Assignment 1 - Correlation

Correlation – Correlation is the measure of dependency of one vector upon another. Correlation forms a symmetric matrix with 1s as diagonal elements because of auto-correlation (correlation between similar vectors). The value of correlation is between -1 to 1, 1 suggests that vectors can be plotted on a 2D graph with a line with positive slope while -1 suggests the line has negative slope. Values close to 0 suggests that the vectors are not very dependent. In this assignment, two datasets of biology domain are taken as examples and we have found the correlation between all the vectors.

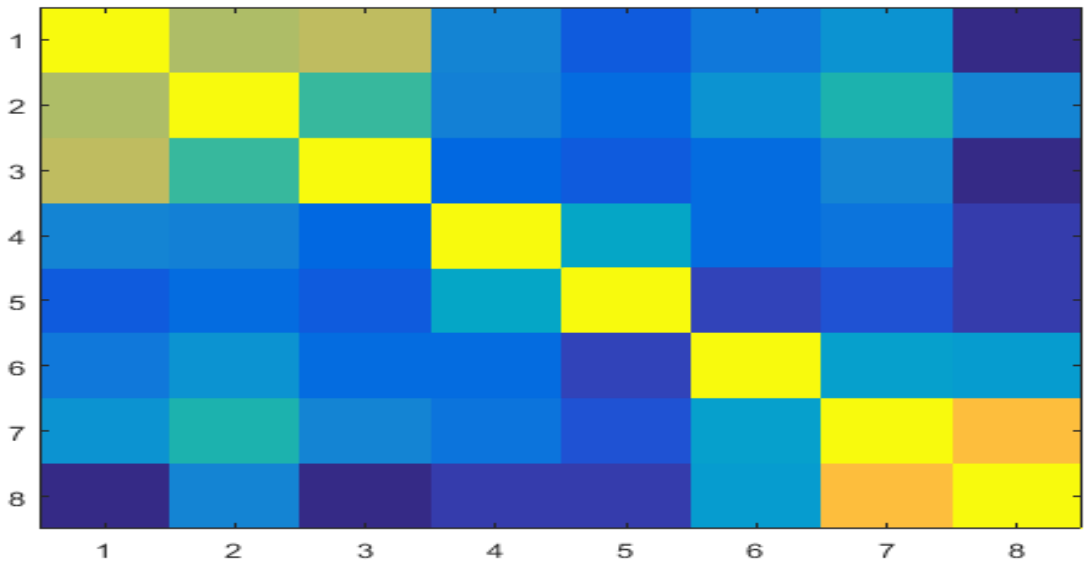
Citation: Dataset 1 – Ecoli (https://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/)

The input features are as follows -:

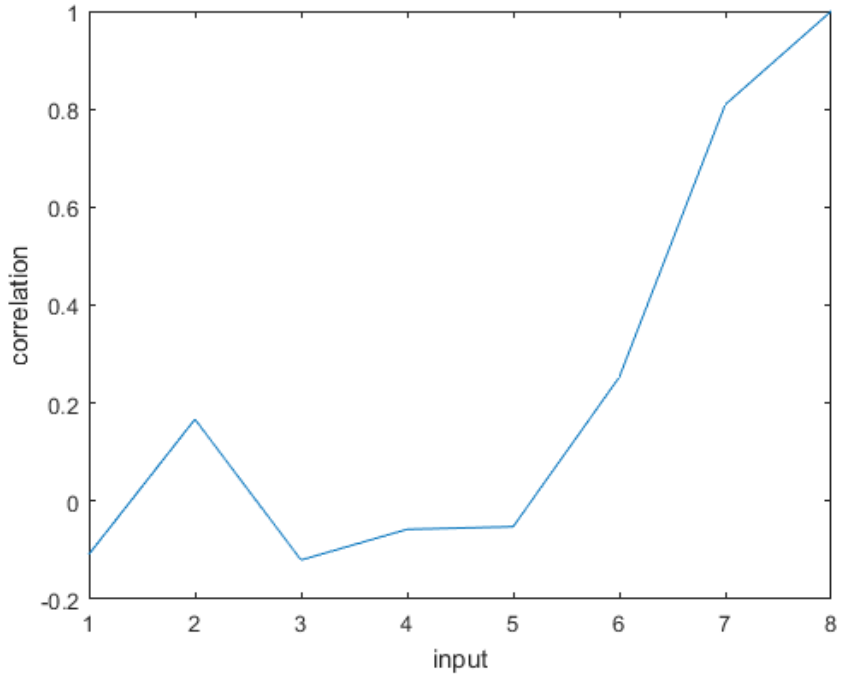
1.Sequence Name: Accession number for the SWISS-PROT database (text)
2. mcg: McGeoch's method for signal sequence recognition. (value ranging 0-1)
3. gvh: von Heijne's method for signal sequence recognition. (value ranging 0-1)
4. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute. (value ranging 0-1)
5. chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute. (value ranging 0-1)
6. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins. (value ranging 0-1)
7. alm1: score of the ALOM membrane spanning region prediction program. (value ranging 0-1)
8. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence. (value ranging 0-1)

