Machine Learning Techniques

DATASCI 420

Lesson 03 Feature Engineering

Categorical Variables

- We call the number of unique values of a categorical variable the number of levels
- Categorical variables with high cardinality
  - A categorical variable has a large number of levels
- Challenges of variables with high cardinality:
  - Random forest model in R can only handle at most 52 levels of a categorical variable

Categorical variables

- Non-numeric variables with a finite number of levels
  - E.g. "red", "blue", "green"
- Some ML algorithms can only handle numeric variables
- Solution 1: One hot encoding

One hot Encoding

feature
red
blue
green
red
red
green
blue

red	blue	green
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1
0	1	0

Dealing with Categorical Attributes

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:

Outlook (overcast, rain, sunny)
Temperature real
Humidity real
Windy (true, false)
Play (yes, no)

Standard Spreadsheet Format

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy	Play	Play
overcast	rain	sunny			TRUE	FALSE	yes	no
0	0	1	85	85	0	1	1	0
0	0	1	80	90	1	0	0	1
1	0	0	83	78	0	1	1	0
0	1	0	70	96	0	1	1	0
0	1	0	68	80	0	1	1	0
0	1	0	65	70	1	0	0	1
1	0	0	64	65	1	0	1	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Problem of One Hot Encoding

- It significantly widens the dataset
  - If you have zip code as a feature in your dataset
  - There are approximately 43,000 zip codes in US
  - It means after one hot encoding, you will have 43,000 binary columns to represent zip codes
  - You may have other categorical variables…
  - Sometimes exceeds the memory limitation

Categorical Variable: Risk Value

- Calculate the risk value of each level of a categorical variable:

$$R_i = \log\left(\frac{\Pr(Y=1 \mid X=x_i)}{\Pr(Y=0 \mid X=x_i)}\right)$$

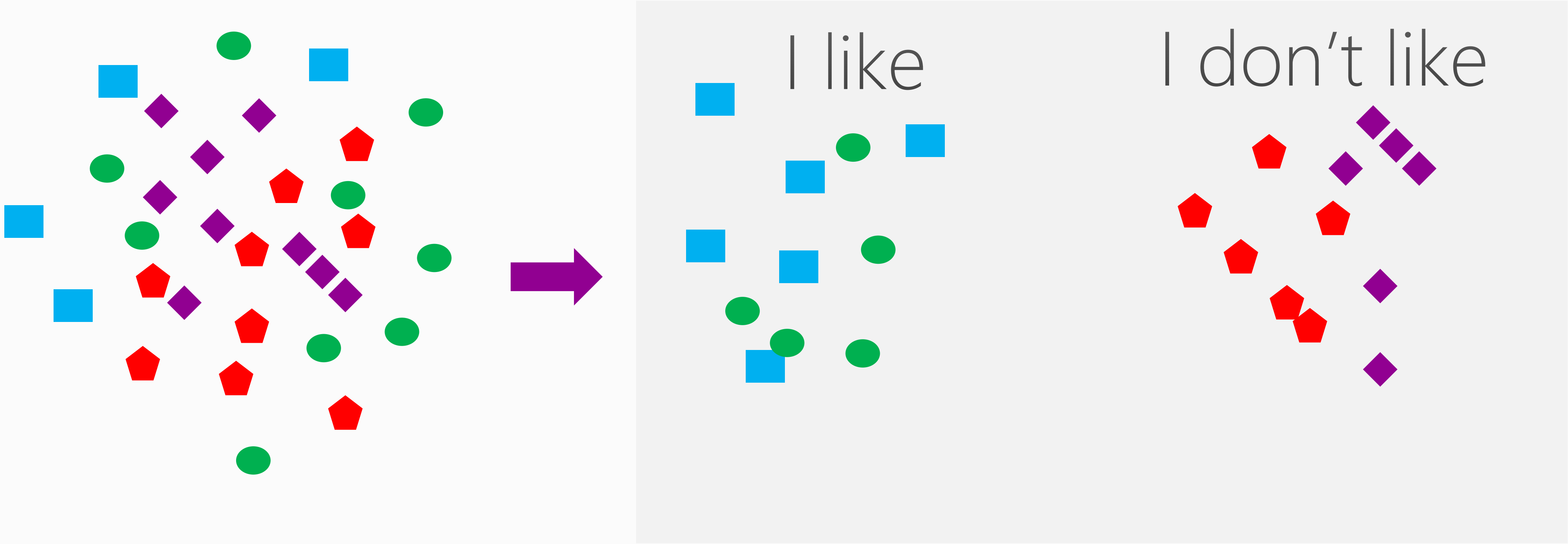
$$\Pr(Y=1 \mid X=x_i) \approx \frac{N_{Y=1\,\&\,X=x_i}}{N_{X=x_i}}$$

