**ASSIGNMENT 3**

Dataset Used:

Ionosphere Dataset:<https://archive.ics.uci.edu/ml/datasets/ionosphere>

* Number of attributes:34
* Number of classes:2
* Number of samples:351
* Train Sample: 50 class A and 50 class B
* Test Sample Size:100
* Total Samples Used :200(100 Test and 100 Train)

I removed column number 1 from every sample and splitted dataset to

Classifier used :LIBSVM module in python with RBG Kernel and C=10

PCA:

+----------------------+---------------------+

| number of components | Variance Captured |

+----------------------+---------------------+

| 1 | 0.23646694012856623 |

| 2 | 0.4004088350189322 |

| 3 | 0.48844572216407456 |

| 4 | 0.547543780879687 |

| 5 | 0.5950936993041743 |

| 6 | 0.6387562673412611 |

| 7 | 0.6768471398141308 |

| 8 | 0.7140982117237195 |

| 9 | 0.7490619073628945 |

| 10 | 0.7799094263537029 |

| 11 | 0.8077252019618576 |

| 12 | 0.832882227207747 |

| 13 | 0.8545066516783146 |

| 14 | 0.8732087896441999 |

| 15 | 0.8910094404700325 |

| 16 | 0.9075690213223373 |

| 17 | 0.9216096622995978 |

| 18 | 0.9342597303458219 |

| 19 | 0.9440964053277263 |

| 20 | 0.9528195576378358 |

| 21 | 0.953544431878194 |

| 22 | 0.9545654570750491 |

| 23 | 0.955983186101519 |

| 24 | 0.9577164460096201 |

| 25 | 0.9597643125333101 |

| 26 | 0.9621460094176243 |

| 27 | 0.9650849315480167 |

| 28 | 0.9688091818636048 |

| 29 | 0.9763136038763311 |

| 30 | 0.9831206775102095 |

| 31 | 0.9882657103871797 |

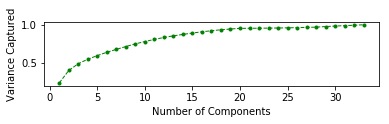
| 32 | 0.9937846971463086 |

| 33 | 1.0 |

+----------------------+---------------------+

(made using Pretty table module)

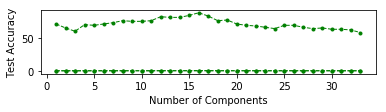
Graph between Variance captured and Number of Components:



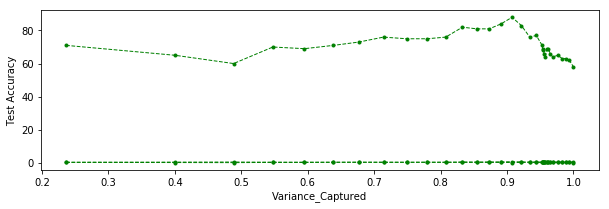
Number of components needed for :

1. 75% variance: 9
2. 85% variance:13
3. 95% variance:18

Graph between Test Accuracy vs Number of components:



Graph between Test Accuracy vs Variance Captured:



