**Fundamentals of Data Science**  
Dimensionality reduction – solutions

Please import (in any program) a set of seeds (google: seeds dataset). This collection represents measurements of several geometrical features of wheat grains - made using X-ray techniques. The last column (values: 1, 2, 3) means the grade of the tested grain. The purpose of this data collection was to verify whether the grain grade can be determined based on the measurement of geometrical features. Unfortunately, the dimensionality of the problem does not allow for its direct analysis.

The course of the task:

1. Please reduce the data dimension to N = 2 using the PCA method. For example, you can use princomp from the Octave package or similar MATLAB function (<http://octave.sourceforge.net/statistics/function/princomp.html)>.

or Python’s sklearn PCA:

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

Note: you must transform the data using the function, then use the first two columns from the score matrix. What are the elements of the latent array?

latent = [10.7933269196671

2.12945511629438

0.0736300329917388

0.0128874947173724

0.00274822667896525

0.00157044979619203

2.96554425025795e-05]

this column vector represents variances calculated along principal components Observe that a lot of total variance is carried with the first two principal components (98% of total variance – to be more precise). It means that it is rational to reduce the dimensionality to 2.

Two dimensional representation (first fifty elements):

[0.663448375781879 -1.41732097557430

0.315666511700423 -2.68922914731690

-0.660499301919611 -1.13150634993143

-1.05527590285231 -1.62119001610855

1.61999921160609 -2.18338442279818

-0.476938007349108 -1.33649436837886

-0.184834720128167 -0.150364411169258

-0.780629616369676 -1.12979882899776

2.28210810225297 -1.36001690130754

1.97854147173832 -1.49468793156681

0.369122946971606 0.886722511461662

-0.711021199508623 -2.10663729852700

-1.21370534732638 0.0946878939133256

-1.16908540571592 -0.742962899015953

-1.19272175563192 -0.953268161727208

-0.508171207012589 0.377958423717976

-1.37469698366276 1.32290558924087

1.05726437573518 -2.01562874734647

-0.150961096956511 -2.02235812659075

-2.46241292804403 0.0737473834990606

-0.631332099877758 -0.718305655107885

-0.689698659667183 -1.11182530515728

1.40769071969915 -2.80658086464545

-2.84267672185914 -2.66880642430202

0.433268215365562 -1.88984464325069

1.81289158323722 -2.60002176250269

-2.02131332039651 -0.608743327724697

-2.19571862287569 -1.49837621604982

-0.744468841188791 -1.06518720842650

-1.50350479946733 -0.368206744610601

-1.52075320185337 -3.06180224858558

0.761190255505459 -0.209488758885235

-0.767738427883690 0.126295451462823

-0.823965933004519 -1.70715019926195

0.439542396447793 -1.52858534253332

1.52205297775685 -1.25609762373484

1.65240524733289 -0.675119440298280

2.47674444868431 -0.451537548389501

0.0115750672757972 -0.596250977233270

-1.11443821870908 2.83345206391614

-1.37160169775339 -1.30108590732987

-1.36349512982823 -1.63960123610541

-1.88302953603366 -1.51985065716355

0.629560566051947 1.10062048000236

0.284412123551472 -0.560641213226648

-0.960044752877241 -2.29723525601891

0.818964617437342 -2.26570275514883

0.196621301164388 -0.740666486130689

0.0153276057361330 -1.02061145814630

0.254235169295421 -1.55022082003377 …

1. Please plot the data in two-dimensional space by coloring the grain species. In the case of Octave/MATLAB, you can use gscatter (requires two dimensions of the file and then a label - allowing the coloring of its elements). For Python use Matplotlib scatter plot (<https://kite.com/python/answers/how-to-color-a-scatter-plot-by-category-using-matplotlib-in-python>). Do such classes are distinguishable?



As you can see classes are distinguishable – it means that dimensionality reduction preserves the informative content of the dataset.