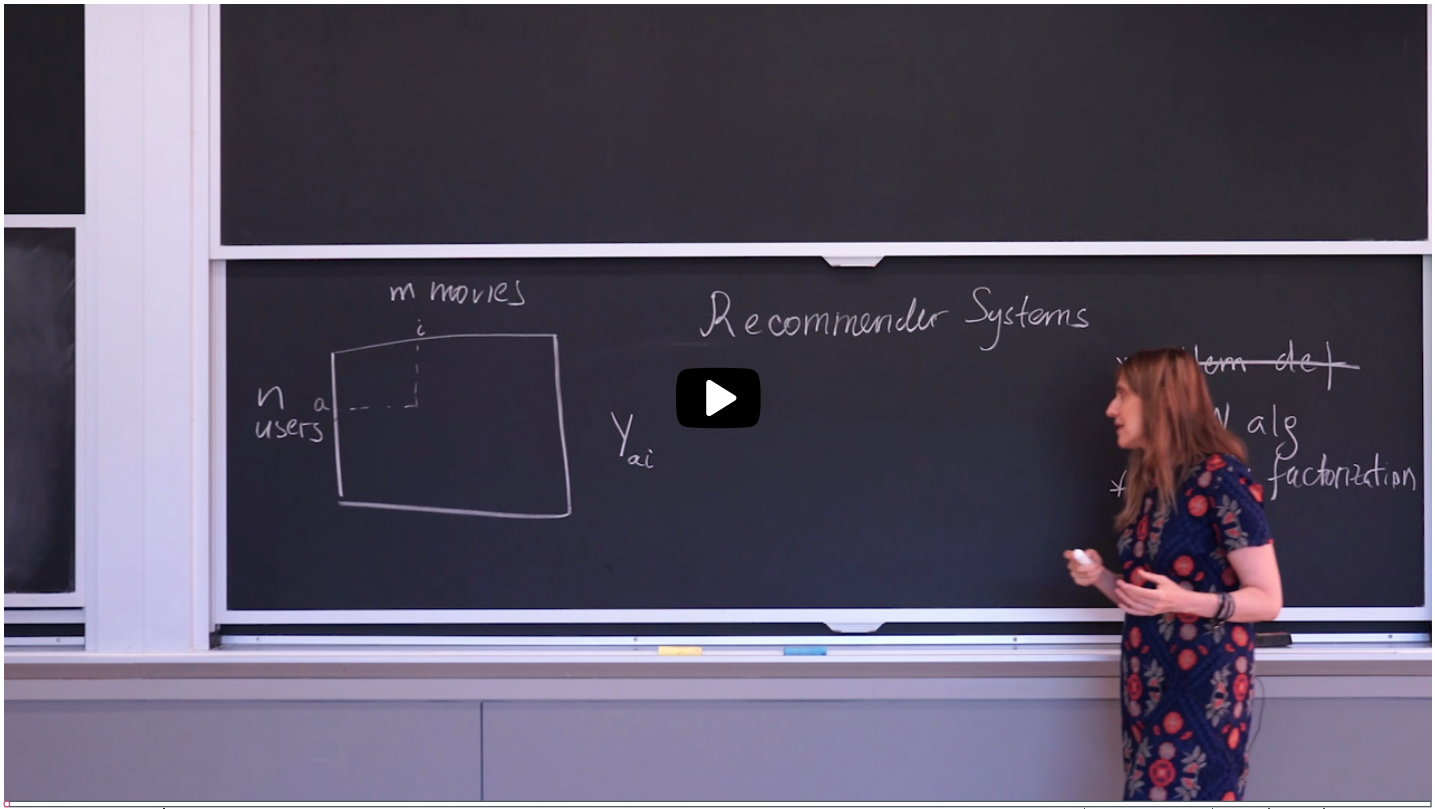
3. K-Nearest Neighbor Method

K-Nearest Neighbor Method



0:00 / 0:00 1.50x

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Our goal in the movie recommender system problem is to predict the movie ranking that a user would give on a movie that (s)he has not seen.

Let m be the number of movies and n the number of users. The ranking Y_{ai} of a movie $i \in \{1, \ldots, m\}$ by a user $a \in \{1, \ldots, n\}$ may already exist or not. Our goal is to predict Y_{ai} in the case when Y_{ai} does not exist.

