Understating the DATA sense & Pre-processing

GROKKERS
AI FOR EVERYONE

Pre-reqs

- Python
- NumPy and PANDAS, SciPy, Visualizations
- Elementary stats and maths
- Some preprocessing steps – may need ML as well, for advanced topics

Background

Numeric Data: Preprocessing involves handling missing values, scaling to a similar range, and possibly normalizing the distribution.

Text Data**:** Common preprocessing steps include text cleaning (removing stop words, punctuation, etc.), tokenization, and vectorization (converting text into numerical form, such as TF-IDF or word embeddings).

Image Data: Techniques like resizing, normalization of pixel values, and data augmentation (creating variations of existing images) are often used.

Time Series Data: Dealing with temporal aspects, handling missing values over time, and creating lag features are important steps in preprocessing time series data.

Topics

- About Data, feature types, tabular form
- General inspection of data quality
- Handling duplicates in data
- Missing value analysis (2 parts)
- Handling Outliers
- Cardinality assessment
- Encoding of discrete data
- Scaling and Normalization
- Handling Skewed Distributions
- Data Imbalance Handling
- **Data Splitting**

About Data

Preparing data tables

Data table

In the context of machine learning, a data table, also known as a dataset or data matrix, refers to a structured arrangement of data

rows typically represent individual instances or observations,

columns represent attributes, features, or variables associated with those instances.

A data table is a fundamental building block in machine learning

17-08-2024

example

columns represent attributes such as age, gender, height, weight

each row represents a person

ID	Age	Gender	Height	Weight	Class
1	25	Male	178	75	A
2	30	Female	162	58	B
3	22	Male	185	82	A
4	28	Female	170	63	B

class label indicating the group the person belongs to

names

Record

Columns

Record, samples, point, case, entity, instance, entry, Objects

Data points, Document, tuple, Transaction, feature vector

Attributes, Features, Variables

Field, Predictors, characteristics

Data science a study of 3 or 4 different disciplines. Hence there are a lot of vocab, often many for the same term!

FEATURE TYPES

- independent variable,
  - sometimes called an experimental or predictor variable,
  - is a variable that is being manipulated in an experiment in order to observe the effect …

- dependent variable, sometimes called an outcome/response/target variable.

Example

- Dependent Variable: Test Mark (measured from 0 to 100)
- Independent Variables:
  - Revision time (measured in hours),
  - Intelligence (measured using IQ score)

- Examples

Sample dataset (temp forecasts)

year	month	day	week	Temp 2 days before	Temp 1 day before	Average temp	Actual temp on that day	Temp forecast by noaa	Temp Forecast by acc	Temp forecast by friend
2016	1	1	Fri	45	45	45.6	45	43	50	29
2016	1	2	Sat	44	45	45.7	44	41	50	61
2016	1	3	Sun	45	44	45.8	41	43	46	56
2016	1	4	Mon	44	41	45.9	40	44	48	53
2016	1	5	Tues	41	40	46	44	46	46	41
2016	1	6	Wed	40	44	46.1	51	43	49	40
2016	1	7	Thurs	44	51	46.2	45	45	49	38
2016	1	8	Fri	51	45	46.3	48	43	47	34
2016	1	9	Sat	45	48	46.4	50	46	50	47
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
2016	1	19	Tues	50	54	47.6	48	47	49	53
2016	1	20	Wed	54	48	47.7	52	44	52	61
2016	1	21	Thurs	48	52	47.8	52	43	51	57

Independent variables

Independent variables

Dependent variable

Sample dataset (loan applications)

