Day 13, Oct-15, 2024

Difference in Scales:

Sparse Regression is sensitive to the units used to train them (so is deep learning).

for example, this means that changing the units of a certain input from meters to milimeters will have a major influence on the fitted model. To address this issue, it is common to try rescaling each variable.

for example, Subtracting the mean and dividing by the deviation

$$z = \frac{x - \mu}{\sigma}$$

will ensure all the variables have approximately the same range. It is also possible to convert any variable x_d to range $[0, 1]$ using the transformation

$$= \frac{x - x_{min}}{x_{max} - x_{min}}$$

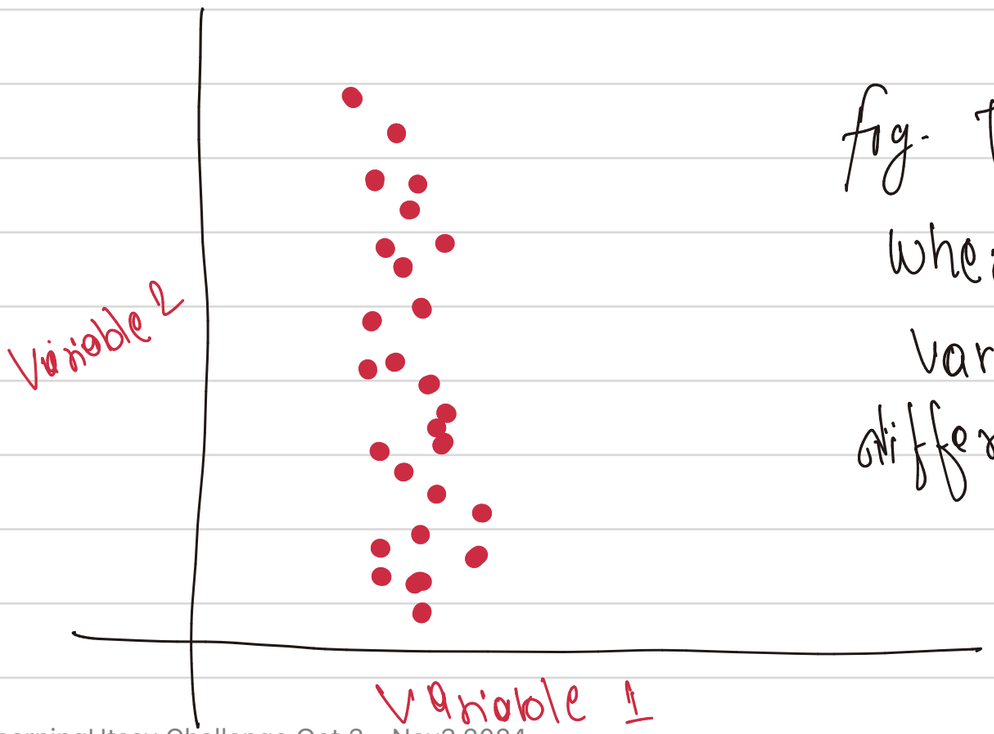


fig. the plot illustrates a 2D scatter plot where two variables (var1 on the x-axis and var2 on the y-axis) showing relationship between different points.

Missing Values:

For some types of data, features have to be dropped for some samples. The sensor might have a temporary breakdown, or a respondent might have skipped a question, for example.

Unfortunately, most methods do not have natural ways for handling missingness.

