Chapter 7

Statistical Models

Recommender systems is inherently statistical. Indeed, the very fact that we discuss the bias-variance tradeoff recognizes the fact that our data are subject to sampling variation, a core statistical notion. In this chapter, we will apply classical statistical estimation methods to a certain *latent variables* model.

7.1 The Basic Model

Again, for concreteness, we'll speak in terms of user ratings of movies. Let (U, I) denote a random (user ID, movie ID) pair. Let u and m denote the numbers of users and movies. Denote the user's rating by Y_{IJ}. The model is additive, postulating that

$$Y_{IJ} = \mu + \alpha_I + \beta_J + \epsilon_{IJ} \tag{7.1}$$

Here μ is an unknown constant, the overall population mean over all users and all movies. The numbers $\alpha_1, \alpha_2, ..., \alpha_u$ and $\beta_1, \beta_2, ..., \beta_m$ are also unknown constants; think of α_i to be the tendency of user i to give harsher ($\alpha_i < 0$) or more generous ($\alpha_i > 0$) ratings, relative to the general population of users, with a similar situation for the β_j and movies. The ϵ term is thought of as the combination of all other affects.

Note that what makes, e.g., α_I random above is that I is random, and similarly for the β_J and ϵ_{IJ}. The α, β and ϵ terms are assumed to be statistically independent, each with mean 0.

So, we model a user's rating of a movie as the sum of latent additive user and movie terms, plus a catch-all "everything else" term.[1] The question then becomes how to estimate μ, and $\alpha_1, \alpha_2, ..., \alpha_u$

[1]What does the word *latent* here mean? Why is μ not "latent"? The answer is that it is a tangible quantity;

and $\beta_1, \beta_2, ..., \beta_m$, where u and m are the numbers of users and movies in our data. We will present two methods.

7.2 Two General Statistical Methods for Parameter Estimation

We'll be using two famous estimation tools from statistics, the Method of Moments and Maximum Likelihood Estimation. We'll introduce those in this section.

7.2.1 Example: Guessing the Number of Coin Tosses

To aovid distracting complexity, consider the following game. I toss a coin until I accumulate a total of r heads. I don't tell you the value of r that I used, only informing you of K, the number of tosses I needed.

It can be shown that

$$P(K = u) = \binom{u-1}{r-1} 0.5^u, \ u = r, r+1, ... \tag{7.2}$$

Say I play the game 3 times, and I tell you $K = 7, 10$ and 9. What could you do to try to guess r?

Notation: We play the game n times, always with the same r, yielding $K_1, K_2, ..., K_n$.

7.2.2 The Method of Moments

The *moments* of a random variable X are the expected values of the powers. E.g. $E(X^3)$ is called the third moment of X.

If we are trying to estimate s parameters, $\theta_1, ..., \theta_s$, we need s moments. We find population expressions for the θ_i in terms of the first s moments of the random variable at hand, setting up s equations that match those expressions to the estimated parameters, $\widehat{\theta}_1, ..., \widehat{\theta}_s$, then solve for the latter, then solve for the latter

Here we have just one parameter, r. It can be shown that in the game example,

$$E(K) = \frac{r}{0.5} = 2r \tag{7.3}$$

we all can imagine finding the overall mean for all users and movies, given enough data. By contrast, the α values' existence depend on the validity of the model. It's similar to the NMF situation, where the postulate postulates existence of a set of "typical" users.

MM involves replacing both sides of an equation like (7.3) by sample estimates, in this case

$$\overline{K} = 2\widehat{r} \tag{7.4}$$

where

$$\overline{K} = \frac{K_1 + ... + K_n}{n} \tag{7.5}$$

and $\widehat{r}$ is our estimate of r.[2]

So the idea of MM is:

1. Find theoretical (i.e. population-level) equations for various expected values, enough to cover the number of parameters being estimated.

2. In those equations, replace expected values and parameters by sample estimates.

3. Solve for the sample estimates.

7.2.2.1 The Method of Maximum Likelihood

To guess r in the game, you might ask, "What value of r would make it most likely to need 7 tosses to get r heads?" You would then find the value of w that maximizes the *likelihood*, defined to be the probability of our observed data under a given value of the parameter(s), in this case

$$\Pi_{i=1}^{n} \binom{K_i}{w-1} 0.5^{K_i} \tag{7.6}$$

In this discrete case you could not use calculus, and simply would use trial-and-error to find the maximizing value of w, which will be our $\widehat{r}$.

7.2.2.2 Comparison: MM vs. MLE

If these two methods were nervous academics, MM would be quite envious of MLE:

- MLE is by far the more widely-used method.

