Inf2B Coursework Report 1

s1813674

14/04/2020

—

Inf2B - Learning

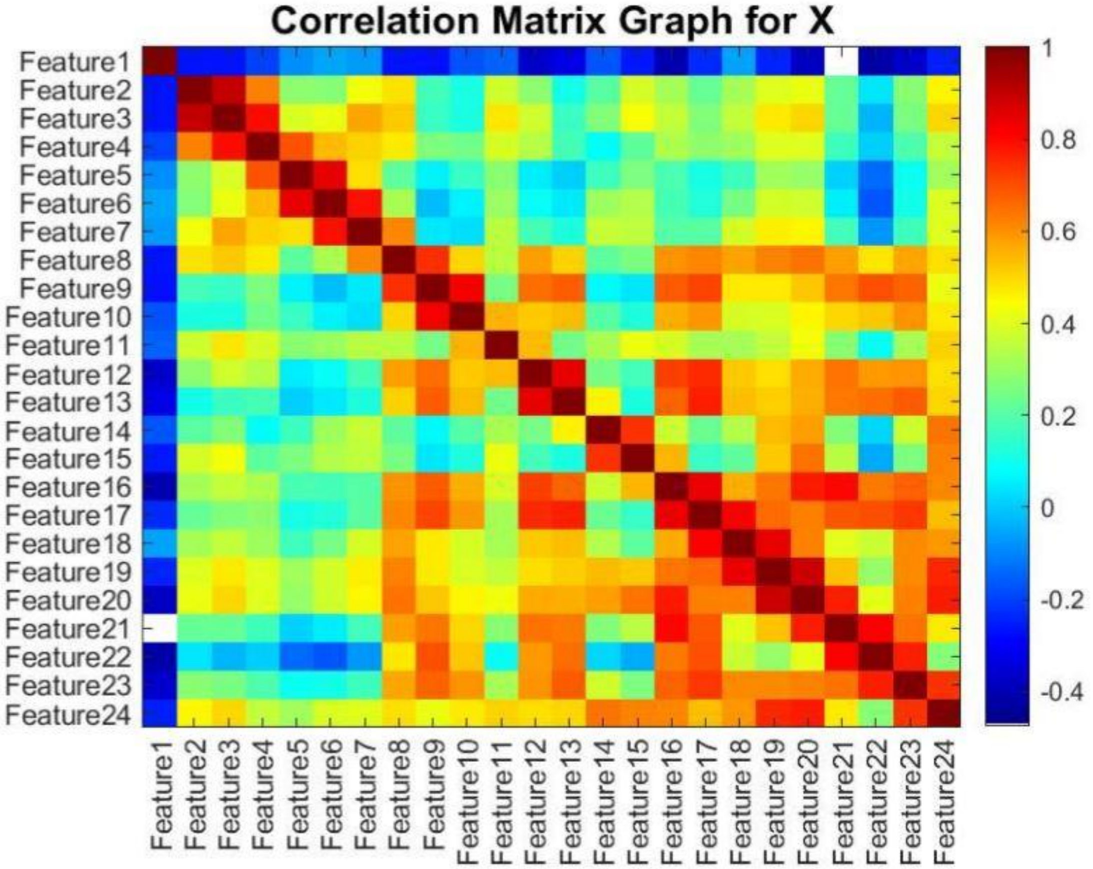
—

Hiroshi Shimodaira

Task 1 – Anuran-Call analysis and classification

Task 1.2

After computing the correlation matrix R, I decided to display the data in the matrix as an image with scaled colors for clarity.

