Chemometrics KLGN10

Module 2: Preprocessing data and Basic MVA

Here we learn to prepare our data for MVA and then do a PCA to become familiar with the results. Assignment 1 will be devoted to preprocessing and PCA. Assignment 2 is your graded assignment combining both sections, along with what you learned in Module 1.

Assignment 1

The Sensory database

This dataset involves perceptions to the consumption of selected fruits and vegetables. What are the relations between perceptions and fiber type? Before we can answer these questions, we need to preprocess the data.

Pre-Import Preprocessing

Open SENSORY.csv (tab separated) and address the following in each spreadsheet:

- Any missing/extra rows or columns?
- Any missing data?
- Odd formats?
- Headers OK?
- Outliers that are obvious?

Edit the spreadsheet, saving as SENSORY_mod.csv. Briefly describe what you changed. How did you address empty or outlier cells? How will you test if your replacement method would influence the experiment?

Post-Import Preprocessing

Import the data into Python as a df and plot for outliers using a lineplot (plt.plot). Test for outliers using Grubb's test. Follow up by doing cluster analysis (dendrogram) to see how well the samples group. This analysis will give us a 'heads-up' on our upcoming PCA. Has cluster analysis identified any potential groupings or miss-classified samples? Make any adjustments to your data and rationalize these changes. Extra tasks might be to color outliers (based on Grubbs for example) on an XY scatterplot of two interesting dependent variables, or binning one of the numeric independent variables (ie making 'low' and 'high' Pressure).

Basic MVA

A quick PCA should be carried out to look at how the dependent variables associate with each other. Plot both the loading variables and scores in a PCA biplot with a title, labelled axes and legend colored by fiber type. What relations do you see? How does this plot relate to the cluster analysis? Show the Scree plot. How many components explain 90% of the variance? Which fruit/veg is most ‘thick’? Is it significant (do an ANOVA/KKW using fiber as the categorical variable and thick as the dependent variable)? Extra analyses might include doing a 3D PCA, doing a corrplot comparing each of the dependent variables, comparing results to a MANOVA, or studying other variables such as Pressure.

Dummy variables

Turn the fiber variable into dummy variables (so a column for each fiber type, with a 1 in rows with that fiber type, otherwise a 0). Replot the PCA adding the dummy variables. This PCA similar to the original? Extra analyses would be to regroup (categorize) important variables, such as turning fiber type to ‘fruits’ vs ‘vegetables’ and repeating the PCA. Once any interesting dependent variables have been identified by PCA, you could also then apply the univariate analyses learned in Module 1.

Assignment 2

Be sure to examine and follow the rubric for this assignment. The report should contain an introduction, methods, results, discussion and conclusion. Answers should be more comprehensive than above. Take care to format figures, tables etc. Justify your actions and conclusions.

A study was done to look at vitamin levels in different sorts of blueberry, grown in different regions of Massachusetts, USA. The goal is to see if year, location and blueberry type have any influence on vitamin levels. Use the data from Blueberry.csv. Use what you have learned in Modules 1+2 to analyse the data.

Consider:

- What blueberry types are most nutritious?
- What blueberry types are most similar to each other?
- What region seems best for blueberry nutrition?
- What year was best for blueberries in terms of nutrition?
- Does temperature or rainfall appear to play a role in blueberry nutrition?
- What might PC1 and PC2 of a PCA represent in this study?
- What is a major issue with this experiment? How might you fix it?