K-Nearest Neighbour

The K-Nearest Neighbor method makes use of ratings by K other "similar" users when predicting Y_{ai}.

Let $\mathrm{KNN}(a)$ be the set of K users "similar to" user a, and let $\mathrm{sim}(a, b)$ be a **similarity measure** between users a and $b \in \mathrm{KNN}(a)$. The K-Nearest Neighbor method predicts a ranking Y_{ai} to be :

$$\widehat{Y}_{ai} = \frac{\sum_{b \in \mathrm{KNN}(a)} \mathrm{sim}(a, b)\, Y_{bi}}{\sum_{b \in \mathrm{KNN}(a)} \mathrm{sim}(a, b)}.$$

The similarity measure $\mathrm{sim}(a, b)$ could be any distance function between the feature vectors x_a and x_b of users a and b, e.g. the euclidean distance $\|x_a - x_b\|$ and the cosine similarity $\cos\theta = \frac{x_a \cdot x_b}{\|x_a\|\,\|x_b\|}$. We will use these similarity measures again in *Unit 4 Unsupervised Learning*.

A drawback of this method is that the success of the K-Nearest Neighbor method depends heavily on the choice of the similarity measure. In the next section, we will discuss collaborative filtering, which will free us from the need to define a good similarity measure.

Discussion

Topic: Unit 2 Nonlinear Classification, Linear regression, Collaborative Filtering (2 weeks):Lecture 7. Recommender Systems / 3. K-Nearest Neighbor Method

Hide Discussion

Add a Post

All Posts

Similarity measure vs distance

discussion posted 8 months ago by **Evgeny8**

May be I'm wrong, but the explanation here is a little bit confusing, because it leads to equivalence of terms **similarity** and **distance**. Hence, in accordance with the formula above we will have higher weight for more distant user.

Probably, it worth to mention, that **similarity** and **distance** are mutually inverse measures, e.g. $\mathrm{sim}(a,b) = \frac{1}{\mathrm{dist}(a,b)+1}$

This post is visible to everyone.

Add a Response

1 response

ptressel (Community TA)
8 months ago - endorsed 8 months ago by **dkonomis** (Staff)

That's a good point, Evgeny8!

Sometimes it may be easier to define a measure of how different two things are, than how alike they are, and then convert the difference value into a similarity value.

Hmm, one example of that might be "edit distance" -- how many one-character changes do you need to make to some text string, to convert it into another? (E.g. "cat" and "car" have an edit distance of 1, "cat" and "dam" have an edit distance of 2, "cat" and "chat" differ by 1, etc.) That's fairly straightforward, and probably easier than defining some similarity directly.

Another might be something like Euclidean distance, but when the scales of each axis don't mean the same thing, so they have to be adjusted. For instance, you might want to use the standard deviation of your data in each direction as the scale for that direction. In that case, you could use the Mahalanobis distance, that uses the inverse covariance matrix to rescale the axes. (Except for a constant, that's the square root of the expression in the exponent of a multivariate Gaussian.)

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^\top \Sigma^{-1} (x_1 - x_2)}$$

A formal, mathematical distance $d(x, y)$ has some requirements: It's non-negative: $d(x, y) \geq 0$. It's zero only if and only if the two points are the same: $d(x, y) = 0 \Leftrightarrow x = y$. It's symmetric: $d(x, y) = d(y, x)$. And it satisfies the triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Not every measure of how different things are will have those properties. I see folks using the word "dissimilarity" instead of distance, so it doesn't exclude perfectly useful measures of difference that don't happen to be formal distances.

"*A formal, mathematical distance*"
It's called measure. See also similarity measure.

posted 8 months ago by **Mark_B2** (Community TA)

I meant distance...with those four properties. Not the same as measure. 😁

posted 8 months ago by **ptressel** (Community TA)

Just for fun, here's another definition of infinitesimal distance, which allows one to integrate to find the curve length: The metric tensor. It is frequently used in studying curved spaces such as the spacetime in general relativity, and also seems to have usage in geometric deep learning. I don't understand the detail as I lack knowledge in differential geometry though.

posted 8 months ago by **FutureStar**

Add a comment

Showing all responses

Add a response:

Preview

Submit