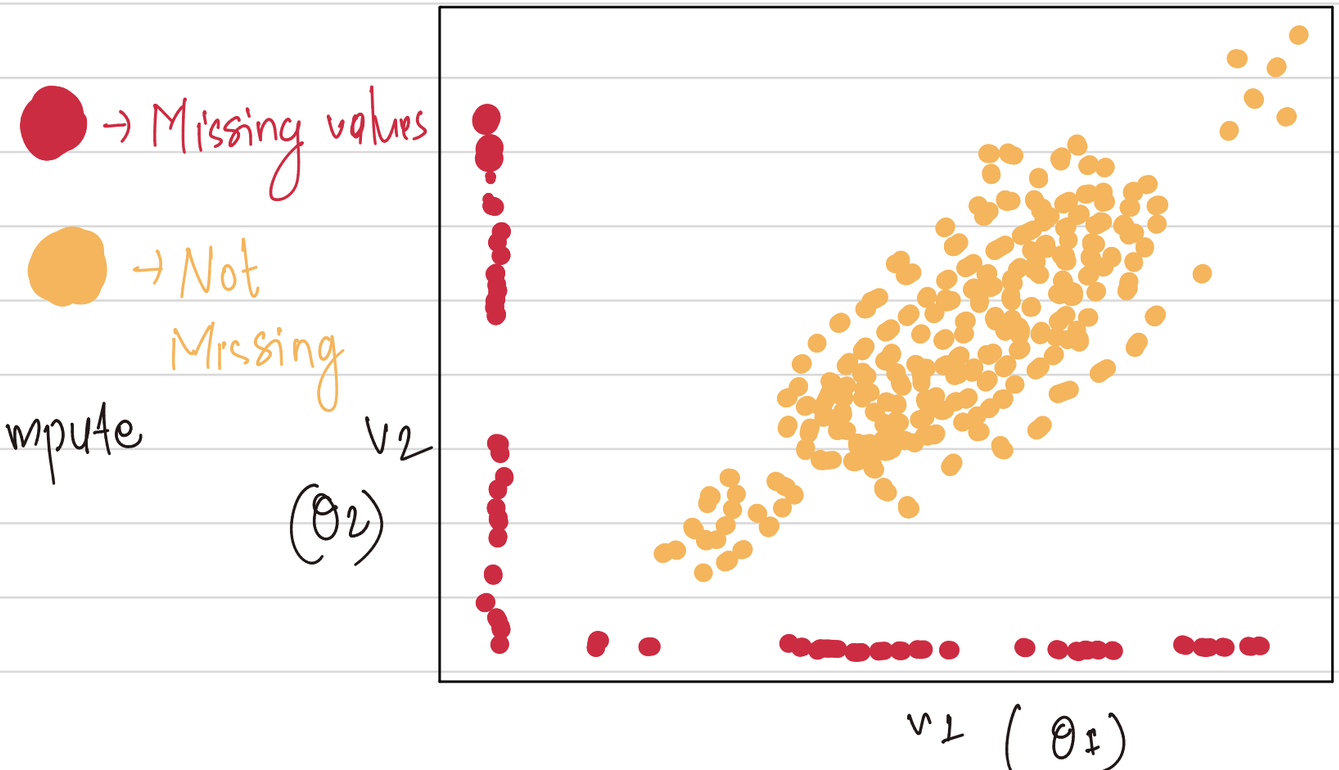
A standard approach to imputation is to replace the missing values in a column using the median of all observed values in that column. This is called median imputation.

#for example in the Tree we had some missing values.

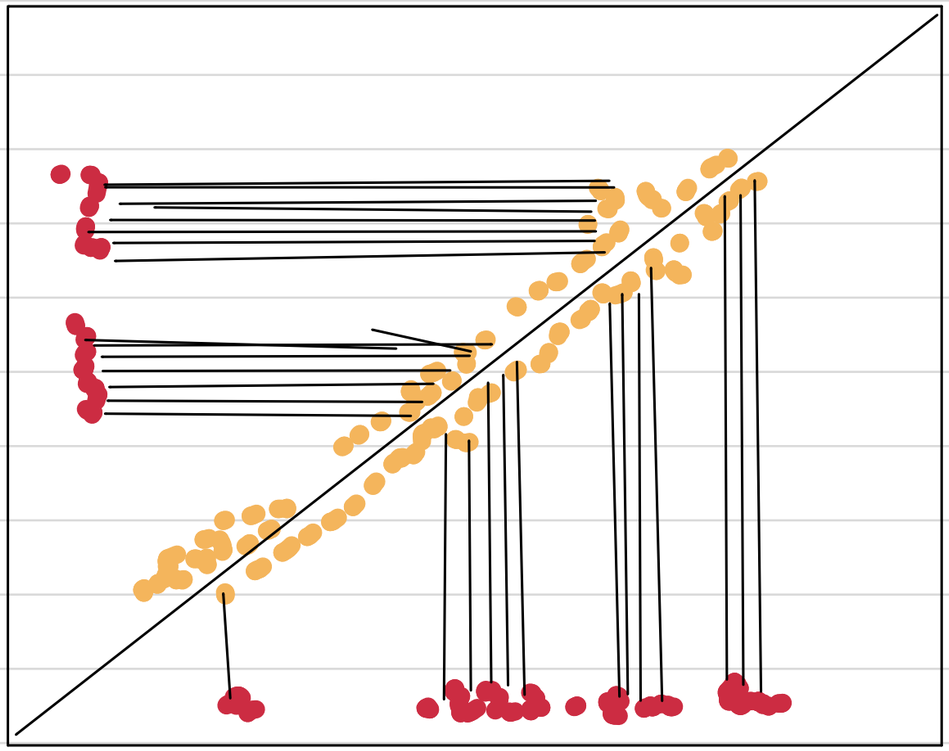
In the above figure, each row and column have observations data we can see Ozone Sensor, Solar Sensor frequently breaks-down.
As, we can see the black lines shows the missing values.
the figure is also called multiple imputations; if we learn, understand the row and column and learn better imputations.

