PRINCETON UNIVERSITY, CENTER FOR INFORMATION TECHNOLOGY POLICY

A Gentle Primer for Recommender Systems Math Stuff

Amy A. Winecoff *

December 3, 2020

0.1 Unnecessary Preamble

On my first day working as a data scientist, my boss handed me a document explaining the team's thinking behind the fashion recommender system they were developing. I was really excited about being a data scientist who worked on recommender systems. Building a recommender system is not like building models for when spark plugs will fail. There is not necessarily a "right" or "wrong" answer; whether or not people find a recommendation useful depends on lots of deeply psychological factors, and I am a trained psychologist. Who better to build a recommendation engine than someone like me? Filled with enthusiasm, I started reading my boss' document. Then I hit the first equation. I specifically remember running into this symbol: $\forall$

I know, right? I had no idea what it meant. For reference, it means "for all" as in "for all people who are terrified of math, $\forall$ is a scary, scary monster." I became convinced I was going to get fired. That a single symbol could strike fear into my heart says as much about me as it does about the symbol; however, I don't think I'm alone. Lots of mathy things that aren't even actual math can present a barrier to people engaging with the data science, machine learning, and recommender systems communities, and that is not great for the data science, machine learning, and recommender systems communities. In this document I want to go over a few concepts, symbols, and definitions that I think could knock down some of those preliminary hurdles. I am going to hold your hand the whole way, so trust me and keep reading.

1 ONE EXAMPLE TO RULE THEM ALL

I hate it when tutorials use letters and symbols in their examples. I get that folks are trying to be generalizable, but I love me a concrete example, so that's what I am going to use in this document. In the before times, Joe, Scott, and I enjoyed doing whiskey tastings (See Figure 1.1).

*I am sure there are some nerds out there who will find errors with my math or my notation or something. If so, shoot me an email at aw0934 [at] princeton [dot] edu.

Wherever possible I will try to avoid abstractions like X or Ω and will instead use examples relevant to whiskey tastings. Hopefully that will give you something to hang your hat (or Glencairn) on. I can't promise you that I won't use *any* symbols or letters, but I can promise to introduce you to them first. In these examples, I'm going to refer to myself, Amy, in the third person here because that's easier, and also because I am a big deal.

Figure 1.1: Scott and Joe Drink All the Whiskey

2 Sets of Things

A lot of times when you read papers and books about recommender systems, they talk about "sets."

- A **set** is an unordered collection of things where nothing is repeated

The set W is a set of whiskeys that Scott, Amy, and Joe have sampled. When we describe sets, we use curly braces because curly braces are really cute. Here, look!

$$W = \{Bodacious\,Bourbon, Stellar\,Scotch, Righteous\,Rye, Triumphant\,Tennessee\} \tag{2.1}$$

I know I tricked you into reading that first equation, but wasn't it sups cute? Anyway, another symbol you should probably get friendly with is this gal: $\in$. She is used to specify that something is a member of a set. For example, if we want to say that Bodacious Bourbon is a member of the set of whiskeys W, we could write:

$$Bodacious\,Bourbon \in W \tag{2.2}$$

You might run across the word "cardinality" when reading about sets. This is just a fancy word for size. The cardinality or size of a set is the total number of things that are in that set. Sometimes mathy people use a symbol that looks suspiciously like an absolute value symbol (absolutely wicked!) to specify a set's cardinality. For example:

$$|W| = 4 \tag{2.3}$$

Some other symbols you might meet in back alley equations include the symbols for the "union" and the "intersection" of two sets.

- The **union** of two sets contains the elements that are in one or the other or both.We use the symbol ∪ to mean union, which is nice because it looks like a "U."

To illustrate this, let's make another set. The set M are the ingredients used to make Joe's version of a Manhattan cocktail.

$$M = \{Righteous\,Rye, dry\,vermouth, bitters, lemon\,peel\} \tag{2.4}$$

This means that:

$$\begin{aligned} W \cup M = \{&Righteous\,Rye, dry\,vermouth, bitters, lemon\,peel,\\ &Bodacious\,Bourbon, Stellar\,Scotch, Triumphant\,Tennessee\}\end{aligned} \tag{2.5}$$

All the elements that are in either M or W are in $W \cup M$. Yay! You're totally nailing it (See Figure 2.1).

Figure 2.1: Gimme a ∪, ∪. Go unions, go unions, go!

- The **intersection** of two sets contains the elements that are in both sets. We use the symbol ∩ to mean intersection. [1]

For example,

$$W \cap M = \{Righteous\,Rye\} \tag{2.6}$$

because Righteous Rye is the only element that is in both sets.