+----------------------+------------------------------------------------+

| number of components | Test Accuracy |

+----------------------+------------------------------------------------+

| 1 | (71.0, 0.29, 0.1891891891891892) |

| 2 | (65.0, 0.35, 0.10176390773405698) |

| 3 | (60.0, 0.4, 0.047619047619047616) |

| 4 | (70.0, 0.3, 0.1736111111111111) |

| 5 | (69.0, 0.31, 0.15486915486915487) |

| 6 | (71.0, 0.29, 0.2352) |

| 7 | (73.0, 0.27, 0.28213333333333335) |

| 8 | (76.0, 0.24, 0.2869269949066214) |

| 9 | (75.0, 0.25, 0.26271542664985287) |

| 10 | (75.0, 0.25, 0.292192613370734) |

| 11 | (76.0, 0.24, 0.27435064935064934) |

| 12 | (82.0, 0.18, 0.4563279857397504) |

| 13 | (81.0, 0.19, 0.41226941226941227) |

| 14 | (81.0, 0.19, 0.4492753623188406) |

| 15 | (84.0, 0.16, 0.5017361111111112) |

| 16 | (88.0, 0.12, 0.6016666666666667) |

| 17 | (83.0, 0.17, 0.4786813186813187) |

| 18 | (76.0, 0.24, 0.35135135135135137) |

| 19 | (77.0, 0.23, 0.3698630136986301) |

| 20 | (71.0, 0.29, 0.26582278481012656) |

| 21 | (69.0, 0.31, 0.2345679012345679) |

| 22 | (68.0, 0.32, 0.21951219512195122) |

| 23 | (66.0, 0.34, 0.19047619047619047) |

| 24 | (64.0, 0.36, 0.16279069767441862) |

| 25 | (69.0, 0.31, 0.2345679012345679) |

| 26 | (69.0, 0.31, 0.2345679012345679) |

| 27 | (66.0, 0.34, 0.19047619047619047) |

| 28 | (64.0, 0.36, 0.16279069767441862) |

| 29 | (65.0, 0.35, 0.17647058823529413) |

| 30 | (63.0, 0.37, 0.14942528735632185) |

| 31 | (63.0, 0.37, 0.14942528735632185) |

| 32 | (62.0, 0.38, 0.13636363636363635) |

| 33 | (57.99999999999999, 0.42, 0.08695652173913043) |

+----------------------+------------------------------------------------+

(1st value is the accuracy)

Accuracy for 9 components:75%

Accuracy for 13 components:81%

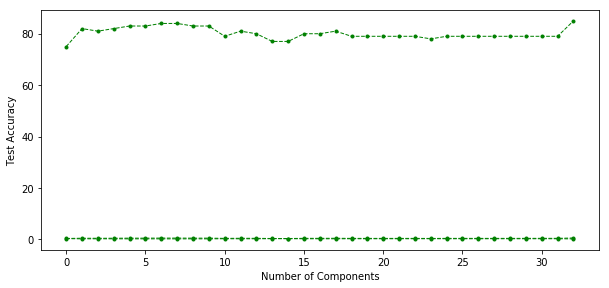
Accuracy for 18 components:77%

Intuition on PCA:

* The accuracy increased with number of components upto the point where almost 92% variance was captured and then started o decrease.
* Initial increase is intuitive as more information will be contained and hence more accuracy
* But the decrease after that could be attributed to increase in VC dimension due to which the probability of converging to ideal classifier decreases.
* Maximum accuracy was obtained for number of components=16(88%)

LDA:

Graph of Test accuracy vs number of components:



+----------------------+-----------------------------------+

| number of components | Test Accuracy |

+----------------------+-----------------------------------+

| 1 | (75.0, 0.25, 0.31709791983764585) |

| 2 | (82.0, 0.18, 0.41025641025641024) |

| 3 | (81.0, 0.19, 0.38455382152861145) |

| 4 | (82.0, 0.18, 0.4096) |

| 5 | (83.0, 0.17, 0.437173825772782) |

| 6 | (83.0, 0.17, 0.44) |

| 7 | (84.0, 0.16, 0.46537842190016104) |

| 8 | (84.0, 0.16, 0.46537842190016104) |

| 9 | (83.0, 0.17, 0.44) |

| 10 | (83.0, 0.17, 0.44430844553243576) |

| 11 | (79.0, 0.21, 0.35350987810004203) |

| 12 | (81.0, 0.19, 0.39208486332109344) |

| 13 | (80.0, 0.2, 0.3694581280788177) |

| 14 | (77.0, 0.23, 0.3064312736443884) |

| 15 | (77.0, 0.23, 0.3013642000826788) |

| 16 | (80.0, 0.2, 0.3652597402597403) |

| 17 | (80.0, 0.2, 0.3694581280788177) |

| 18 | (81.0, 0.19, 0.39208486332109344) |

| 19 | (79.0, 0.21, 0.3476643241008681) |

| 20 | (79.0, 0.21, 0.3476643241008681) |

| 21 | (79.0, 0.21, 0.3476643241008681) |

| 22 | (79.0, 0.21, 0.3476643241008681) |

| 23 | (79.0, 0.21, 0.3476643241008681) |

| 24 | (78.0, 0.22, 0.32666666666666666) |

| 25 | (79.0, 0.21, 0.35350987810004203) |

| 26 | (79.0, 0.21, 0.35350987810004203) |

| 27 | (79.0, 0.21, 0.35350987810004203) |

| 28 | (79.0, 0.21, 0.35350987810004203) |

| 29 | (79.0, 0.21, 0.35350987810004203) |

| 30 | (79.0, 0.21, 0.3607893607893608) |

| 31 | (79.0, 0.21, 0.36967032967032964) |

| 32 | (79.0, 0.21, 0.36967032967032964) |

| 33 | (85.0, 0.15, 0.49979600163198695) |

+----------------------+-----------------------------------+

* Accuracy for 1 component:75%
* Accuracy for 9 components:83%
* Accuracy for 13 components:80%
* Accuracy for 18 components:81%

Table for variance Captured and Accuracy:

+--------------------+-----------------------------------+

| Variance Captured | Test Accuracy |

+--------------------+-----------------------------------+

| 0.9999999999999974 | (75.0, 0.25, 0.31709791983764585) |

| 0.9999999999999977 | (82.0, 0.18, 0.41025641025641024) |

| 0.9999999999999978 | (81.0, 0.19, 0.38455382152861145) |

| 0.9999999999999982 | (82.0, 0.18, 0.4096) |

| 0.9999999999999986 | (83.0, 0.17, 0.437173825772782) |

| 0.9999999999999988 | (83.0, 0.17, 0.44) |

| 0.9999999999999989 | (84.0, 0.16, 0.46537842190016104) |

| 0.9999999999999991 | (84.0, 0.16, 0.46537842190016104) |

| 0.9999999999999992 | (83.0, 0.17, 0.44) |

| 0.9999999999999992 | (83.0, 0.17, 0.44430844553243576) |

| 0.9999999999999992 | (79.0, 0.21, 0.35350987810004203) |

| 0.9999999999999992 | (81.0, 0.19, 0.39208486332109344) |

| 0.9999999999999992 | (80.0, 0.2, 0.3694581280788177) |

| 0.9999999999999994 | (77.0, 0.23, 0.3064312736443884) |

| 0.9999999999999994 | (77.0, 0.23, 0.3013642000826788) |

| 0.9999999999999998 | (80.0, 0.2, 0.3652597402597403) |

| 0.9999999999999998 | (80.0, 0.2, 0.3694581280788177) |

| 0.9999999999999998 | (81.0, 0.19, 0.39208486332109344) |

| 0.9999999999999998 | (79.0, 0.21, 0.3476643241008681) |

| 0.9999999999999998 | (79.0, 0.21, 0.3476643241008681) |

| 0.9999999999999998 | (79.0, 0.21, 0.3476643241008681) |

| 0.9999999999999998 | (79.0, 0.21, 0.3476643241008681) |

| 0.9999999999999998 | (79.0, 0.21, 0.3476643241008681) |

| 1.0 | (78.0, 0.22, 0.32666666666666666) |

| 1.0 | (79.0, 0.21, 0.35350987810004203) |

| 1.0 | (79.0, 0.21, 0.35350987810004203) |

| 1.0 | (79.0, 0.21, 0.35350987810004203) |

| 1.0 | (79.0, 0.21, 0.35350987810004203) |

| 1.0 | (79.0, 0.21, 0.35350987810004203) |

| 1.0 | (79.0, 0.21, 0.3607893607893608) |

| 1.0 | (79.0, 0.21, 0.36967032967032964) |

| 1.0 | (79.0, 0.21, 0.36967032967032964) |

| 1.0 | (85.0, 0.15, 0.49979600163198695) |

+--------------------+-----------------------------------+

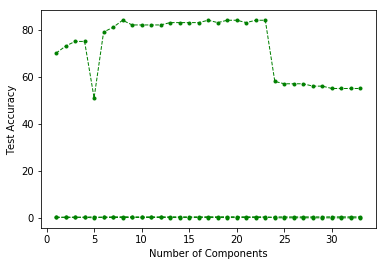
Intuition:

* Even for projection on 1 component accuracy of 75% was obtained showing that most of variance was captured by this component(.9999).
* So the accuracy almost remained constant throughout even with increase in number of components no appreciable change was observed.
* Better accuracy compared with linear pca for same number of components.

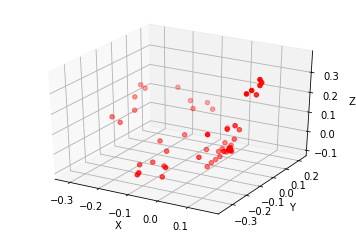
Kernel PCA:

Number of components needed for 80% variance: 6

Graph between Test Accuracy vs Number of components:



3-D projection on 3 components:(Accuracy obtained: 74%)



Number of points in graph above=100

Intuition:

* At number of components=5 there was an unexpected dip in accuracy(I don’t know why ) this happened as this happened only at this point)
* Accuracy with 3 components in kernel pca was 74% and with linear pca 60% so clearly more variance is captured in kernel pca in very few number of components compared to linear pca.

Sangner’s rule(summation to n ):