Figure shows an example

Where correlation between two variables might be used to impute their missing values (exactly one is present)



Geometric Intrepretation on finding the missing value (Imputation)



① Find the Red dot:
This is the data with missing number

② Look at the Orange dot: These have both numbers

③ follow the Line: the black connects the red dot to the nearest orange dot.

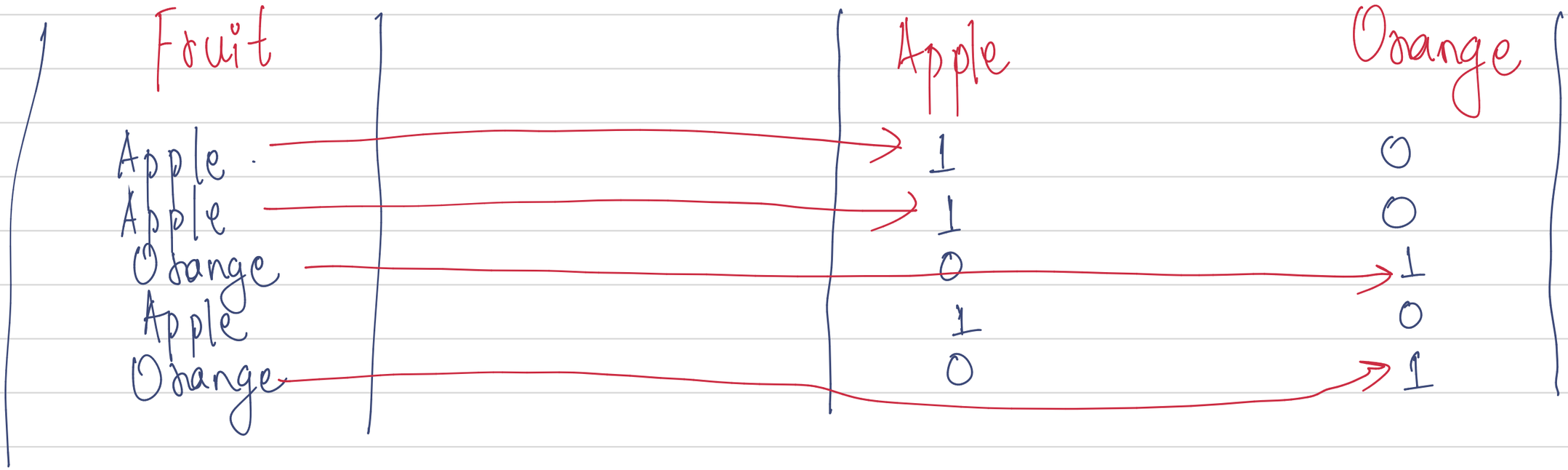
④ Copy the Original Dot's value: the missing number for the red dot is the same as the orange dot it's connected to.

word meaning.
Imputation: the assignment of value to something by inference from the value of the products.

Be especially careful with numerical values that have been used as a proxy for missingness (eg. sometimes -99 is used as a placeholder) these should not be treated as actual numerical data, they are fact missing!

Categorical Inputs:

from the last set of notes, only tree-based methods can directly use categorical variables as inputs. for other methods, the categorical variables needed to be coded. for example, variables with two levels can be converted to 0's & 1's, and variables with K levels can only be hot-coded,



Sometimes a variable has many levels. For example, a variable might say which city a user was from. In some cases, a few categories are very common, with the rest appearing only a few times.