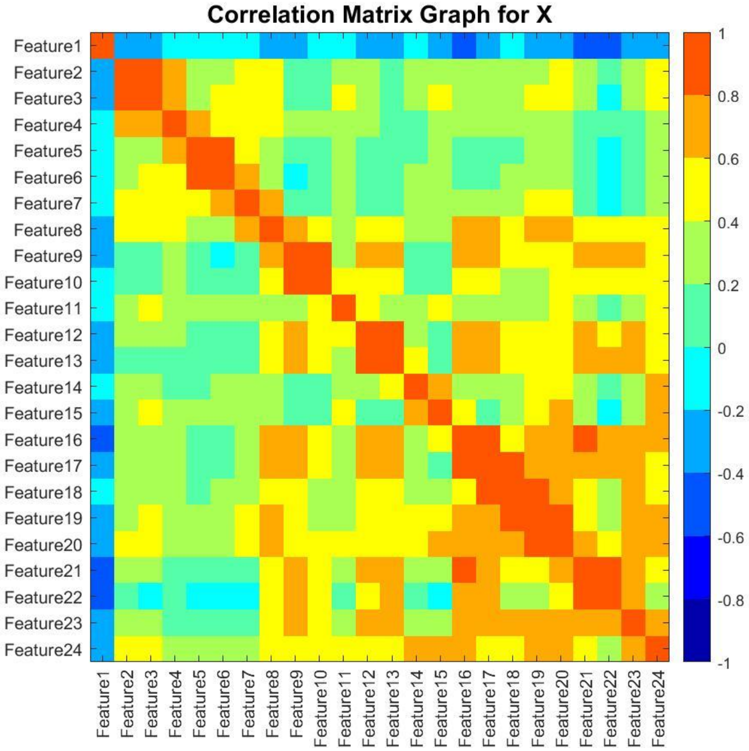


For starters, we will point out that we don't need to consider the upper triangular half of our graph because it's symmetric data. We also don't need to describe the main diagonal since the correlation coefficient will always be 1 since it only shows that each variable always perfectly correlates with itself.
We will be talking about strong/weak negative/positive correlation. A correlation is strong when its absolute value is closer to 1 and weak when it's closer to 0. Since a correlation is a form of dependency, when it's weak a shift in one variable is less likely to affect the other one then if it was strong.

When a correlation is negative, that means that when one variable goes up the other goes down and when it's positive, it means that both variables move in the same direction. It is important to recall that correlation does not imply causation.

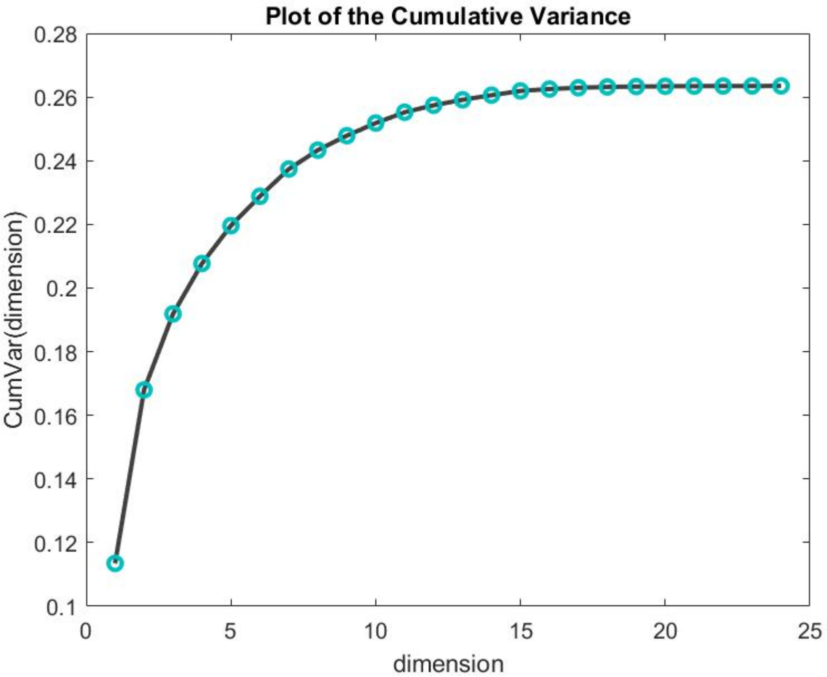
The first noticeable thing in this graph is that the 1st feature has a correlation coefficient smaller than 0 no matter what other feature we compare it with (correlation ranging from -0.4 to 0). It is also the only feature that has no linear relationship with another one (white square) as the correlation between the 1st feature and the 21st feature is 0. But overall, Feature 1 has a weak negative correlation with the other features.