+----------------------+-------------------------------------------------+

| number of components | Test Accuracy |

+----------------------+-------------------------------------------------+

| 1 | (52.0, 0.48, 0.02040816326530612) |

| 2 | (59.0, 0.41, 0.07161803713527852) |

| 3 | (65.0, 0.35, 0.09652509652509653) |

| 4 | (60.0, 0.4, 0.0744047619047619) |

| 5 | (59.0, 0.41, 0.06352941176470588) |

| 6 | (57.99999999999999, 0.42, 0.053156146179401995) |

| 7 | (61.0, 0.39, 0.07293550331525014) |

| 8 | (62.0, 0.38, 0.07484407484407485) |

| 9 | (54.0, 0.46, 0.021739130434782608) |

| 10 | (61.0, 0.39, 0.06453333333333333) |

| 11 | (66.0, 0.34, 0.17344173441734417) |

| 12 | (61.0, 0.39, 0.08575478384124734) |

| 13 | (63.0, 0.37, 0.07643600180913614) |

| 14 | (62.0, 0.38, 0.0661764705882353) |

| 15 | (70.0, 0.3, 0.16233766233766234) |

| 16 | (61.0, 0.39, 0.048574869530309116) |

| 17 | (60.0, 0.4, 0.1111111111111111) |

| 18 | (56.00000000000001, 0.44, 0.024390243902439025) |

| 19 | (66.0, 0.34, 0.10305958132045089) |

| 20 | (69.0, 0.31, 0.14728682170542637) |

| 21 | (63.0, 0.37, 0.07643600180913614) |

| 22 | (63.0, 0.37, 0.09013333333333333) |

| 23 | (75.0, 0.25, 0.2501000400160064) |

| 24 | (61.0, 0.39, 0.10698496905393456) |

| 25 | (60.0, 0.4, 0.0625) |

| 26 | (66.0, 0.34, 0.10865874363327674) |

| 27 | (68.0, 0.32, 0.1330049261083744) |

| 28 | (65.0, 0.35, 0.0989010989010989) |

| 29 | (72.0, 0.28, 0.1936) |

| 30 | (60.0, 0.4, 0.05482456140350877) |

| 31 | (64.0, 0.36, 0.08506944444444445) |

| 32 | (67.0, 0.33, 0.1194708557255064) |

| 33 | (67.0, 0.33, 0.12398112398112399) |

+----------------------+-------------------------------------------------+

+----------------------+-------------------------------------------------+

| number of components | Test Accuracy |

+----------------------+-------------------------------------------------+

| 1 | (51.0, 0.49, 0.0015360983102918587) |

| 2 | (56.99999999999999, 0.43, 0.029535864978902954) |

| 3 | (57.99999999999999, 0.42, 0.06060606060606061) |

| 4 | (60.0, 0.4, 0.05827505827505827) |

| 5 | (67.0, 0.33, 0.11601766358892011) |

| 6 | (50.0, 0.5, nan) |

| 7 | (65.0, 0.35, 0.09301364200082679) |

| 8 | (66.0, 0.34, 0.10509031198686371) |

| 9 | (62.0, 0.38, 0.057692307692307696) |

| 10 | (62.0, 0.38, 0.0975609756097561) |

| 11 | (71.0, 0.29, 0.17703733440385389) |

| 12 | (65.0, 0.35, 0.12) |

| 13 | (56.99999999999999, 0.43, 0.02766798418972332) |

| 14 | (68.0, 0.32, 0.1330049261083744) |

| 15 | (59.0, 0.41, 0.06352941176470588) |

| 16 | (59.0, 0.41, 0.08273748723186926) |

| 17 | (63.0, 0.37, 0.07643600180913614) |

| 18 | (61.0, 0.39, 0.10698496905393456) |

| 19 | (69.0, 0.31, 0.14585858585858585) |

| 20 | (70.0, 0.3, 0.16420361247947454) |

| 21 | (71.0, 0.29, 0.17703733440385389) |

| 22 | (67.0, 0.33, 0.12147961328289197) |

| 23 | (71.0, 0.29, 0.17703733440385389) |

| 24 | (57.99999999999999, 0.42, 0.04) |

| 25 | (70.0, 0.3, 0.1610305958132045) |

| 26 | (70.0, 0.3, 0.16420361247947454) |

| 27 | (65.0, 0.35, 0.0989010989010989) |

| 28 | (66.0, 0.34, 0.10666666666666667) |

| 29 | (65.0, 0.35, 0.09652509652509653) |

| 30 | (68.0, 0.32, 0.1314935064935065) |

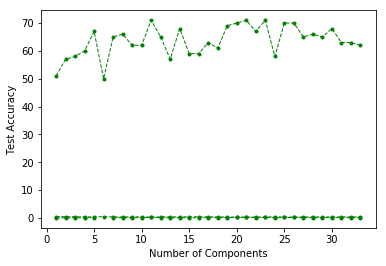
| 31 | (63.0, 0.37, 0.08574327752409944) |

| 32 | (63.0, 0.37, 0.07428571428571429) |

| 33 | (62.0, 0.38, 0.07484407484407485) |

+----------------------+-------------------------------------------------+

The two tables are the accuracies when the same code was run twice.It is clear that it converges to a different vector each time it is run



Variance: I didn’t understand what is meant by this as we can find only the weights in sanger’s rule(for each dimension of w we have to run it in a loop).

Conclusion:

Linear PCA works well and was able to capture 90% variance in 13 components and hence can be used to reduce the dimension of the dataset to 13,even the accuracy reduced when we increased the number of components above 15.LDA seems to work better as it classifies considering the direction in which variance between classes is maximized so here also LDA was found to have better accuracy than PCA for the same number of components.Kernel PCA seems to work better than LDA above a certain number of components(got the highest accuracy of 88%) and obviously better than linear PCA.Oja’s Rule was converging at different vectors in each time you run the code since ‘w’ is initialized randomly and due to the fact that we are taking summation to n(n=number of components).

HARI KRISHNAN CK

2016PH10562