Continuous Data

1 Introduction

Figure 1: freeimages.com

Introduction

Today, we are going to work with a popular wine data set containing measurements for different attributes of wines. This makes it a perfect example to learn on how to handle continuous data. The data set itself is available to you in Moodle as `winedata.csv`, whereas information pertaining to the vineyard each wine in the data set is from is saved to `winelabels.csv`.

Before you start: If you haven't already, make sure to add seaborn and scikit-learn to your Python interpreter as you are going to need it to complete this exercise.

simon.schindler@fh-salzburg.ac.at
maximilian.schirl@fh-salzburg.ac.at
martin.uray@fh-salzburg.ac.at

2 Lab Assignments

As always, you are asked to finish the three assignments below until the next session.

Data Exploration and Cleaning ✓

For your first assignment, you are asked to complete the following steps after loading the `winedata.csv` using pandas:

1. Try to get an idea of what exactly is stored in `winedata.csv`. Check the size of the dataset (columns and rows) and see if you can find out more about the columns and its values.
2. Call the describe() method of your `DataFrame`. What do you see?
3. Check for missing values (NaNs), calculate their occurences and treat missing values correctly according to one of the methods shown to you in the lecture of DAPIL. Be prepared to defend your choice!

Apply and Visualize Normalization ✓

One technique often used in data preprocessing, especially when dealing with features in differing value ranges, is normalization. Through normalization, differences in value ranges for multiple columns in a data set can be compensated by (e.g.) transforming each feature in such a way that its mean is 0 with standard deviation 1. While this sounds difficult, it is actually pretty easy to do, and we'll use scikit-learn's StandardScaler for it. So in short:

1. Use seaborn's boxplot() function **before** applying normalization on your cleaned data set.
2. Transform the data by correctly applying the `StandardScaler`.
3. Redraw the `boxplot` and be prepared to talk about the differences you can make out!

Apply and Visualize Dimensionality Reduction✓

Trying to figure out similarities between the data points (wines) in our data set is a challenging task. As it stands, we have thirteen features describing each wine. If we want to use figures to visualize the relationships between the different wines, we need a way to reduce the dimensionality of the dataset to two dimensions. There are many different ways to do this. You may, for example, simply select two features from the thirteen by hand to achieve this. This approach has some limitations though, as you lose all the information present in the other eleven columns.
Other techniques are PCA and TSNE, both well known and state-of-the-art methods to implement dimensionality reduction. For now, knowing how to apply these techniques on a

simon.schindler@fh-salzburg.ac.at
maximilian.schirl@fh-salzburg.ac.at
martin.uray@fh-salzburg.ac.at

data set is enough. However, if you want to know a bit more about how these techniques work, check the scikit-learn documentation or consult your lecturer. Complete the following steps to finish this assignment:

1. Instantiate an object of your preferred method for dimensionality reduction and don't forget to parametrize this object: make sure to set the **number of target dimensions to 2**.
2. Visualize the resulting dimensions in a scatter plot.
3. Load `winelabels.csv` and add color to your scatter plot indicating the vineyard each data point was produced in. Do not forget to add a **legend** to your plot.
4. Be prepared to explain what you see in our next session!

3 Homework

At the beginning of the next lab session, you will get a chance to indicate which of the **lab assignments** (✓) you completed and some of you will be asked to present their solution in class. Each student will be asked to present their solutions **at least twice** over the course of the semester. Other than that, no additional upload is needed. Please see the Course Syllabus for details on how this marking of assignments and their presentation affects your grade.

simon.schindler@fh-salzburg.ac.at
maximilian.schirl@fh-salzburg.ac.at
martin.uray@fh-salzburg.ac.at