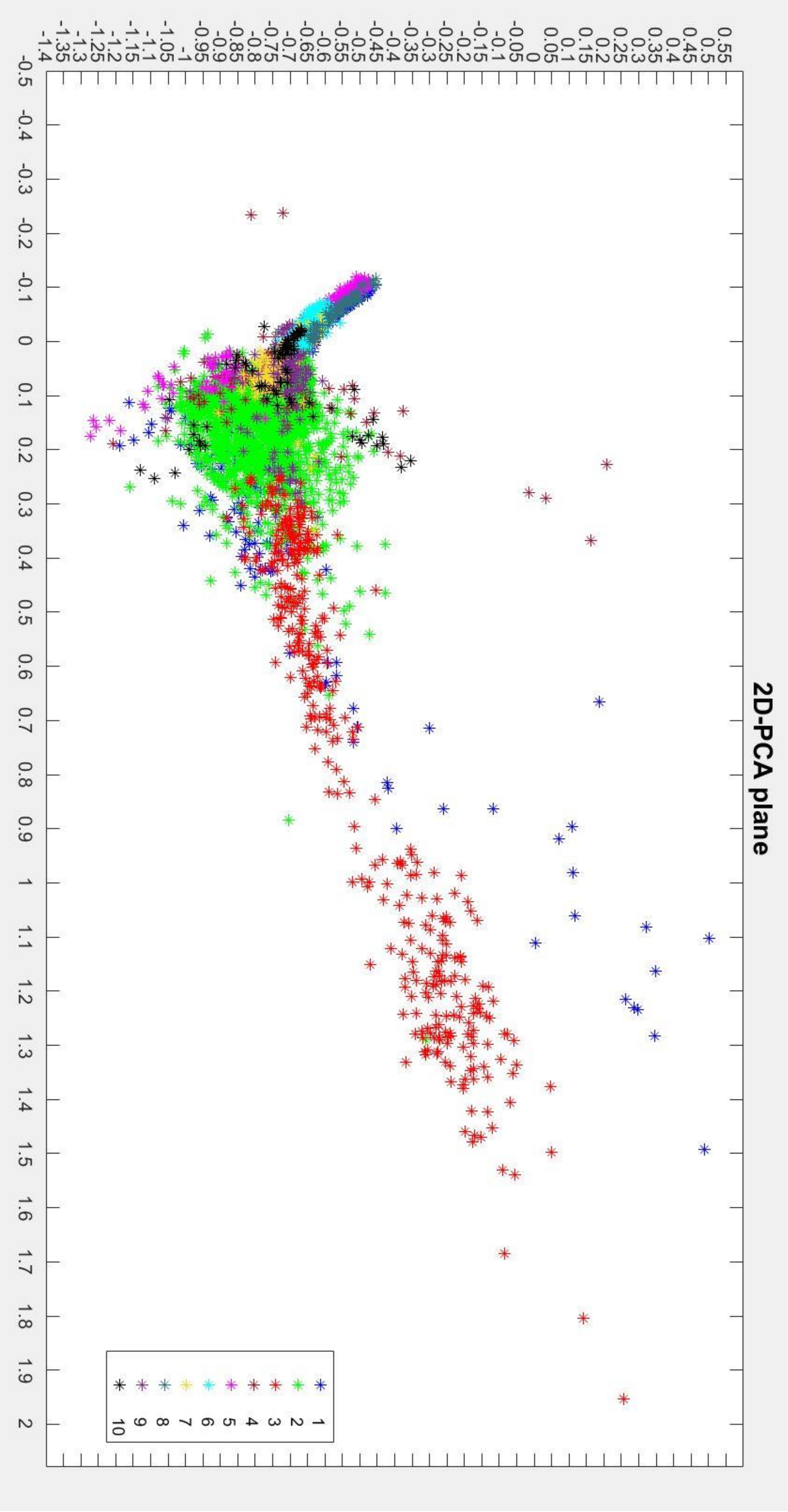
To make the description of the other features easier and now that we have clarified what we mean by weak/strong positive/negative correlation, I've changed the coloring such that it has bigger and clearer intervals:



We define the following intervals: $[\pm 0, \pm 0.2]$ = very weakly correlated (this can almost mean that there are no linear relationship between the two features), $[\pm 0.2, \pm 0.4]$ = weakly correlated, $[\pm 0.4, \pm 0.6]$ = moderately correlated, $[\pm 0.6, \pm 0.8]$ = strongly correlated , $[\pm 0.8, \pm 1]$ = very strongly correlated. The shades of blue say negatively correlated while the other shades mean positively correlated. Thanks to this simplified graph, we can easily see the correlation between the 24 features.