[1] Annoyingly, the symbol for intersection does not look like an "I" because they already gave the "I" symbol to another math thing just to get your goat.

3 STUFF YOU MIGHT DO WITH SETS OF THINGS

Usually, if you're bothering to talk about sets, it's probably because you want to do something with their elements. For example, if all the things in your set are numbers, then you might want to add them. This is not to say that you might not want to add a bunch of whiskeys (See Figure 1.1), but TBH I don't really know how to write an equation for that, so we're going to need some numbers. Here are Amy's ratings, presented as your garden variety table:

Whiskey	**Rating**
Bodacious Bourbon	5
Stellar Scotch	4
Righteous Rye	4
Triumphant Tennessee	3

Table 3.1: Amy's Ratings of the Whiskeys.

But we could also show Amy's ratings using what recommender systems people call a "user-item matrix," which is basically a grid of ratings where the rows correspond to users and the columns correspond to items. Here is a user-item matrix containing only Amy's ratings, which we will call R_A:

	Bodacious Bourbon	Stellar Scotch	Righteous Rye	Triumphant Tennesse
Amy	5	4	4	3

Table 3.2: Amy's Ratings as the User-Item Matrix, R_A.

Maybe you want to sum up all Amy's whiskey ratings for funsies. To write that in a mathy way, meet your new pal $\sum$, which stands for a summation. When you use the summation symbol, you put the starting point of stuff you want to sum at the bottom and the stopping point of stuff you want to sum at the top. For example, if I wanted to sum the number 2 raised to the 1st, 2nd, 3rd, and 4th power, I could write:

$$\sum_{i=1}^{4} 2^i \tag{3.1}$$

This is the same thing as:

$$2^1 + 2^2 + 2^3 + 2^4 \tag{3.2}$$

Cool, cool. Now we're going to get wild and combine some symbols. We create an equation for summing up all the entries in Amy's user-item matrix. In plain language, we're saying, for each position (AKA index AKA i) in Amy's user-item matrix R_A, add all of the ratings. In equation form:

$$\sum_{i \in R_A} r_i \tag{3.3}$$

which leads us to sum:

$$5 + 4 + 4 + 3 = 16 \tag{3.4}$$

In this case, there is no stopping point specified at the top of the summation sign, because the specification at the bottom lets you know all the elements you will be summing. Let's get even fancier here, and combine a few more things. If you want to get the average, μ (pronounced "mew"), of all Amy's whiskey ratings, you could write:

$$\mu_A = \frac{\sum\limits_{i \in R_A} r_i}{|R_A|} \tag{3.5}$$

which corresponds to:

$$\mu_A = \frac{5+4+4+3}{4} = 4 \tag{3.6}$$

Remember, Amy was not the only one drinking whiskey. Scott and Joe tasted them too (See Figure 1.1). That is to say, we've got a set of whiskey "users" U that includes Amy, Scott, and Joe.

$$U = \{Amy, Scott, Joe\} \tag{3.7}$$

I know I already showed you Amy's ratings, but I really want you to like me, so I'm going to put all the ratings together to make it easy. Here are everyone's ratings, displayed as a user-item matrix:

	Bodacious Bourbon	Stellar Scotch	Righteous Rye	Triumphant Tennessee
Amy	5	4	4	3
Scott	5	5	3	2
Joe	2	4	5	1

Maybe, just maybe, you want to calculate the average whiskey rating for all users in the set U (i.e., for Amy, Scott, and Joe). We can write the following general equation:

$$\mu_u = \frac{\sum\limits_{i \in R_u} r_i}{|R_u|} \quad \forall u \in U \tag{3.8}$$

In plain language, that means for any user u in the set U, their average rating is the sum of all their ratings divided by their total number of ratings. Now let's actually do it for Amy (again), Scott, and Joe.

$$\mu_A = \frac{5+4+4+3}{4} = 4 \tag{3.9}$$

$$\mu_S = \frac{5+5+3+2}{4} = 3.75 \tag{3.10}$$

$$\mu_J = \frac{2+4+5+1}{4} = 3 \tag{3.11}$$

4 Necessary Interlude

You have just read a bunch of equations. Good for you! As a reward for your hard work so far, here is a picture of Joe snuggling my dog while sporting a truly excellent mustache (See Figure 4.1). This is an objectively fantastic picture. You're welcome.

Figure 4.1: Can You Even?

