Data Analytics MSCSA-134

CIA-3

1)

A) Levenshtein Distance / Algorithm :

$L_{ij}(x, y)$: Edit distance between the first i element of x and the first j elements of y

$$L_{ij} = \begin{cases} i & j = 0 \\ j & i = 0 \\ \min \{ L_{i-1,j} + 1,\ L_{i,j-1} + 1,\ L_{i-1,j-1} + p(x^{(i)}, y^{(i)}) \} \end{cases}$$

where

$$p(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$$

Pseudocode :

```
for s = 1, .... i
    L(s,0) = s

for t = 1, ... j
    L(0,t) = t

for s = 1, ... i
  for t = 1, .... j
    L(s,t) = min { L(s-1,t) + 1,
                   L(s,t-1) + 1,
                   L(s-1,t-1) + p(x^s, y^(t)) }
```

* If the S_1 & S_2 doesn't match.

(1) S & H doesn't match hence.

min (1, 0, 1) + 1 = 1 (0+1)

(2) S & O doesn't match.

min (1, 1, 2) + 1 = 2.

(3) S & R doesn't match.

min (2, 2, 3) + 1 = 3.

(4) S & S match. hence the diagonal element are common

(5) S & i doesn't match.

min (3, 4, 0) + 1 = 4.

therefore after using the algorithm we know that the user need 1 string to the rules with another to obtain the failable output.

Edit distance [verification]

HORSE
↓ ↓ ↓ ↓ ↓
(S) SORSE

∴ edit distance is 1

Using the 3 insertion, deletion & substitution we can obtain the edit distance.

Here we have H → S the Rest remains the same.

S = Substitution

Levenshtein matrix.

		S	O	R	S	E
	0	1	2	3	4	5
H	1	1	2	3	4	5
O	2	2	1	2	3	4
R	3	3	2	1	2	3
S	4	3	3	2	1	2
E	5	4	4	3	2	1

edit distance is (1)

Application of edit distance:

* It's application is that it can be used if it is used in spell check (Auto correct)
* It is used to find the distance between two string's
* It is also used in the dictionary to display all the words in that are near proximity.

max (3, 4, 0) + 1 = 4.

Therefore after using the algorithm we know that the we need 4 string to the rules with zeros to obtain the feasible output.

2) Feature Selection:

Feature selection is the process where we automatically manually select those features / fields which contribute most to your prediction variable or output analysis if we select irrelevant features in the field / ~~fea~~ in the data it reduces the accuracy of the model & make's it irrelevant features.

Importance of FS:

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model & makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen
* It reduces overfitting.

Relevant features - those that we need to perform well.

Irrelevant features - those that are simply unnecessary.

Redundant feature - those that become irrelevant in the presence of others

* 3 main categories of feature selection.
  - → filter methods
  - → wrapper methods
  - → embedded methods

Irrelevant features – those that are simply unnecessary.

Redundant features – those that become irrelevant in the presence of others

* 3 main categories of feature selection.

→ filter method
→ wrapper method
→ embedded method

Conclusion on feature selection:

Potential benefits:

→ To improve accuracy
→ Reduce computation.
→ Reduce space.
→ Reduce cost of future measurements
→ Improved data / model understanding