For the first situation, one solution is to replace the level of the category with the average value of the response within that category.

This is called response coding.

For the second, it's possible to lump all the rare categories into an "Other" column.

Note:

To detect the types of issues discussed here, a good practice is to Compute i) Summary statistics and

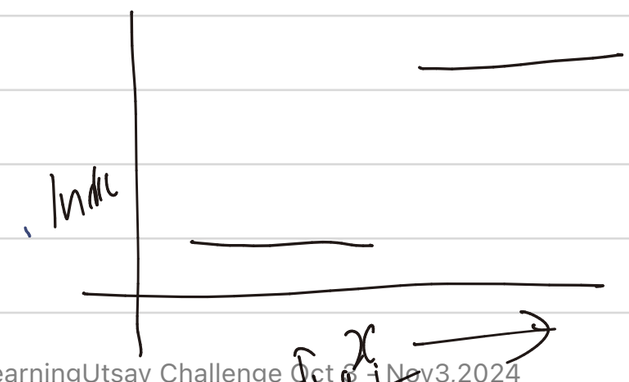
ii) a histogram / barplot for every feature in the dataset. Even if there are a few hundred features, it's possible to skim the associated summaries relatively quickly (just a few seconds for each).

#(Skim → an act of reading something quickly)

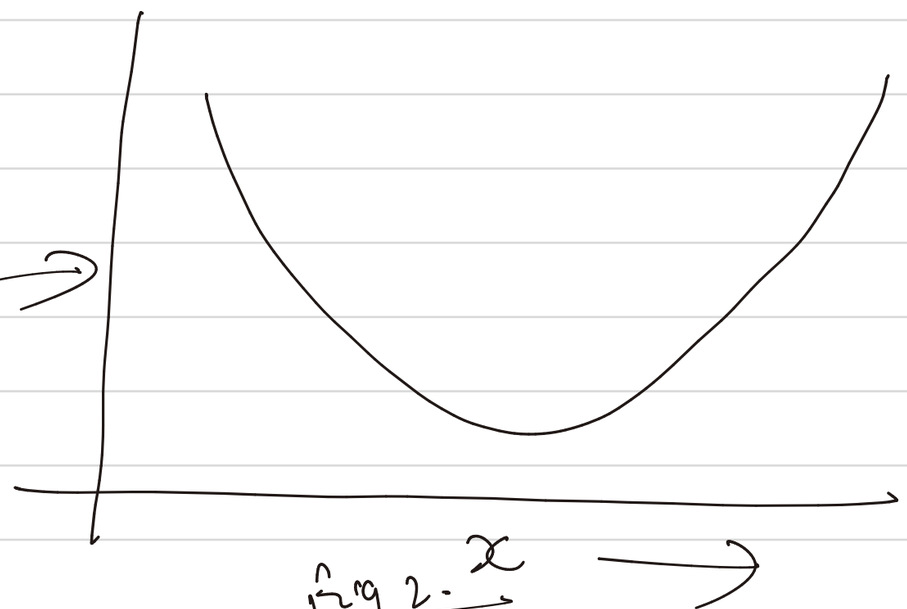
→ Instead of coding hundreds of columns for datasets variable, take average and use that as predictor which will be categorical predictor.

→ Preprocessing tips: always try to plot the datas in different plot types so that we can draw inferences and make visualization of datasets that may help us to find the visual elements, outliers or missing values. If does make sense.

Featurization:



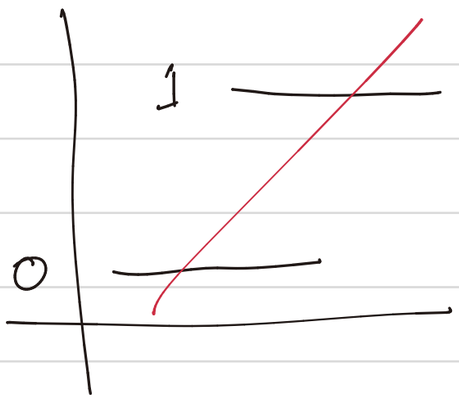
quadratic or Convex Curve.