The output is as follows –

cp (cytoplasm), im (inner membrane without signal sequence), pp (perisplasm), imU (inner membrane, uncleavable signal sequence), om (outer membrane), omL (outer membrane lipoprotein), imL (inner membrane lipoprotein. Thus, in all there are 7 classes. But to find the correlation, I had to convert each class into a number ranging from 1 to 7. Thus, cp : 1, im :2 and so on.



This above is the image scan of the correlation matrix where the values exactly yellow has correlation of 1 which means they both can be plotted on a linear line. Of course, that would be the case as it's the auto-correlation (correlation between same elements). With the yellow color getting lighter, it shows that they are relatively less positively correlated (correlation coefficient is positive but small) than before. The values in blue are either less correlated or are negatively correlated (dark blue).



The above figure gives the correlation between features and outcomes where last column is outcome. We can see that features X6 and X7 are closely related to Y. And also as Y is linear to Y it gives correlation coefficient as 1.

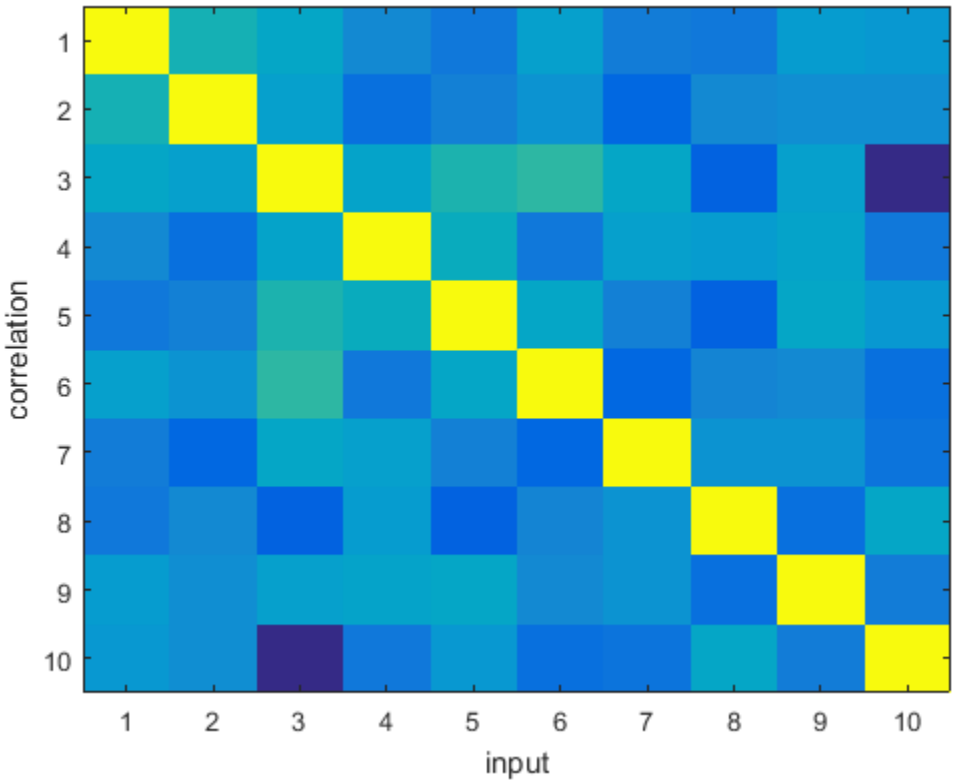
Citation: Dataset 2 – Fertility (https://archive.ics.uci.edu/ml/machine-learning-databases/fertility)

