Content-based recommenders

Friday, 26 March 2021 07:07

- Based on features of items - similar to typical ML models
- Find correlations between feature items and users

Recommendations for active user u

Explicit feedback

Position	Item	Predicted rating
1	Rocky	4.9
2	Interstellar	4.7
3	Shrek	4.1
4	Star Wars	3.6
⋮	⋮	⋮

Implicit feedback

Position	Item	Probability / score
1	Rocky	0.95
2	Interstellar	0.91
3	Shrek	0.85
4	Star Wars	0.72
⋮	⋮	⋮

Typically items the user has not interacted with are evaluated

Explicit feedback

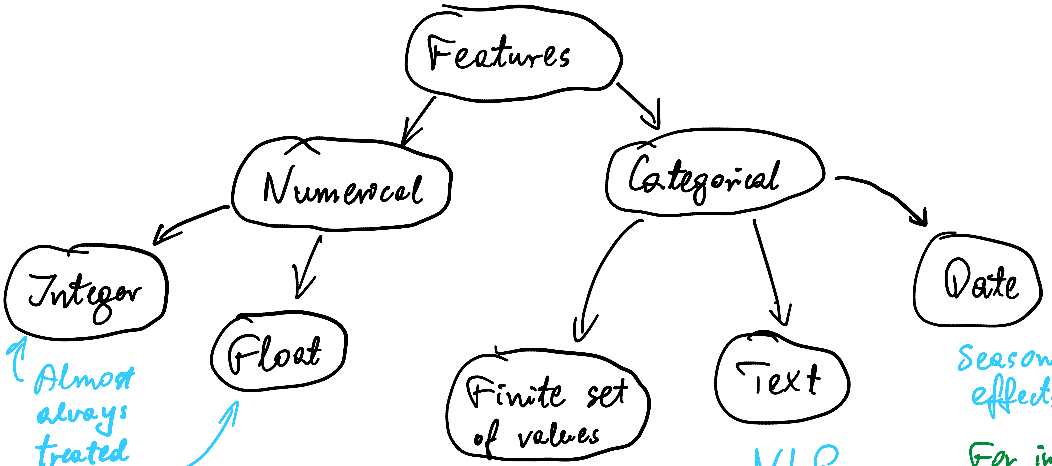
Any regression ML model can be used as a recommender in the explicit feedback case

Implicit feedback

Any classification ML model returning probabilities can be used as a recommender in the implicit feedback case

Non-personalized: one model for all users
Personalized: - one model per cluster of users
- one model per user

Types of features



Almost always treated identically

For instance:
- length of a movie
- box-office result
- number of beds in a hotel room

Typically small

For instance:
- movie genres
- hotel room types
- ids

NLP

For instance:
- movie description

Seasonal effects

For instance:
- arrival date

Categorical finite sets of values - one-hot encoding

One-hot encoding transforms a single column with N possible values into N columns with binary values

movie	genre
movie 1	sci-fi
movie 2	drama
movie 3	comedy
movie 4	sci-fi
⋮	⋮

→

movie	sci-fi	drama	comedy
movie 1	1	0	0
movie 2	0	1	0
movie 3	0	0	1
movie 4	1	0	0
⋮			

NLP

- separate field
- n-grams
- vectorization
- embeddings

Dates

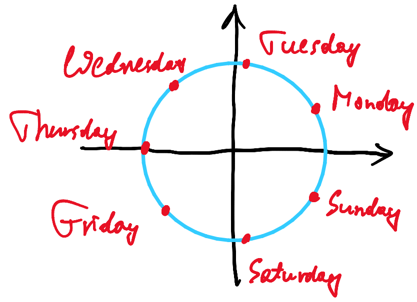
Examples:

One-hot encoded day of week

One-hot encoded month

Day of week projected on a circle

	Monday	Tuesday	Wednesday	...
Monday →	1	0	0	...
Tuesday →	0	1	0	...