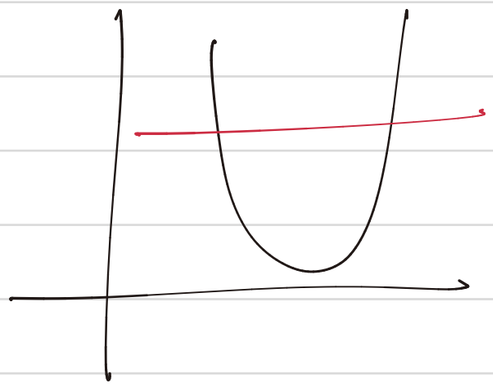


So, what'll be the best fit line Curve in both figure?

So, the featurization is the process of converting data (row data like texts, images, or numbers) into a format numerical features that can be used by machine Learning Algorithms. Transforming the input data into a set of features, which are the characteristics or attributes of the data.

→ Text featurization: Converting words into numbers (words-embeddings)
→ Image featurization: Extracting the features like edges, shapes, color
→ Numerical featurization: Attributes like averages or percentages from raw numbers.





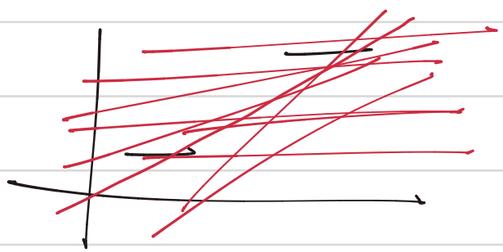
Does it make sense using these best-fit line?

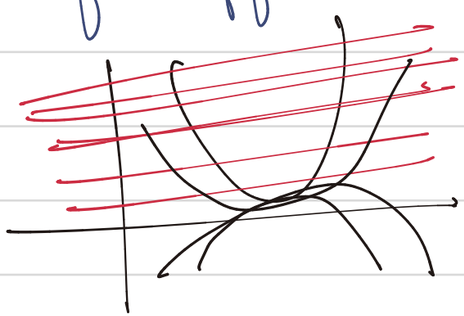
⇒ Not very good

best-fit

→ there is a way to adapt the above ^line which means linear Regression can do well in both situations by introducing new features related to datasets.

Intuition behind this is the decision boundry for classification & Regression problem and also the use of different functions to separate data points.

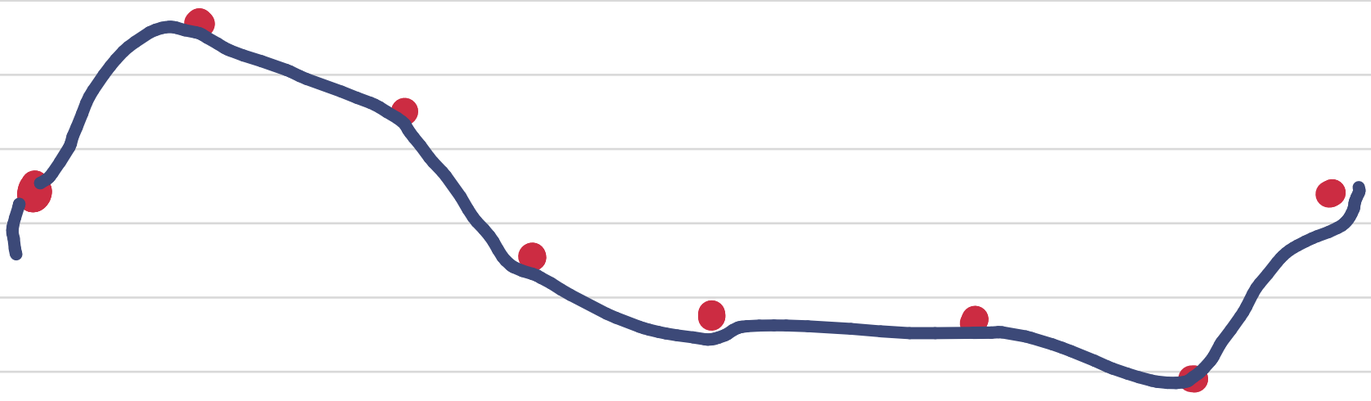




Decision Boundary Based on the f(x).

- Longitudinal Measurements:

Imagine we want to classify a patient's recovery probabilities based on a series of lab test results. We don't want to use the raw measurements we can use the trend, the range, or the "Spikeness" for example.



values over time

t_1	t_2	t_3	t_4	----
1.3	2.4	1.6	2.1	

Fig. The raw measurements associated with a series, Imagine that we need to classify these series into groups of different shapes.

