DATA.ML.100 Introduction to Pattern Recognition and Machine Learning
TAU Computing Sciences
Exercise - Week 4: *Bayes classification (female-male dataset)*

Be prepared for the exercise sessions (watch the demo lecture). You may ask TAs to help if you cannot make your program to work, but don't expect them to show you how to start from the scratch.

1. **Male and female – Bayesian classifier** (40 points)

   Download the male and female height and weight measurements:

   - `male_female_X_train.txt`
   - `male_female_X_test.txt`
   - `male_female_y_train.txt`
   - `male_female_y_test.txt`

   (a) Height and weight histograms (10 points)
   Compute the histograms of the male height, female height, male weight and female height measurements using the NumPy `histogram()` function. For the both use 10 bins and fixed ranges ($[80, 220]$ for height and $[30, 180]$ for weight).

   Plot two histograms, one for height and another for weight, that includes the both classes.

   Estimate visually which measurement is likely better for classification.

   Return the following items:

   - Python code: <surname>_male_female_histogram.py
   - A full desktop screenshot that includes a terminal window executing your code: <surname>_male_female_histogram_desktop.png
   - The histogram plot 1: <surname>_male_female_histogram_plot_height.png
   - The histogram plot 2: <surname>_male_female_histogram_plot_weight.png

   (b) Baseline classifier (10 points)

   Make a classifier that assigns a random class to each test sample and computes its classification accuracy. Print the accuracy.

   Make another classifier that assigns the most likely class (highest a priori) to all test samples. Print the accuracy.

   Return the following items:

   - Python code (single file): <surname>_male_female_baseline.py
   - A full desktop screenshot that includes a terminal window executing your code: <surname>_male_female_baseline_desktop.png

   (c) Bayes classifier with non-parametric distribution (20 points)

   Compute and print the prior probabilities for male and female.

   Compute class likelihoods, $p(\text{height}|\text{male})$, $p(\text{weight}|\text{male})$, $p(\text{height}|\text{female})$ and $p(\text{weight}|\text{female})$ for all test samples. This can be done by using the bin min/max values returned by NumPy `histogram()` function. You can calculate the centroid of each bin and assign each test sample to the closest bin. After knowing the bin index, the likelihood can be computed using the *count* vector provided by the same `histogram()` function.

   Classify all test samples and compute the classification accuracy. Print accuracies for height only, weight only, and weight and height together (multiply likelihoods).

   Return the following items:

   - Python code (single file): <surname>_male_female_bayes.py
   - A full desktop screenshot that includes a terminal window executing your code: <surname>_male_female_bayes_desktop.png