Missing Values

UNIVERSITY OF
LIVERPOOL

Missing values

- What if we do not know the value of a particular feature of an instance?
- **Example**: the height of the 10-th student is missing although other feature values are known for the 10-th student and all the other students in the dataset
- May be it is important to know this missing value because the height is an essential feature for our classification task such as obesity.

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	205	85
11	184	80

Handling missing values #1: discard

- **Discard** the entire training instance
  - might get away because of the redundancy in the dataset. We might have another student with the same height as student no 10 for whom we have measured the height.
  - might not get away because student no 10 was the only student who had obesity. By ignoring student no 10, we loose all the information we had about the positive class.

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
11	184	80

Handling missing values #2: fill in values by hand

- Re-annotate the data or re-measure the instances with missing feature values
  - A reliable method to overcome the missing value problem
  - Might not be possible in practice because we no longer have access to the subjects.
  - Too slow (manual work)
  - Costly (manual work)
  - There might be lots of data points with missing feature values

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	205	85
11	184	80

Handling missing values #3: set “missingValue”

- We consider “missing” as another category for the feature, and set some constant indicating that the value is missing such as “**missingValue**”.
- Not possible for numerical data. Does “0” mean the value was actually measured and turned out to be zero, or was it simply missing (possibly non-zero) in the dataset?
- Does not actually solve anything!

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	missingValue	85
11	184	80

Handling missing values #4: replace with the mean

- Compute the mean of the available feature values for the entire training dataset and replace the missing values by this “sample” mean.
- This might be a good option if the data points with missing values are representative samples in the dataset.
- But if those data points are outliers this method is inaccurate

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	**178**	85
11	184	80

Handling missing values #5: predict

- We can train a new classifier to first predict the missing values in data instances and then train a second classifier to predict the target class using all (original + missing values predicted) the data points.

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	205	85
11	184	80

Handling missing values #6: accept missing values

- Just leave the data points with missing values as they are, and let the algorithm (eg. classifier) deal with the missing values in an appropriate manner
- The classifier might first try to come up with a rule to classify data without using features that have missing values. If it can do so with high accuracy, then we are fine. Nothing to worry about missing values.

ID	Height (cm)	Weight (kg)
1	175	60
2	168	80
3	183	85
4	178	65
5	203	90
6	185	78
7	155	98
8	180	75
9	173	65
10	205	85
11	184	80