Featurized Series (predictive)

Slope	Range	n_max	n_min .....
-1.2	4	2	2 .....

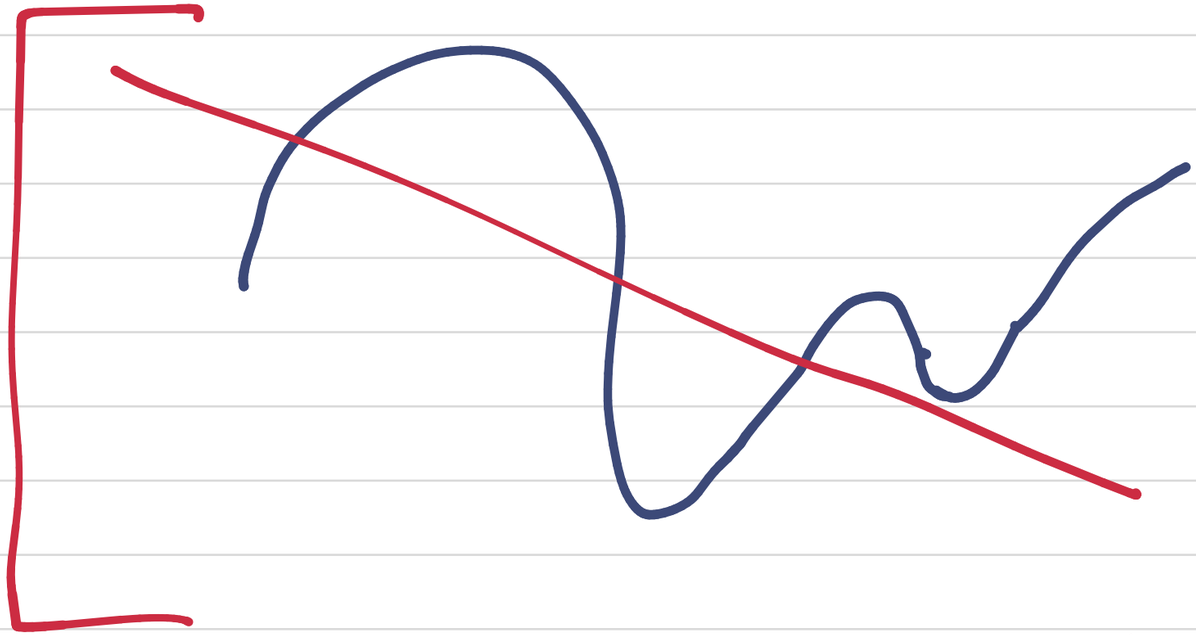


Fig: An example of featurization of a longitudinal series

More featurization.

The difficulty of deriving useful features in image data was one of the original motivations for Deep learning.

By Automatically deriving useful features in image data, deep learning replaced a whole suite of more complicated Image features (eg: Hog):

So, the lesson is that featurization means creating, transforming, manipulating raw input to features such that the Model can adopt & work as expected.

From Textbook on

Knowledge Representation in AI.

"Information Alone is not Knowledge."

(1) Information: data or a sequence of words, details, brief description about Something.

(2) Facts: truths of the world or environment, accumulated by Agent

(3) Knowledge: useful and meaningful information that help AI Agent to make rational and optimal decision, perform tasks/action

(4) Representations: format to represent the facts, perform Some sort of operations and better to approach a clean way to present.