Loan_ID	Gender	Married	Number of Dependents	Education	Self Employed ?	Applicant Income	Co-applicant Income	Loan Amount	Loan Amount Term	Credit History	Property Area
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban
LP001824	Male	Yes	1	Graduate	No	2882	1843	123	480	1	Semiurban
LP002928	Male	Yes	0	Graduate	No	3000	3416	56	180	1	Semiurban
LP001814	Male	Yes	2	Graduate	No	9703	0	112	360	1	Urban
LP002244	Male	Yes	0	Graduate	No	2333	2417	136	360	1	Urban
LP001854	Male	Yes	3+	Graduate	No	5250	0	94	360	1	Urban
…	Male	Yes	0	Graduate	No	3500	1667	114	360	1	Semiurban
LP001647	Male	Yes	0	Graduate	No	9328	0	188	180	1	Rural
LP001871	Female	No	0	Graduate	No	7200	0	120	360	1	Rural
LP001379	Male	Yes	2	Graduate	No	3800	3600	216	360	0	Urban
LP002789	Male	Yes	0	Graduate	No	3593	4266	132	180	0	Rural
LP001578	Male	Yes	0	Graduate	No	2439	3333	129	360	1	Rural
LP001318	Male	Yes	2	Graduate	No	6250	5654	188	180	1	Semiurban
LP001259	Male	Yes	1	Graduate	Yes	1000	3022	110	360	1	Urban
LP002804	Female	Yes	0	Graduate	No	4180	2306	182	360	1	Semiurban

Independent variables

Dependent variable ?

SAMPLE DATASET (automobiles)

KMs per liter	cylinders	displacement	horsepower	weight	acceleration	year	origin	Model name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
24	4	113	95	2372	15	70	3	toyota corona mark ii
22	6	198	95	2833	15.5	70	1	plymouth duster
18	6	199	97	2774	15.5	70	1	amc hornet
21	6	200	85	2587	16	70	1	ford maverick
27	4	97	88	2130	14.5	70	3	datsun pl510

Dependent variable ?

Sample dataset (diabetes related)

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Diabetic ? (1 = Yes, 0= No)
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

Dependent variable ?

Sample dataset (cancer related)

code	Clump Thickness	Cell Size	Cell_Shape	Adhesion	Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Cancer Stage
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4

Dependent variable ?

Sample dataset (air quality)

Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
3/10/2004	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
3/10/2004	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
3/10/2004	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
3/10/2004	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
3/10/2004	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
3/10/2004	23:00:00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
3/11/2004	0:00:00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603
3/11/2004	1:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0	0.7702
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
3/11/2004	9:00:00	2.2	1351	87	9.5	960	129	1079	101	1583	1028	10.5	60.6	0.7691
3/11/2004	10:00:00	1.7	1233	77	6.3	827	112	1218	98	1446	860	10.8	58.4	0.7552
3/11/2004	11:00:00	1.5	1179	43	5.0	762	95	1328	92	1362	671	10.5	57.9	0.7352
3/11/2004	12:00:00	1.6	1236	61	5.2	774	104	1301	95	1401	664	9.5	66.8	0.7951
3/11/2004	13:00:00	1.9	1286	63	7.3	869	146	1162	112	1537	799	8.3	76.4	0.8393
3/11/2004	14:00:00	2.9	1371	164	11.5	1034	207	983	128	1730	1037	8.0	81.1	0.8736

Attribute Information:

0 Date (DD/MM/YYYY)
1 Time (HH.MM.SS)
2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)

7 True hourly averaged NOx concentration in ppb (reference analyzer)
8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12 Temperature in Â°C
13 Relative Humidity (%)
14 AH Absolute Humidity

Sample data - jobs

Month	Total Filled Jobs
2004M07	1795610
2004M08	1792770
2004M09	1809590
2004M10	1815580
2004M11	1856360
2005M04	1871630
2005M05	1867870
2005M06	1857260
2005M07	1858360
2005M08	1856320
2005M09	1876270
2005M10	1866920
2011M10	1903630
2011M11	1940200
2011M12	1983070
2012M01	1865540
2012M02	1932380

Independent variables ?

Dependent variable ?

