SOFTWARE ENGINEERING PROGRAMME
UNIVERSITY OF OXFORD
www.softeng.ox.ac.uk

ASSESSMENT

Student: Nicholas Drake

Course: Classical Machine Learning

Date: 9th September 2024

Grade: 74

Rob Collins

January 6, 2025

REPORT

You start your analysis by importing the data into Pandas and applying various summary functions to get a view of the data. You notice quickly that some of the features include null values and objects. Some of these objects contain a string that represents an error. You correctly apply the missingno library to visual visualise missing data. Of course, this would not indicate places where there are error values. You make a reasonable judgement about deleting the rows containing errors. It does indeed represent a small proportion of the overall total amount of data. You also apply imputation to the response variables and this demonstrates that you understand this process.

You use a seaborne KDE plot to show the distribution of each feature. You both plot the distribution and the transformed log of the distribution. But it is not clear to me that you do anything with this chart data. What does it tell you? To me it is very obvious that there are some significant outliers for most of the features in this data-set. You do however provide a logical and well-referenced argument for imputing based on the median. You also consider the use of multivariate feature imputation. I think you overstate the case to say that it is dangerous to use this if you do not know the distribution. But you have documented your rationale and for that reason I concur with your decision.

1. Is the wheat in this dataset homogeneous, random or does it fall into distinct types?:

You begin by first describing the data and then using a scatter-matrix. This is a reasonable of approach to understanding the data and you immediately identify the huge range of some of the features. You then plot the scatter-matrix for subset of the data and provide a description of the various kind of distributions that one might often expect. You point out that the data is highly skewed. Whereas actually I would suggest, that this really indicates significant outliers. You can see this as single points in the off-diagonal cells of the matrix. You also mention the existence of these outliers and confirm that using a violin plot. Your discussion around the decision to remove the outliers or not is completely valid. It is always tempting just remove them. But you are quite correct to mention the in some cases they will represent a really important aspect of the data. So, your rationale is effective in this case.

Rob Collins **January 6, 2025**

You treat the data as for separate groups depending on the nature of the outliers. This is not wrong, but I think you are working too hard. Simply removing data outside of the particular extreme range in all cases would have essentially achieved the same ends and been rather simpler to execute. As previously, it is good to see that you have provided references to support your arguments.

Your updated violin plots confirm that you have done an effective job of removing outliers. They further demonstrate that many of the features have bimodal or multi-modal distributions. This is, a first clue that the data may be clustered.

You consider standardizing the data. You correctly point out that this are commonly applied approach and you also provide very good reasoning for why you decide not to apply it in this case. I was pleased to see that you are really applying critical thinking about the processing you were doing on this data. It is all too easy simply to follow a wrote process rather the thinking about the actual data and how it realistically needs to be processed.

You decide to apply PCA to the data. This is one approach to data reduction and you do provide a good description of what it does and provide references into the literature. You do not exactly explain how PCA would help you identify groups in the data, other than it was somehow simplify the data. Of course, the PCA data is rather less interpretable in terms of the original features without transforming back into the original feature space. So, your decision to use PCA is that somewhat at odds with your previous rationale for not standardizing the data. Further PCA is a common enough technique at this point in analysis and I accept your rationale.

It is somewhat impressive that you lay down the algorithm for PCA. However, this is the kind of case that I feel should be followed up with a viva voce; if I asked you to describe the process of calculating first a covariance matrix and then performing an Eigen decomposition, would you feel confident to do that?

Both the Scree-plot and your chart of cumulative explained variance do indicate that there are relatively fewer dimensions underlying this data than the full data-set would suggest. But you have not yet, at this point, argued how this world demonstrate that the data is clustered into ‘types’. At this point in my review, I retain an open mind since I think there is a route to success here. But at this point you have not argued your method.

Rob Collins **January 6, 2025**

You then use a power transformer. But again, I do not following the process by which you intend to demonstrate that there are distinct types in this data-set.

Finally, you produce the scatter-matrix of your transformed data. This does indeed visually indicate clusters within the transformed space.

You continue and identify a series of different clustering algorithms. Your discussion around this is reasonable although in practice I think you would find that all of them would have worked with some success. Actually, the wheat types are very well clustered into mostly disjointed sets in the higher dimensional data-space.

You produce a silhouette chart and also an elbow plot to discover that there are in fact eight clusters. Then you use the SNS pair plot which I think gives a very clear and attractive display of the nicely separated clusters. This is indeed in my mind a good demonstration that there are distinct types of wheat.

2. If the wheat is homogeneous or random, then identify the overall characteristics of the complete sample: Your conclusion is correct. There are definitely different types of wheat in this dataset.