Task 1.3




2D-PCA plane
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
-0.6
-0.65
-0.7
-0.75
-0.8
-0.85
-0.9
-0.95
-1
-1.05
-1.1
-1.15
-1.2
-1.25
-1.3
-1.35
-1.4
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
* 1
* 2
* 3
* 4
* 5
* 6
* 7
* 8
* 9
* 10

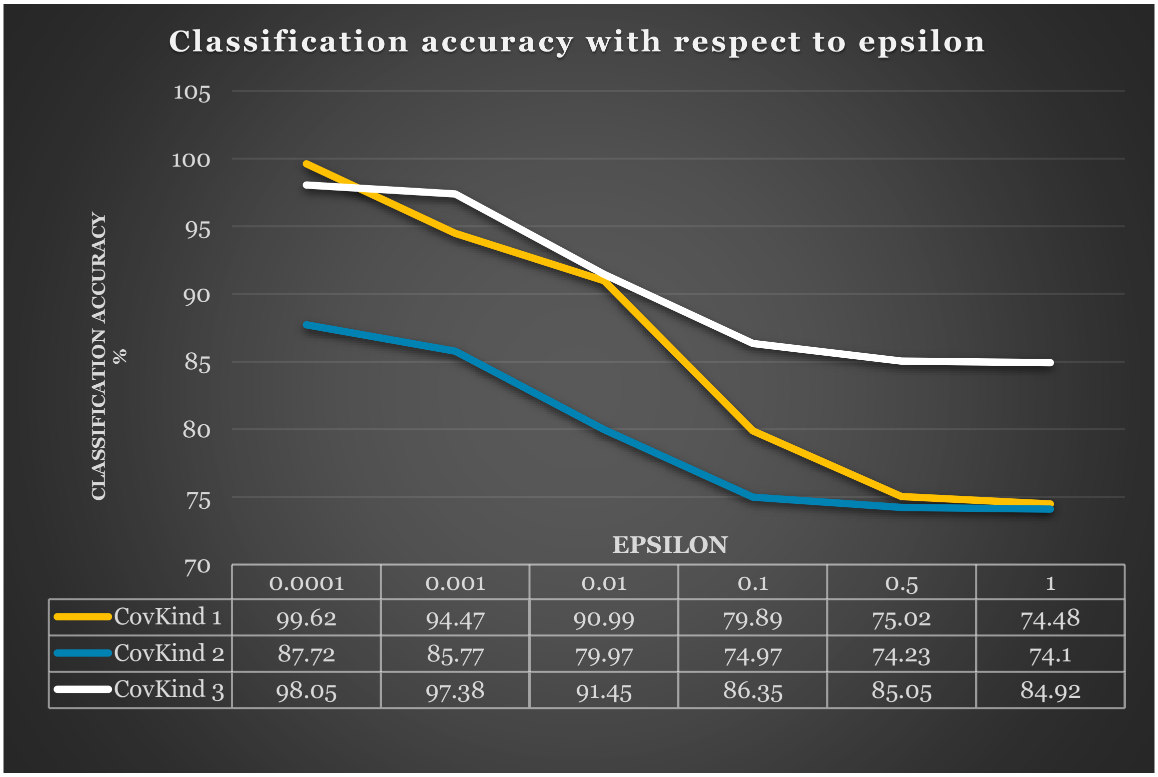
Task 1.4

Since the number of correctly classified patterns is obtained by summing the numbers on the leading diagonal. I got the accuracy (correct classification rate) by summing the leading diagonal of the final confusion matrix (where each element is a relative frequency).

Classification rate table depending Epsilon, KFolds and CovKind

Epsilon	KFolds	CovKind	Accuracy
0.01	5	1	*90.99 %*
0.01	5	2	*79.97 %*
0.01	5	3	*91.45 %*

Task 1.5



	0.0001	0.001	0.01	0.1	0.5	1
CovKind 1	99.62	94.47	90.99	79.89	75.02	74.48
CovKind 2	87.72	85.77	79.97	74.97	74.23	74.1
CovKind 3	98.05	97.38	91.45	86.35	85.05	84.92

All covariance kinds have decreasing accuracy when epsilon is large and come closer to a perfect accuracy when epsilon is really small (see 0.0001). The Full Covariance kind (CovKind 1) is the one that gets closest to perfect accuracy but it is really close to the Shared Covariance Kind (CovKind 3) whereas the Diagonal Covariance Kind (CovKind 2) always has the lowest accuracy (which is probably coming from the assumption that all elements of the data set are uncorrelated, making everything but the diagonal equal to 0). The Full Covariance kind is also the one that decreases the fastest when compared to the other kinds, it is the most responsive to the change in epsilon. After the value 0.5, it looks like the accuracy decreases way more slowly however it remains relatively high for the Shared Covariance Kind (10% loss in accuracy) whereas it's around 75% for the other kinds (25% loss for CovKind1 and 10% for CovKind2). Another observation would be that both CovKind 2 and 3 have the same changes in variation although CovKind 3 has a much higher accuracy. Hence, I believe epsilon needs to be really small in order to get close to perfect accuracy.