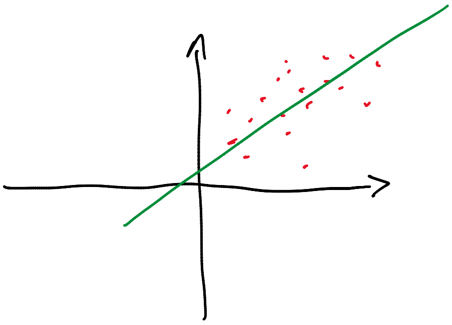


$$\text{Thursday} \to \begin{array}{cc} [-1, & 0] \\ x & y \end{array}$$

Models

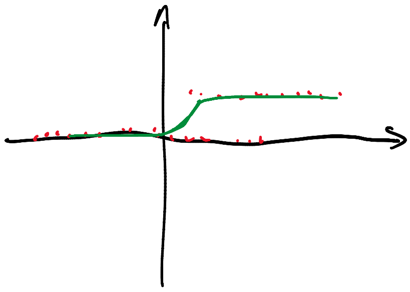
Linear

$$\hat{y} = f(x \mid \theta) = \theta_0 + \theta_1 x_1 + .. + \theta_n x_n$$



Logistic

$$\hat{y} = f(x \mid \theta) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1 - \dots - \theta_n x_n}}$$



$x_1, x_2, \dots, x_n$ – numerical variables

$\theta_1, \theta_2, \dots, \theta_n$ – trained parameters

$$\hat{y} = \begin{cases} \text{real ratings} & : \text{explicit feedback} \\ \text{real binary interactions} & : \text{implicit feedback} \end{cases}$$

Other very popular models

- SVR
- XGBoost
- Random Forest (RF)
- Decision Tree
- Naive Bayes
- Artificial Neural Networks (ANN)

Tuning hyperparameters

Many models have tunable parameters (hyperparameters)

$$\hat{y} = f(x \mid \theta, \mathcal{T})$$

- set $\mathcal{T}$
- train θ on the training set
- evaluate on the validation set

} Iterate for many $\mathcal{T}$

- choose the best $\mathcal{T}$
- evaluate the model on the test set

TF-IDF Term Frequency - Inverse Document Frequency

Based on relative frequencies of feature values for a given user vs all users

user	concatenated genres
1	sci-fi, drama, sci-fi, sci-fi
2	comedy, comedy, drama
3	sci-fi, action, comedy
4	comedy, sci-fi, sci-fi

$tf(1, \text{sci-fi}) = 3$ $tf(1, \text{drama}) = 1$

$tf(2, \text{comedy}) = 2$ $tf(2, \text{drama}) = 1$

$tf(3, \text{sci-fi}) = 1$ $tf(3, \text{action}) = 1$ $tf(3, \text{comedy}) = 1$

$tf(4, \text{comedy}) = 1$ $tf(4, \text{sci-fi}) = 2$

$idf(\text{sci-fi}) = \ln \frac{4}{3}$ $idf(\text{drama}) = \ln \frac{4}{2} = \ln 2$

$idf(\text{comedy}) = \ln \frac{4}{3}$ $idf(\text{action}) = \ln \frac{4}{1} = \ln 4$

$tf\text{-}idf(1, \text{sci-fi}) = 3 \cdot \ln \frac{4}{3}$ $tf\text{-}idf(1, \text{drama}) = 1 \cdot \ln 2$

$tf\text{-}idf(2, \text{comedy}) = 2 \cdot \ln \frac{4}{3}$ $tf\text{-}idf(2, \text{comedy}) = 1 \cdot \ln \frac{4}{3}$

$tf\text{-}idf(3, \text{sci-fi}) = 1 \cdot \ln \frac{4}{3}$ $tf\text{-}idf(3, \text{action}) = 1 \cdot \ln 4$ $tf\text{-}idf(3, \text{comedy}) = 1 \cdot \ln \frac{4}{3}$

$tf\text{-}idf(4, \text{comedy}) = 1 \ln \frac{4}{3}$ $tf\text{-}idf(4, \text{sci-fi}) = 2 \cdot \ln \frac{4}{3}$

To get an item score take its features tf-idf average for a given user

Example:

movie : sci-fi, action
user : 1

$$\text{score} = \frac{\text{tf-idf}(1, \text{sci-fi}) + \text{tf-idf}(1, \text{action})}{2}$$

$$= \frac{3 \cdot \ln \frac{4}{3} + 0}{2} = \frac{3}{2} \ln \frac{4}{3}$$