CSE 597G Presentation
4/27/2000, 5:30 pm

An Introduction to Collaborative Filtering

Presented by
Anirudh Modi,
Liang Xia
and
Mei Lu

Overview

- What is Collaborative Filtering?
- How to model it?
- Methods?
  - Repeated clustering
  - Gibbs sampling
- Various Applications, etc

What is Collaborative Filtering?

- Collaboration [n.]: The act of working together; cooperating.
- Selection based on overlap of interests. Similar to giving out recommendations to a friend.
  - e.g., I like romantic movies, and I know that my friend X likes some of the same romantic movies. Thus, if I come across a new romantic movie which I like, I recommend it to him/her, and chances of him/her liking it are quite high.
  - A somewhat better analogy: a group of friends working together to decide what gift to buy for another friend's birthday: a farely complex process if you try to formalize it.

What is Collaborative Filtering?

- Real problems are more complex. Consider a problem of recommending a particular movie to a particular person given a database of each.
  - People attributes: age, sex, country of origin
  - Movies attributes: directors, actors, genre [note: directors and actors are facts, whereas *genre* is derived attribute based on the movie's story/enactment, hence is not necessarily unique and is sometimes omitted in such techniques].
- How do we go about recommending one to another?

Methods

•**Statistical Model**: Treat *people* and *movies* as separate classes:

	Batman	Rambo	Andre	Hiver	Whispers	Star Wars
Tom			y			y
Dick			y	y		y
Harry				y	y	
Mr. X	y					y
Mr. Y	y	y				y

	action	foreign	classic
intellectual	0/6	5/9	2/3
fun	3/4	0/6	2/2

Note: most current methods don't do the above classification

Methods

- The statistical model has 3 sets of parameters:
  - P_k = probability a random person is in class *k*
  - P_l = probability a random movie is in class *l*
  - P_{kl} = probability a random person in class *k* likes a movie in class *l*
- P_k and P_l are base rates for the classes, and P_{kl} is estimated from them in the table shown.

Repeated Clustering

- Cluster people and movies separately.
  - Cluster people based on movies they watched and then cluster movies based on people that watched them.
  - The people can then be re-clustered based on the number of movies in each movie cluster they watched.
  - The above process is repeated several times till some convergence is obtained.
- Clustering provides generalization beyond individual movies to groups, and thus should help with sparse data, but it also *smears out* data, and thus may over-generalize.

Gibbs Sampling

- One cannot reclassify a person in a single one person-movie event, as this would make the data inconsistent! He has to be simultaneously reclassified in all other events he occurs in. *Expensive!*
- Gibbs sampling solves this It makes it easy to change the class of a person or movie- and change it simultaneaously in all the events in which they occur.
- Ugly equations which you do not want to know (says who? …says me!) ☺

Applications

- Widely used by online e-tailers to target advertisement (mails) and products effectively.
  - Several e-tailers ask you to fill in a form describing your interests, etc at the time of registration.
  - This information is used for initial classification of the buyers. Their buying habits provide additional information for better classification.
  - e.g., if CDNow notices that a certain person never buys a CD greater than $5, they may label him “thrift” and recommend more titles around that price in the future. The recommendations can be further filtered based on the genre of CDs the customer buys, etc.

Applications

- Product rating by online e-tailers:
  - Feedback is collected when the e-tailers realize that the product has reached the customer and he has had a reasonable time to use it. The data is used to rate the product and based on the rating/feedback, more specific recommendations can be made for other similar items.
  - eBags.com, ValueAmerica.com, Netflix.com use this method.

Applications

- Search engine advertisements:
  - Banner ads are shown based on the query entered by the user.
  - The query is classified, and the ads from the class the query belongs are popped up.
  - Based on the response to the banner ads for a particular query, reclustering is done to give better classification for future queries.
  - Not been implemented yet, but is the future
- News-reading software like PointCast, etc use similar techniques to target the most relevant news to the users based on their profiles.

Some notes

- Collaborative filtering can have some disadvantages when used by e-tailers.
  - Often, a single person may buy stuff for friends or others in the household which may actually be of little or no interest to the person himself.
  - e.g., ***CDNow*** consistently recommends *BackStreet Boys* and similar titles to me, just because I bought 3 of their CDs for a cousin….and I hate that kind of music!!!
  - A better model can be used to correct this which allows a same person to be in several classes and have widely varying tastes. A better feedback model can also do away with such irrelevant data.