Sample data - text

Tweet d	Airline sentiment	Retweet count	text	tweet_location
570306133677760000	neutral	0	=@VirginAmerica What @dhepburn said.	
570301130888122000	positive	0	@VirginAmerica plus you've added commercials to the experience... tacky.	
570301083672813000	neutral	0	@VirginAmerica I didn't today... Must mean I need to take another trip!	Lets Play
570301031407624000	negative	0	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp; they have little recourse	
….	….	….	…. ….	….
570300767074181000	negative	0	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA	
570300616901320000	positive	0	@VirginAmerica yes, nearly every time I fly VX this â€œear wormâ€•wonâ€™t go away :)	San Francisco CA
570300248553349000	neutral	0	@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP	Los Angeles
570285904809598000	positive	0	@VirginAmerica Thanks!	San Francisco, CA
570282469121007000	negative	0	=@VirginAmerica SFO-PDX schedule is still MIA.	palo alto, ca
570277724385734000	positive	0	@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but great things about Virgin America. #29DaysToGo	west covina
570276917301137000	negative	0	@VirginAmerica I flew from NYC to SFO last week and couldn't fully sit in my seat due to two large gentleman on either side of me. HELP!	this place called NYC
570270684619923000	positive	0	I â�ï¸�flying @VirginAmerica. â˜ºï¸�ðŸ‘•	Somewhere celebrating life.
570267956648792000	positive	0	@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly with only you.	Boston | Waltham
570265883513384000	negative	0	=@VirginAmerica why are your first fares in May over three times more than other carriers when all seats are available to select???	

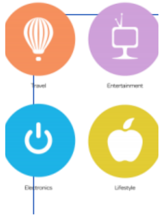
Independent variables ?

Dependent variable ?

- Feature types

Feature types

Numerical Features: age, temperature, height, and income



Categorical Features: variables that represent different categories or labels.

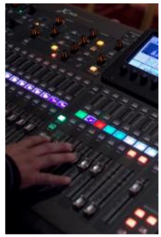
Textual Features: such as documents, sentences, or paragraphs.

Date and Time Features represent temporal information.

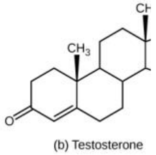
Boolean Features: take on two possible values: True or False.

Geospatial Features: location-based data, such as latitude, longitude, postal codes, or addresses.

Image and Video Features: computer vision tasks.



Audio Features: speech recognition, music analysis, and more.



Derived Features: features are created by transforming or combining existing features.

17-08-2024

Numerical variables

Continuous Numerical Variables:

- **Definition:** can take an infinite number of possible values within a given range.
- **Examples:**
  - **Height**: The height of a person can be any value within a certain range and is not restricted to specific discrete values.
  - **Temperature**: Temperature can be measured with great precision, allowing for an infinite number of possible values.

Discrete Numerical Variables:

- **Definition:** can only take specific, separate values, often integers.
- **Examples:**
  - Number of Siblings: The count of siblings is a discrete numerical variable, as it can only be a whole number.
  - Number of Cars in a Parking Lot: The count of cars is discrete, and you can't have a fraction of a car.

Numerical variables

Ratio Variables:

- have a true zero point, meaning that a value of zero indicates the absence of the variable.
- Examples include height, weight, and income.

Interval Variables:

- while numeric, lack a true zero point. Zero does not imply the absence of the variable.
- Examples include temperature measured in Celsius or Fahrenheit.