[2] It is standard to use the "hat" symbol to mean "estimate of."

- MLE can be shown to be optimal in a certain sense. (Roughly, it has the smallest possible variance of all estimators, when n is large.)
- Various aspects of MLE and related topics are famous enough to be named after people, e.g. Fisher information (yes, the significance testing Fisher) and the Cramer-Rao lower bound.

On the other hand:

- Often MM makes fewer assumptions than MLE. That will be the case for us in the RS application below, a major point.
- MM is easier to explain. MLE has the same "What if...?" basis that p-values have, rather confusing.
- MM is actually the basis for the 2013 Nobel Prize in Economics! Lars Peter Hansen won the prize for his development of the Generalized Method of Moments estimation tool.

7.3 MM Applied to (7.1)

As you'll see, MM is arguably the more useful of the two methods in this particular setting.

7.3.1 Derivation of the Estimates

The expected values in Section 7.2.2 can be conditional. So, from (7.1), write

$$E(Y_{IJ} \mid I = k) = \mu + \alpha_k + E(\beta_J | I = k), \ k = 1, 2, ..., u \tag{7.7}$$

But since I and J are independent, we have

$$E(\beta_J | I = k) = E(\beta_J) = 0, \ k = 1, 2, ..., u \tag{7.8}$$

so

$$E(Y_{IJ} \mid I = k) = \mu + \alpha_k, \ k = 1, 2, ..., u \tag{7.9}$$

Now we must find our sample estimate of the left-hand side, and equate it to $\mu + \alpha_k$.

But the natural estimate of $E(Y_{IJ} \mid I = k)$ is simply the mean rating user k gave to all movies she rated.

Moreover, the natural estimate of μ is the average rating given to all movies in our data.

So we now have our $\widehat{\alpha}_k$. The derivation of the $\widehat{\beta}_l$ is similar.

7.3.2 Relation to Linear Model

For simplicity, consider the call

```
lm(rating ~ userID-1)
```

omitting the movies. Think of what will happen with the matrix A and the vector D in Section 3.4.5.

Recall that the -1 in the above call means we do not want an intercept term. In that case, **lm()** will produce u dummy variables rather than $u - 1$. This will help clarify the situation.

So, in the matrix A, column i will be the vector of 1s and 0s in the dummy for user i, $i = 1, ..., u$. Now consider the (i, i) element in $A'A$. It's the dot product of row i in A' and column i in A, thus the dot product of column i in A and column i in A. That will in turn be the sum of some 1s — actually, n_i 1s, where n_i is the number of ratings user i has made.

Meanwhile, the same reasoning says that for $i \neq j$, element (i, j) in $A'A$ is 0, since two dummy vectors coming from the same categorical variable will never have a 1 in the same position.

Putting all that together, we have that

$$(A'A)^{-1} = \mathrm{diag}(\frac{1}{n_1}, ..., (\frac{1}{n_u}) \tag{7.10}$$

a diagonal matrix with the indicated elements.

What about $A'D$ in (3.4.5)? Similar reasoning shows that its m^{th} element is the sum of all the ratings given by user m.

Putting this all together, we find that the m^{th} estimated coefficient returned by **lm()** will be the average rating given by user m — exactly the same as MM gave us!

7.4 MLE Applied to (7.1)

Here, the α, β and ϵ terms are assumed to have Gaussian distributions. Continue to assume they have mean 0, and denote their variances by σ_1^2, σ_2^2 and σ^2, respectively.

Since Gaussian distributions are continuous, the likelihood function L involves density functions rather than probabilities. L will then be a complicated expression involving various instances of the standard normal density,

$$\phi(w) = \frac{1}{\sqrt{2\pi}} e^{-0.5w^2} \tag{7.11}$$

μ, σ_1^2, σ_2^2, σ^2, the α andβ terms, and the user ratings. The expression is maximized with respect to μ, σ_1^2, σ_2^2 and σ^2, to obtain the corresponding estimates. Of course, an iterative procedure is used for the maximization.

What, then, do we get from this?

- We get predicted values for all user/movie combinations, as with MM.
- We get estimates of the σ_i^2, of interest as we can tell whether there is more variation in users or in movies – or maybe not much in either, so we can't expect to predict well.[3]
- On the other hand, the Gaussian assumption is extremely strong, especially in view of the fact that ratings (in this case) are integers from 1 to 5.

7.5 Conclusion

MM seems a better bet here. It makes fewer restrictive assumptions, it is fast computationally, and jibes with an **lm()** analysis.

[3]It is possible to get variance estimates using MM as well.