence}}{\textit{number of people who viewed A}}. \end{aligned} \tag{5}$$

However cooccurrence based similarity suffers from the problem of recommending items that are globally popular. Essentially, globally popular items (*i.e.* those that are viewed or purchased by lots of people), will naturally have the highest cooccurrences, and so will end up being recommended with lots of items, despite them not being very specific to each item. We instead must find items to recommend that are anomalously relevant.

Thus, a better way to phrase the question of similarities is: "*are viewing item A and* *purchasing item B independent processes?*", *i.e.* ***if***

$$p(\text{bought } B|\text{viewed } A) \approx p(\text{bought } B|\text{not viewed } A) \approx p(\text{bought } B),$$

then items A and B are not related, and should not be recommended together. Conversely, if these conditional probabilities are very different, this indicates an anomalously relevant pair of items.

To test whether or not these conditional probabilities are related , we begin with the contingency table

	B	$\sim B$
A	k_1	$n_1 - k_1$
$\sim A$	k_2	$n_2 - k_2$

Here, $\sim B$ means "did not purchase B" etc. So here, we have n_1 people who viewed A, of whom k_1 also purchased B, while n_2 people did not view A, of whom k_2 purchased B.

The question before the house is whether $p(B|A) \overset{?}{=} p(B| \sim A)$, *i.e.* whether the two rows of this contingency table are generated by the same probability distribution. Assuming the process of purchasing items is a Bernoulli process (for each user), the number of success (purchases of B) comes from a binomial distribution $Bin(n, p)$. Dividing our dataset into those who viewed A (X_{BA}) and did not view A ($X_{B\sim A}$), The null and alternative hypotheses are

$$H_0 : p(B|A) = p(B| \sim A), \text{ i.e. } X_{BA} \sim Bin(n_A, p_0) \text{ and } X_{B\sim A} \sim Bin(n_{\sim A}, p_0) \tag{6}$$

$$H_1 : p(B|A) \neq p(B| \sim A), \text{ i.e. } X_{BA} \sim Bin(n_A, p_1) \text{ and } X_{B\sim A} \sim Bin(n_{\sim A}, p_2) \tag{7}$$

The likelihood ratio is then

$$\lambda = \frac{\max_p \left[\binom{n_1}{k_1} p_0^{k_1}(1-p_0)^{n_1-k_1} \cdot \binom{n_2}{k_2} p_0^{k_2}(1-p_0)^{n_2-k_2}\right]}{\max_{p_1,p_2} \left[\binom{n_1}{k_1} p_1^{k_1}(1-p_1)^{n_1-k_1} \cdot \binom{n_2}{k_2} p_2^{k_2}(1-p_2)^{n_2-k_2}\right]} \tag{8}$$

Some simple manipulations yield maximum likelihood estimates for p_0, p_1 and p_2 of

$$p_0^* = \frac{k_1 + k_2}{n_1 + n_2}, \quad p_1^* = \frac{k_1}{n_1}, \quad p_2^* = \frac{k_2}{n_2} \tag{9}$$

And so we find the *log-likelihood ratio* to be

$$\boxed{-2\log\lambda = 2\left[\log L(p_1^*, k_1, n_1) + \log L(p_2^*, k_2, n_2) - \log L(p_0^*, k_1, n_1) - \log L(p_0^*, k_2, n_2)\right]} \tag{10}$$

where

$$\log L(p, k, n) = k \log p + (n - k) \log(1 - p) \tag{11}$$

The crucial leap is to then use this test statistic $(-2 \log \lambda)$ as a similarity metric for determining which items should be recommended together, *i.e.* we are not interested in accepting or rejecting the null hypothesis, but we simply use the test statistic to rank item similarities. Here, a large $-2 \log \lambda$ indicates a λ much smaller than 1, and so distributions that are more likely to have different means, *i.e.* better item similarity.