Noisy Data

UNIVERSITY OF
LIVERPOOL

Noisy data

- By “noisy data” we mean random errors scattered in the data
  - eg: due to inaccurate recording, data corruption
- Why is it a problem?
  - Overfitting
    - If we assume the noisy values are correct values then we will learn classifiers to predict the noise as well. This could not be possible because it is “random” noise OR we might end up learning a classifier that learns “too much” from the train data and does not generalize to test data
      - This phenomenon is called “**over-fitting**” (fitting too much to the train data such that the model does not work well outside the given train dataset)

Detecting noisy data

- Some noise will be very obvious:
  - data has incorrect type (string in numeric attribute)
  - data is very dissimilar to all other entries (in one instance a feature has value 10; in all others it is the range $[0,1]$)
- Some incorrect values might not be obvious
  - eg. typing 0.52 instead of 0.25

Handling noisy values

- Manual inspection and removal
- Use clustering on the data to find instances or features that lie outside the main clusters (outlier detection) and remove them
- Use linear regression to determine the function, then remove those that lie far from the predicted value
- Ignore all values that occur below a certain frequency threshold
  - effective for detecting misspellings in text
- If noisy data points can be identified and removed, we can apply missing value techniques to fill the missing features.