5 Similarity Between Things

I'm going to go off on a little bit of a high school math tangent (har, har!). Remember how we used to calculate the length of a hypotenuse of a triangle? Yep, me neither. Here is an equation to jog your memory:

$$a^2 + b^2 = c^2 \tag{5.1}$$

This is the Pythagorean theorem, and it can be used to determine how far apart two things are. You might ask, "How?" Or "When is this going to be over?" Not much longer, I promise.

Imagine we have two points in a space with two axes, x and y (See Figure 5.1). If we want to know how far apart these points are, we can make a right triangle out of them (See Figure 5.2). Then we can use the Pythagorean theorem to calculate the hypotenuse, which will give us the distance between the two points. In Figure 5.2, the line indicated by a is how far apart the two points are on the x-axis. The line indicated by b is how far apart the two points are on the y-axis. We use the values of a and b to calculate c (See Eq. 5.2).

$$\begin{aligned} (3-1)^2 + (50-30)^2 &= c^2 \\ 4 + 400 &= c^2 \\ \sqrt{404} &= c \\ 20.1 &= c \end{aligned} \tag{5.2}$$

Let's imagine that instead of being some abstract numbers, these values and points relate to our whiskey example. Let's say that we are plotting two whiskeys, Bodacious Bourbon (BB) and Triumphant Tennessee (TT), in terms of their price and quality (See Figure 5.3).

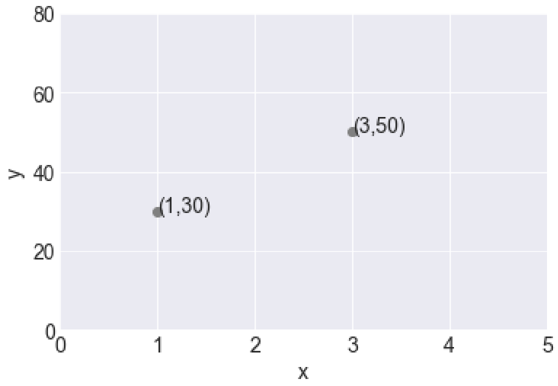

Figure 5.1: Points in Space, Fascinating!

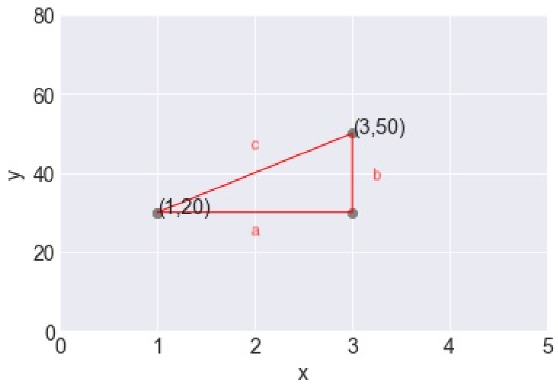

Figure 5.2: Triangle from Points, Fascinating!

We can use the distance between these points as a measure of their similarity. Whiskeys that are really close together in space can be considered very similar whereas whiskeys that are really far apart in space can considered very dissimilar. Note that when we're talking about distance, similarity is higher when the distance is smaller.

When we calculate distance this way, we often call it the Euclidean distance. You may encounter a formula for the Euclidean distance, D, written something like this:

$$D(a,b) = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2} \tag{5.3}$$

This means, for the features (i) of the items a and b, take their differences and square them. Add up all these squared feature differences, then take the square root. We go through all of this squaring and unsquaring because we don't really care about whether the difference is positive or negative, we just care about how large the difference is. Think about this, if you walked some number of feet from your house, that distance should be positive, regardless of whether you walked east, west, south, or north. Squaring the differences solves this problem since all the squared distances will be positive. We then take the square root of the sum to get it back in the original units. You might think, "Well, why not just take the absolute value then?" Sick dude, you just reinvented what is known as the Manhattan distance [2]. For our example, to calculate

[2] Not to be confused with the Manhattan cocktail

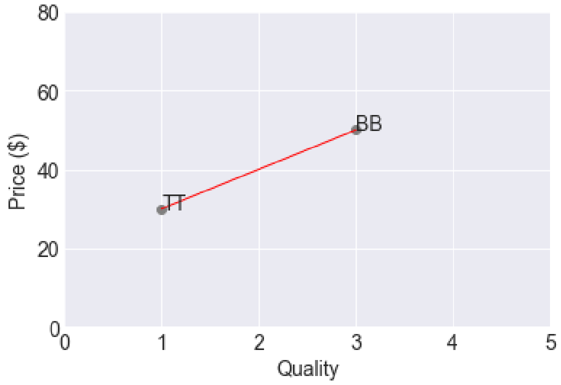

Figure 5.3: This is the Same Plot as Before with New Labels

the Euclidean distance, the values we would subtracting, squaring, summing, then unsquaring would be the two whiskeys' prices and qualities. I'll do it for you just to gain some more trust.