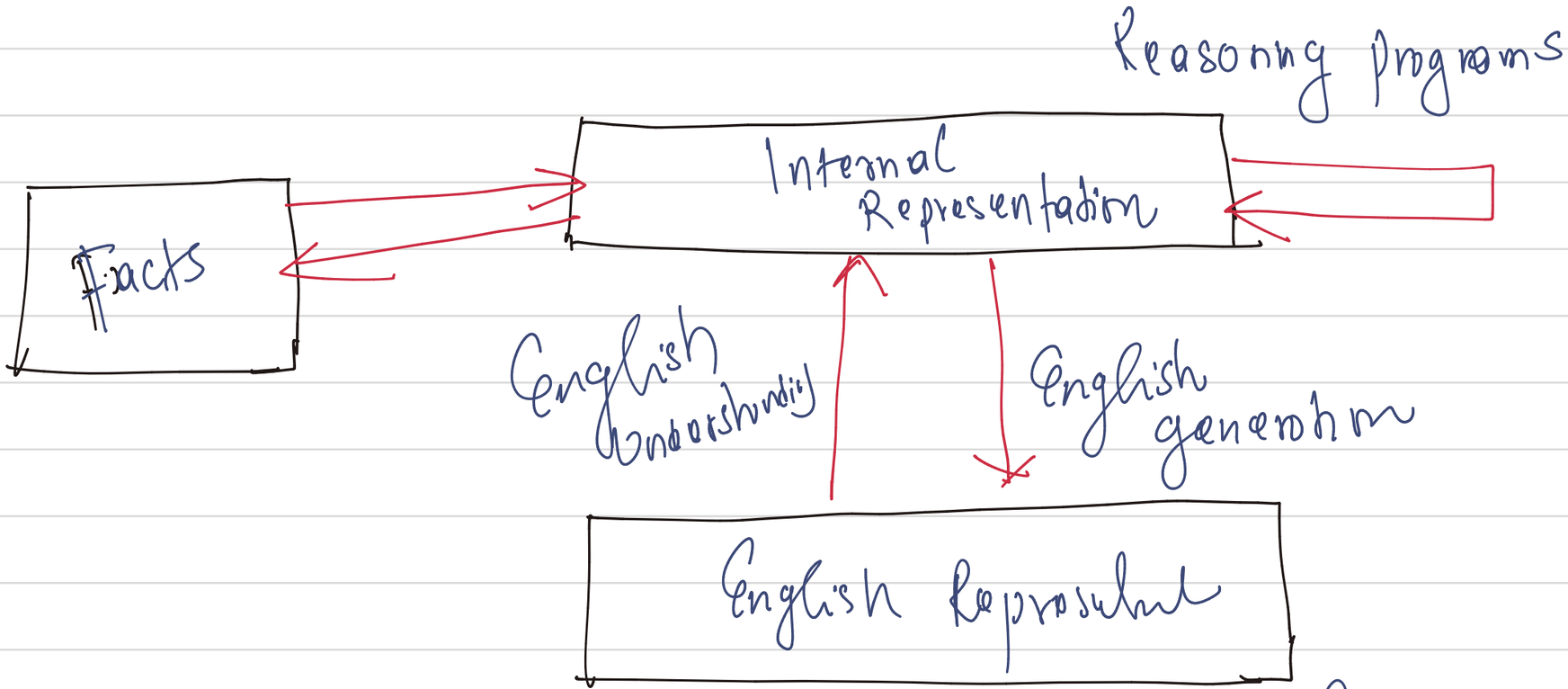


fig. Mappings between facts & representations.

Two types of Mapping Representations

1) forward Mapping Representations – maps from facts to the representation

2) Backward Mapping Representations – maps from representation to the facts.

Propositional Logic:

Propositions: A declarative statement that holds either true or false into a single value.

Sentences → Atomic | Complex.

Atomic Sentence → True | False | P | Q | R ....

I am Computer Scientist. P(x): x is Computer Scientist

Here x can be any entity.

Operators are: ¬ → NOT

∨ ⇒ OR

⇒ Implies

⇐⇒ bidirectional, ∧ - AND

operator precedence,
¬, ∧, ∨, ⇒, ⇐⇒

Complex Sentences have: → (Sentence) | [Sentence]

| ¬Sentence

| Sentence ∧ Sentence

| Sentence ⇒ Sentence

| Sentence ⇐⇒ Sentence

| Sentence ∨ Sentence.

¬ → Not.

⇐⇒ if and only if

⇒ if then it must be true.

(1) ¬ → A dog is not cat.

¬(D∧C)

D =1 Dog

C=1 Cat

∧ → AND

(2) ∨ → A dog or cat is pet-animal where D ⇒ Dog pet animal

C=1 Cat pet animal

(D∨C)

③ ∧ → A dog and cat has tail.

(P∧C)

④ ⇒ (Implications)

If it Rains then the road gets wet.

P → Rains

Q → Road gets wet.

P → Q if P is true Q must be true.

⑤ P ⟷ B → you are citizen of a country. if and only if you are citizen you have right to vote.

or the light is on if and only if the switch is flipped.