Homework assignment 2:
Data visualization and Probability and Statistics

Objective: The overall objective is to get an understanding of the many ways data can be visualized. Upon completing this exercise you should be familiar with histograms, boxplots, and scatter plots.

Material: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, section 3.3

Important: The following points is how you hand-in the homework assignment.

- Provide clear and complete answers to the questions below (not hidden somewhere in your source code), and make sure to explain your answers / motivate your choices. Please make as PDF file.
- Source code, output graphs, derivations, etc., should be included, and zipped together with the PDF file.
- Hand-in: upload to Blackboard.
- Include name, student number, assignment (especially in filenames)
- For problems or questions: use the BB discussion board or email.

2.1 Visualizing wine data

We will in this part of the exercise consider two data sets related to red and white variants of the Portuguese "Vinho Verde" wine [1]. The data has been downloaded from `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`. Only physicochemical and sensory attributes are available, i.e., there is no data about grape types, wine brand, wine selling price, etc. The data has the following attributes:

This exercise is based upon material kindly provided by the Cognitive System Section, DTU Compute, `http://cogsys.compute.dtu.dk`. Any sale or commercial distribution is strictly forbidden.

#	Attribute	Unit
1	Fixed acidity (tartaric)	g/dm^3
2	Volatile acidity (acetic)	g/dm^3
3	Citric acid	g/dm^3
4	Residual sugar	g/dm^3
5	Chlorides	g/dm^3
6	Free sulfur dioxide	mg/dm^3
7	Total sulfur dioxide	mg/dm^3
8	Density	g/cm^3
9	pH	pH
10	Sulphates	g/dm^3
11	Alcohol	% vol.
12	Quality score	0-10

Attributes 1−11 are based on physicochemical tests and attribute 12 on human judging. The data set has many observations that can be considered outliers and in order to carry out analyses it is important to remove the corrupt observations.

The aim of this exercise is to use visualization to identify outliers and remove these outliers from the data. It might be necessary to remove some outliers before other outlying observations become visible. Thus, the process of finding and removing outliers is often iterative. The wine data is stored in a MATLAB file, `Data/wine.mat`.

2.1.1 Load the data into Python using the `scipy.io.loadmat()` function. This data set contains many observations that can be considered outliers. Plot a box plot and a histogram for each attribute to visualize the outliers in the data set. From prior knowledge we expect volatide acidity to be around 0-2 g/dm^3, density to be close to 1 g/cm^3, and alcohol percentage to be somewhere between 5-20% vol. We can safely identify the outliers for these atributes, searching for the values, which are a factor of 10 greater than the largest we expect. Identify outliers for volatide acidity, density and alcohol percentage, and remove them from the data set. Plot new box plot and histogram for these attributes and compare them with initial ones.

Hints:

- You can use function zscore to standardize your data before you plot a boxplot.
- You can use function subplot to plot several plots on one figure

2.1.2 Make scatter plots between attributes and wine quality as rated by human judges. Can you identify any clear relationship between the attributes of the wine and

wine quality? Which values of these attributes are associated with the high quality wine?

Hints:

- You can calculate correlation coefficient between attributes and wine quality using function pearsonr(x, y) in module scipy.stats.stats to measure the strength of association.

2.2 Visualizing the handwritten digits

In this part of the exercise we return to the hand written digit data sets (`http://yann.lecun.com/exdb/mnist/`).

2.2.1 Load `zipdata.mat` by typing `loadmat('Data/zipdata.mat')`. There are two data sets containing handwritten digits `testdata` and `traindata`. Here, we will only use `traindata`. The first column in the matrix `traindata` contains the digit (class) and the last 256 columns contain the pixel values.

Create the data matrix `X` and the class index vector `y` from the data. Remove the digits with the class index 2-9 from the data, so only digits belonging to the class 0 and 1 are analyzed. Visualize the first 10 digits as an image. Script `ex2_1_1.py` contains an example of visualizing digit as an image.

Next, compute the principal component analysis (PCA) of the data matrix. Now, using the PCA, create a new data matrix `X` overwriting the old data matrix. The new data matrix should have 4 attributes corresponding to PC1-PC4. Reconstruct the initial data using PC1-PC4. Visualize the first 10 digits as an image for the reconstructed data and compare them with images for original data.

Make a matrix of scatter plots of each combination of two attributes against each other. Make a 3-dimensional scatter plot of three attributes. Plot elements belonging to different class in different colors.

Hints:

- Script `ex2_1_1.py` can help you to visualize digits as an image
- To compute a PCA subtract the mean from the data, $\mathbf{Y} = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}$, and calculate the SVD i.e., $\mathbf{Y} = \mathbf{USV}^T$.
- To project the data onto the first four principal components use `Z = Y*V[:,:3]`
- To reconstruct the data from projection use the following formula: $\mathbf{W} = \mathbf{Z} * \mathbf{V}[:, \mathbf{0} : \mathbf{3}]^T + \boldsymbol{\mu}$

2.3 Probability and Statistics

The aim of this exercise is to learn how to calculate basic statistics in Python.

2.3.1 A study of a very limited population of Aliens reveals the following number of body appendages (limbs):

$$2, 3, 6, 8, 11, 18$$

i. Find the mean m and the standard deviation σ of this population.

ii. List all possible samples of two aliens without replacement ("zonder terugleggen"), and find each mean. Do the same with samples of four aliens.

iii. Each of the means above is called a sample mean. Find the mean of all the sample means (denoted by m_x) and the standard deviation of all the sample means (denoted by σ_x) for both the $N = 2$ and $N = 4$ samples.

iv. Verify the Central Limit Theorem: (i) compare the population mean with the mean of both sample means; (ii) compare the population standard deviations divided by the square root of the sample size with the standard deviation of both sample means (i.e., does $\sigma_x \approx \sigma/\sqrt{N}$). BTW, a better approximation for small population sizes is $\sigma_x = \sigma/\sqrt{N} \times \sqrt{(M-N)/(M-1)}$ with $M = 6$ the size of the original population.

v. Plot the distribution of the population and the distributions of both sample means using histograms. What happens to the shape of the sample means distribution as the sample size (N) increases?

Hints:

- You can use methods mean() and std() of NumPy array class in Python to calculate mean and standard deviation
- You can use method itertools.combinations(v,n) of module itertools to find all possible samples of a vector v taking n elements at a time.

References

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.