The input data is as follows- :

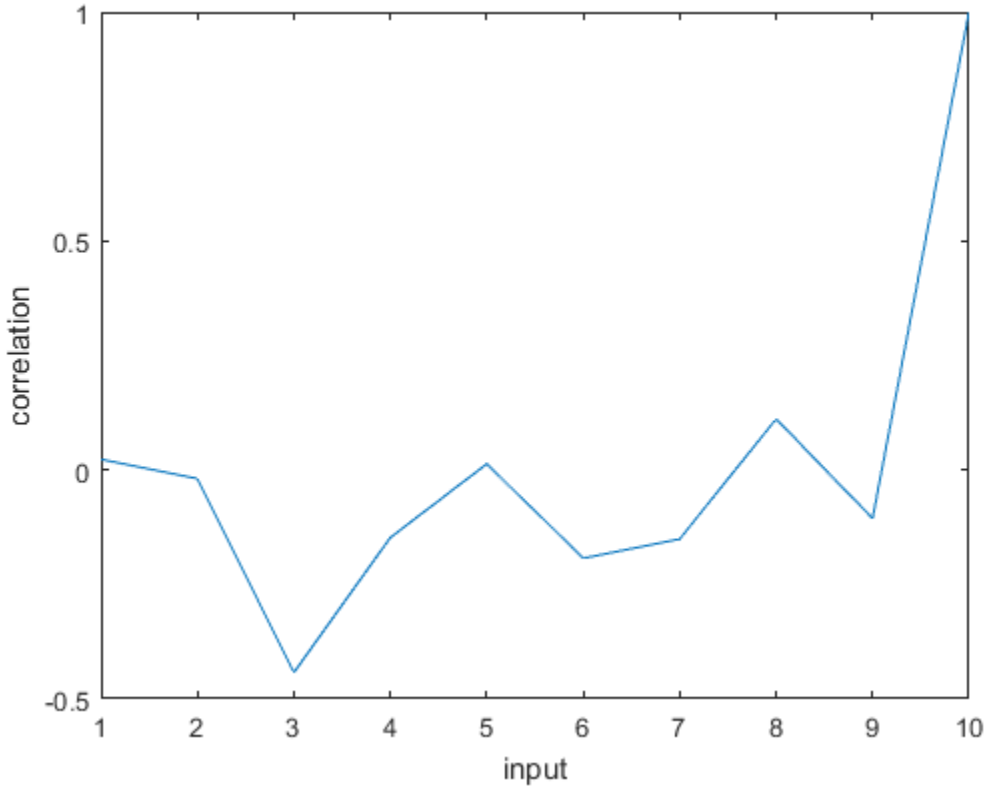
1. Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1) .
2. Age at the time of analysis. 18-36 (0, 1) .
3. Childish diseases (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1).
4. Accident or serious trauma 1) yes, 2) no. (0, 1) .
5. Surgical intervention 1) yes, 2) no. (0, 1) .
6. High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1).
7. Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1) .
8. Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1).
9. Number of hours spent sitting per day ene-16 (0, 1) .

Output is as follows-

Output: Diagnosis normal (N), altered (O). I changed the N to 0 and O to 1 to find the correlation matrix.



This above is the image scan of the correlation matrix where the values exactly yellow has correlation of 1 which means they both can be plotted on a linear line. Ofcourse, that would be the case as it's the auto-correlation (correlation between same elements). We see less positive correlation in this case. Its dependent on the data. It's good that we don't have much correlation. It avoids the problem of over training a curve.



Because we don’t see much correlation in this dataset, we see that values are between -0.5 to 0.5.