DISCRETE VARIABLES

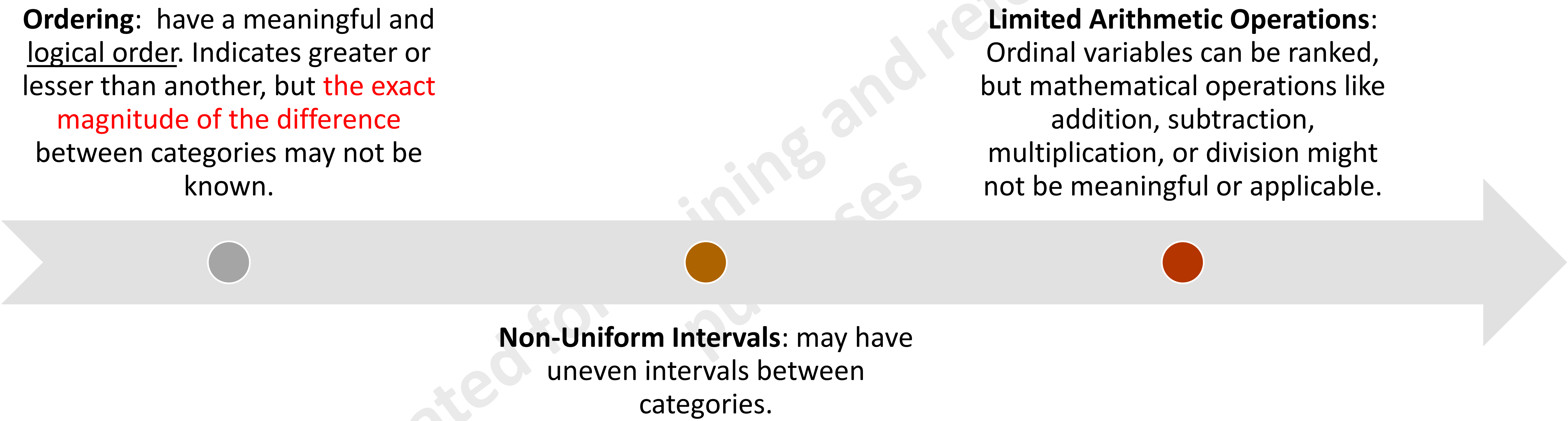
- a type of numerical variable that can only take specific integer values within a certain range.
- values are often counted, and there are gaps between them.
- can't be subdivided further.

- Examples of discrete variables:
  - The number of students in a classroom
  - The number of cars in a parking lot
  - The number of items purchased

CATEGORICAL VARIABLES

- also known as a **qualitative** variable, represents data in categories or labels that have no inherent numerical order or ranking.
- categories are usually distinct and don't have mathematical operations applied to them.
- characteristics or attributes that fall into different groups.
- **Examples** of categorical variables:
  - Colors (red, blue, green, etc.).
  - Types of fruits (apple, banana, orange, etc.).
  - Payment methods (credit card, cash, PayPal).

Ordinal variables - a type of categorical variable



Ordering: have a meaningful and logical order. Indicates greater or lesser than another, but the exact magnitude of the difference between categories may not be known.

Non-Uniform Intervals: may have uneven intervals between categories.

Limited Arithmetic Operations: Ordinal variables can be ranked, but mathematical operations like addition, subtraction, multiplication, or division might not be meaningful or applicable.

Examples of ordinal variables:

- Education Levels: (High School, Associate's, Bachelor's, Master's, PhD)
- Socioeconomic Status: (Low, Middle, High)
- Customer Satisfaction Ratings: (Poor, Fair, Good, Very Good, Excellent)
- Pain Intensity Levels: (Mild, Moderate, Severe)

Nominal variables - a type of categorical variable

Ordering: no inherent order or ranking between them.

Non-Uniform Intervals: may have uneven intervals between categories.

No Arithmetic Operations: do not support arithmetic operations like addition, subtraction, multiplication, or division, as the categories have no numerical meaning.

Examples of nominal variables:

- Colors: (Red, Blue, Green, etc.)
- Types of Fruits: (Apple, Banana, Orange, etc.)
- Payment Methods: (Credit Card, Cash, PayPal)
- Countries: (USA, Canada, France, etc.)

binary

has only two categories or levels.

categories are often represented as "1" and "0," "Yes" and "No," or "True" and "False."

Examples of dichotomous variables:

- Gender: (Male, Female)
- Smoker: (Yes, No)
- Voter: (Voted, Did Not Vote)
- Married: (Married, Not Married)

Test your understanding

Measurements	Nominal	Ordinal	Dichotomous	Interval	Ratio
Favorite candy bar	☑				
Weight of luggage					☑
Year of your birth				☑	
Egg size (small, medium, large, extra large, jumbo)		☑			
Military rank		☑			
Number of children in a family					☑
Jersey numbers for a football team	☑				
Shoe size				☑	
Number of emergency room patients					
Tumor size					
Religious preference: Buddhist, Muslim, Jewish, Christian, Hindu	☑				
Likert Scale: strongly disagree, disagree, neutral, agree, strongly agree.					
IQ (intelligence scale)					

Thanks !!

AI FOR EVERYONE

Grokkers

CONCEPTS AND EQUATIONS MADE SIMPLE