- Use risk values to replace categorical levels in the data
- Avoids widening the dataset
- Converts the categorical values to numerical values, preferable by many models
- The higher risk value of a level, the higher probability that the target variable = 1

Feature Engineering

- Better to have a fair modeling method and good variables, than to have the best modeling method and poor variables
- Insurance Example: People are eligible for pension withdrawal at age 59 ½. Create it as a separate Boolean variable!
- Advanced methods exist for automatically examining variable combinations, but they can be computationally very expensive!

Example of Feature Engineering



Recency, Frequency, and Magnitude (Monetary)

- A set of features very popularly used in customer churn problem, and other domains where analog exists
- Recency: How long has it passed since the most recent interaction between the user and the system
- Frequency: In the past given time period (1 month, 1 week, etc.), how many times the user interacted with the system
- Monetary: In the past given time period, how much (time, money, etc.) the user has spent on the system

Example of RFM Calculation

UserId	Age	Address	Column 0	Transactio	Timestamp	ItemId	Quantity	Value
1113	K	F	118152	904890	11/12/2000 0:00	4710000000000	2	29
1113	K	F	118153	905431	11/12/2000 0:00	4900000000000	3	391
1113	K	F	118154	1000113	11/26/2000 0:00	4900000000000	1	111
1113	K	F	118155	1000416	11/26/2000 0:00	7620000000000	1	268
1113	K	F	118156	1000417	11/26/2000 0:00	4710000000000	1	179
1113	K	F	118157	1018276	11/27/2000 0:00	4710000000000	1	14
1113	K	F	118158	1019142	11/27/2000 0:00	4720000000000	1	224
1113	K	F	118159	1019267	11/27/2000 0:00	4710000000000	1	65
1113	K	F	118160	1019384	11/27/2000 0:00	4710000000000	1	116
1113	K	F	118161	1019478	11/27/2000 0:00	4710000000000	1	116
1113	K	F	118162	1019482	11/27/2000 0:00	4710000000000	1	89
1113	K	F	118163	1282039	1/6/2001 0:00	4710000000000	1	188
1113	K	F	118164	1284131	1/6/2001 0:00	4710000000000	1	28
1113	K	F	118165	1284189	1/6/2001 0:00	4710000000000	2	84
1113	K	F	118166	1284585	1/6/2001 0:00	37000440147	1	47
1113	K	F	118167	1284765	1/6/2001 0:00	4900000000000	1	169
1113	K	F	118168	1284951	1/6/2001 0:00	9560000000000	1	28
1113	K	F	118169	1285516	1/6/2001 0:00	4710000000000	2	84
1250	D	D	243280	1494035	2/4/2001 0:00	4720000000000	1	148
1250	D	D	243281	1494721	2/4/2001 0:00	4720000000000	1	179
1250	D	D	243282	1494852	2/4/2001 0:00	4910000000000	1	309
1250	D	D	243283	1495078	2/4/2001 0:00	4720000000000	2	98
1250	D	D	243270	1451064	2/10/2001 0:00	4710000000000	1	89
1250	D	D	243271	1451293	2/10/2001 0:00	4710000000000	1	65
1250	D	D	243272	1451301	2/10/2001 0:00	723000000000	1	65
1250	D	D	243273	1451534	2/10/2001 0:00	20480349	1	395
1250	D	D	243274	1451641	2/10/2001 0:00	4710000000000	2	44
1250	D	D	243275	1451863	2/10/2001 0:00	4710000000000	1	28
1250	D	D	243276	1452120	2/10/2001 0:00	4710000000000	2	26
1250	D	D	243277	1452219	2/10/2001 0:00	4710000000000	2	44
1250	D	D	243278	1452444	2/10/2001 0:00	4710000000000	1	28
1250	D	D	243279	1452672	2/10/2001 0:00	4710000000000	1	65

- If we set the checkpoint 11/20/2000
- Recency: 11/12 – 11/20 = 8 days
- Frequency: 2 (2 transactions)
- Monetary Value: 29+391 = 420
- Monetary Quantity: 2+3 = 5

- If we want the frequency and monetary in the most recent 7 days:
- Frequency = 0
- Monetary Value and Quantity = 0