Computing

CS5228: Knowledge Discovery and Data Mining

Tutorial 1 — Data, Data Cleaning, Data Preprocessing

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44
5	male	group-c	UM	some college	xxxxx	8 to 9	none	76.0	78.0	0.75
6	female	group b	OK	Associate's Degree	xxxxx	6 to 7	none	71.0	83.0	0.78
7	female	group-b	PA	Some College	xxxxx	6 to 7	completed	88.0	95.0	0.92
8	male	group-b	GA	Some College	xxxxx	> 9	none	40.0	43.0	0.39
9	male	group d	LA	high school	xxxxx	6 to 7	completed	64.0	64.0	0.67
10	female	group b	KY	High School	xxxxx	7 to 8	none	38.0	60.0	0.50
11	male	group c	AZ	Associate's Degree	xxxxx	< 6	none	58.0	54.0	0.52
12	male	group d	OK	associate's degree	xxxxx	8 to 9	none	40.0	52.0	0.43
13	female	group-b	OR	High School	xxxxx	7 to 8	none	65.0	81.0	0.73
14	male	group-a	SD	some college	xxxxx	6 to 7	completed	78.0	72.0	0.70
15	female	group a	KY	master's degree	xxxxx	8 to 9	none	50.0	53.0	0.58
16	female	group-c	CO	Some High School	xxxxx	< 6	none	69.0	75.0	0.78
17	male	group-c	UM	high school	xxxxx	7 to 8	none	88.0	89.0	0.86
18	female	group b	KY	Some High School	xxxxx	> 9	none	18.0	32.0	0.28
19	male	group-c	WI	Master's Degree	xxxxx	6 to 7	completed	46.0	42.0	0.46
20	female	group-c	NE	associate's degree	xxxxx	6 to 7	none	54.0	58.0	0.61

Question 1

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

1. **Types of Attributes.** For each attribute, decide whether it is *nominal*, *ordinal*, *interval*, or *ratio*. For which attributes might this decision not so clear?

	id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
nominal											
ordinal											
interval											
ratio											

Question 1

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

1. **Types of Attributes.** For each attribute, decide whether it is *nominal*, *ordinal*, *interval*, or *ratio*. For which attributes might this decision not so clear?

	id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
nominal	✔	✔	✔	✔		✔		(✔)			
ordinal					✔		✔	(✔)			
interval											
ratio									✔	✔	✔

Question 2

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

2. **Data Cleaning.** Just by looking at these 20 samples, which data cleaning steps seem recommended? Note that this refers only to preprocessing steps to remove potential noise from the dataset, not any steps that might further benefit a subsequent analysis (arguably, there is no clear distinction, but we cover these steps in the next questions).

Question 2

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

2. **Data Cleaning.** Just by looking at these 20 samples, which data cleaning steps seem recommended? Note that this refers only to preprocessing steps to remove potential noise from the dataset, not any steps that might further benefit a subsequent analysis (arguably, there is no clear distinction, but we cover these steps in the next questions).

Solution

- Normalize ethnicity (e.g. ”group c” vs ”group-c”)
- Normalize parent education (e.g., convert to all lowercase)
- Adjust scale of write score

Question 3

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

3. **Attribute/Feature Importance.** For each attribute, assess its importance (or relevance, usefulness, etc.) for a subsequent analysis. Let's assume we want to predict students' *_scores based on the other attributes.

Question 3

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

3. **Attribute/Feature Importance.** For each attribute, assess its importance (or relevance, usefulness, etc.) for a subsequent analysis. Let's assume we want to predict students' *_scores based on the other attributes.

Solution

- remove id (just an ”artificial” attribute)
- remove email (it's redacted anyway, and so of no use)
- probably remove state (we only have 1,000 sample and there a 50+ states, so the information w.r.t. students' state is very sparse and far from representative

Question 4

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

4. **Additional Data Preprocessing.** Many to most off-the-shelf data mining algorithms for clustering or classification/regression require numerical data as input. How does this affect the analysis of this dataset, and what can we do to address this using data preprocessing?

Question 4

id	gender	ethnicity	state	parent_education	email	sleeping_hours	prep_course	math_score	read_score	write_score
1	female	group b	MO	Bachelor's Degree	xxxxx	7 to 8	none	72.0	72.0	0.74
2	female	group c	VA	Some College	xxxxx	< 6	completed	69.0	90.0	0.88
3	female	group b	PR	master's degree	xxxxx	7 to 8	none	90.0	95.0	0.93
4	male	group a	CT	associate's degree	xxxxx	7 to 8	none	47.0	57.0	0.44

4. **Additional Data Preprocessing.** Many to most off-the-shelf data mining algorithms for clustering or classification/regression require numerical data as input. How does this affect the analysis of this dataset, and what can we do to address this using data preprocessing?

Solution

- gender and prep_course seem to be both binary attributes, so 0/1 encoding should do just fine
- parent education can be converted into simple ranking of numerical values (e.g., 1, 2, ...)
- state needs to be encoded, but one-hot encoding is problematic (large number of possible values)
- converting the sleeping hours strings into a numerical estimate would probably be useful

Notes

Notes

Notes