3. If there are distinct types of wheat in the dataset, then identify the characteristics of each type of wheat: You applying k-means appropriately and realize that the cluster centroids are indeed a useful way of characterizing the various types of wheat. A small difficulty I have here is that you are still working with the transformed data-set. So, you are not exactly characterizing the wheat in terms of the original features but rather in the transformed space. I think there are lots of circumstances and which this would be a very reasonable thing to do (for example in cases when there are very large amounts of data). In this case the dimensionality is relatively low and the overall types of size is not huge. It would have perhaps been simpler to work in the un-transformed space. In which case the cluster centroids would characterized each of the wheats in terms of those original features. So, I am not saying your approach is incorrect but I do think the exact method is arguable. If I was doing this I probably would have elected to work in the untransformed space.

You say “this data is still very difficult to interpret as it is just a bunch of

numbers". This is very casual language and it is best to avoid such language in academic writing and in technical papers. There are but much better ways of saying this.

The PCA components and then the loadings in each dimension. You perform this correctly but I do not think you do a good job at explaining exactly why you are attempting to do this. Which part of the questions does it relate to how does it help you characterize each of the different wheats?

You say "What we care about, however, is not so much what features make up each principal component but rather what features distinguish each component from one another".

But I feel a confusion has arisen here. It seems as if you are trying to characterize the difference between the transformed dimensions, whereas actually you should be trying to characterize the difference between the wheats (probably in terms of their cluster centroids). So, I think the PCA has not done you a favour here because it has somewhat clouded the argument and confused the process that you are trying to follow. To be clear, the seven PCA dimensions do not have some direct one-to-one correspondence with the eight clusters (self-evidently!). Each of the cluster centroids can be characterized in seven dimensions and in that seven-dimensional space, the clusters are distinct.

Actually, all I think that is really required at this point is to cluster the wheats into those eight clusters (probably in the original high-dimensional space) and then to tabulate or chart the cluster centroids, for example by showing their mean and or standard deviation in each dimension.

You say "we find a decent level of separation". Again, this is very informal and does not really mean very much at all in the context of a rigorous analysis of the data.

Finally, you do produce a series of bar charts showing the composition of each cluster with respect to each of the features. To me this is the answer we have been searching for. Each of those plots represents a unique signature for a type of wheat.

I noticed that you also do revert to working in all 40 dimensions. So, it seems that you have arrived at the same place I would have but via a somewhat circuitous route. (I do accept of course that it gave you the opportunity to

demonstrate that you understood PCA which if nothing else is a good bit of examination strategy on your part!)

4. What is the relationship between the input variables (first 40 features) and the 5 response variables?

You correctly perform a train / test split on the data to enable quality checking and you immediately realise that this is a regression problem. You provide a good explanation of your chosen quality metrics and it is good to see them discussed beyond simply applying them.

You quickly identify that there appears to be a good fit on the data. You go somewhat beyond the course material by calculating the residuals. I do not believe that I presented this in the course material and it is good that you have done some reading around this subject and realize the importance of doing this. You have applied this correctly and I agree with your conclusion.

You have further applied a polynomial regression of order 2 with not dissimilar results. You are conclusion under the circumstances is not unreasonable.

5. Are there specific features which 'drive' specific responses?

You immediately identify that this part of the question can be addressed by looking at the feature coefficients for each of the responses. The bar charts you produce are a very convincing demonstration of this.

I wonder if you missed an opportunity at this point. Had you produced the correlation heat-map individually for each type of wheat and additionally included the response variables I believe it would have been very apparent that there was a strong relationship between certain features and certain responses. This would have constituted a useful summary of your bar charts shown previously.

You also provide quite a deep discussion around multicollinearity at this point. Again, I think this provides a good indicator that you are have either thought rather deeply about this or done some reading around the subject.

Considering your discussion of the Variance Inflation Factor: this is significantly beyond the material presented during the course. Well done for

discovering and including this.

You have also recognised that LASSO is a good option here and the provided a very good rational around that.

6. If there are different types of wheat present in the dataset, then which type of wheat has the greatest impact on each response variable?: It is not clear to me why you went back to doing PCA at this point and you do not provide a clear rationale for doing so.

It does not seem to me that you quite managed to solve this last part of the problem. The logic is that certain features have a greater impact on certain response variables. Additionally certain types of wheat measure higher or lower in each of those features. So, there is a two-step process required here: from wheat to composition and composition to response variable.

Overall, this was a very good piece of analysis. It was well documented well-argued and executed professionally. In some cases, you went well beyond the course material as presented and provided good references to support your arguments. In a couple of places, I think you went slightly off track. However, you managed to get back on track each case. I think you were very close to solving the very last part of the problem and I think you had all of the information required to do so. Well done.

Rob Collins **January 6, 2025**