$$\begin{aligned} D(BB,TT) &= \sqrt{\sum_{i=1}^{N}(BB_i - TT_i)^2} \\ D(BB,TT) &= \sqrt{(3-1)^2 + (50-30)^2} \\ D(BB,TT) &= \sqrt{404} \\ D(BB,TT) &= 20.1 \end{aligned} \tag{5.4}$$

In recommender systems, we often base recommendations on how similar entities–like users or items–are to each other based on some distance/similarity metric. Euclidearn and Manhattan distance are two examples, but there are others such as the Pearson correlation (r), which you might have heard of before. Correlation is nice because it ranges from -1 to 1, with 0 meaning there is no correlation between two things, 1 meaning there is a perfect positive correlation between things, and -1 meaning there is a perfect negative correlation between things. I'll come back to this in a minute.

6 Making Predictions Using the Similarity Between Things

Back to our whiskey story, imagine that Joe and Scott went whiskey tasting without Amy. This is lame. Still, they both tried a new whiskey, called Most Excellent Whiskey (ME for short), that Amy has not tried. Scott tells Amy it is terrible (rating = 1), but Joe tells Amy it is fantastic (rating = 5). Who should Amy trust? Amy could base her assessment on how similar her preferences are to each of her compatriots' preferences. To do this, she's going to need you to bust out some distance metrics. We can calculate the correlation between Amy and Scott's ratings (i.e., the similarity between their ratings) with the following equation:

$$Sim(A,S) = \frac{\sum\limits_{i \in R_A \cap R_S} (r_{A,i} - \mu_A) \times (r_{S,i} - \mu_S)}{\sqrt{\sum\limits_{i \in R_A \cap R_S} (r_{A,i} - \mu_A)^2} \times \sqrt{\sum\limits_{i \in R_A \cap R_S} (r_{S,i} - \mu_S)^2}} \tag{6.1}$$

I know this one is kind of a lot, but you have already calculated several parts of this. You know that $\mu_A = 4$ and that $\mu_S = 3.75$. Now you just have to run through all the whiskeys that Amy and Scott have both rated (i.e., $R_A \cap R_S$), which includes all the whiskeys in set W.

$$Sim(A,S) = \frac{[(5-4)\times(5-3.75)]+[(4-4)\times(5-3.75)]+[(4-4)\times(3-3.75)]+[(3-4)\times(2-3.75)]}{\sqrt{(5-4)^2+(4-4)^2+(4-4)^2+(3-4)^2}\times\sqrt{(5-3.75)^2+(5-3.75)^2+(3-3.75)^2+(2-3.75)^2}} = 0.81 \tag{6.2}$$

If you do the same thing to calculate the correlation between Amy's ratings and Joe's ratings, you will get 0.22. Which is to say,

$$Sim(A,S) > Sim(A,J) \tag{6.3}$$

Or in human language, the similarity between Amy's ratings and Scott's ratings is higher, so she should probably trust him more. In fact, you can use the correlations you just calculated to get a prediction for how much Amy will probably like Most Excellent Whiskey. Note that when we have a predicted value, we put a little triangle hat over it like this: $\hat{R}$. Isn't that fun?! Here is a formula for calculating a prediction for what Amy would rate Most Excellent Whiskey:

$$\begin{aligned} r_{A,\hat{M}E} &= \frac{Sim(A,S)\times r_{S,ME} + Sim(A,J)\times r_{J,ME}}{Sim(A,S)+Sim(A,J)} \\ r_{A,\hat{M}E} &= \frac{(0.81\times 1)+(0.22\times 5)}{0.81+0.22} \\ r_{A,\hat{M}E} &= 1.85 \end{aligned} \tag{6.4}$$

You can predict that Amy would rate Most Excellent Whiskey a 1.85. Snap, you just made a recommender system. Seriously though, what you just did is more or less what you do to build what is known as a user-user collaborative filter. Of course, not all recommender systems are user-user collaborative filters. There are a lot of recommender system algorithms out there with varying levels of complexity. In the papers we will likely be reading in the group, there will be a lot of equations for algorithms, algorithm evaluations, and all matter of other ungodly things. When you encounter such equations, know that you have already worked through quite a few here, so you will be able to survive. Even if you don't fully get what is going on in an equation, understanding some part of it can still be really helpful. This kind of thing happens to me all the time.

Cheers,

Amy