

MTAT.03.227 Machine Learning
Spring 2015 / Exercise session I
Nominal score: 10p
Maximum score: 15p
Deadline: 17th of February 16:15 EET

1. The aim of this exercise is to propose and test various hypotheses about countries and their flags. To complete this task, you should use basic data frame manipulation operations and basic plotting functions. First, download the Flags Data Set from UCI Machine Learning repository <http://archive.ics.uci.edu/ml/datasets/Flags>. Use `read.csv` with right arguments to read in the data about flags (file `flag.data` with description `flag.names`).
 - (a) Replace numerical code with descriptive labels given in `flag.names` and report how are the attributes `religion`, `landmass`, `language` split by drawing bar plots (**1p**).
 - (b) Use box plot to visualise and compare the areas and population of countries and their dependence on the landmass and religion. More precisely, study where are the countries larger in Europe or in Africa. Compare similarly countries with different religions (**1p**).
 - (c) To get more fine-grained overview of different groups of countries and compute 10, 20, ..., 90 % quantiles for population, area, and population density attributes. After that compare corresponding quantiles visually by drawing a `qqplot` with 10, 20, ..., 90 % quantiles. Give some sort of interpretation to the result (**1p**).
 - (d) Compute absolute and relative support, confidence and coverage for the following rules (**1p**):
 - `religion = Marxist` \implies `Flag = Red`
 - `religion = Muslim` \implies `Flag = Green`
 - `language = English` \implies `Flag has saltires`
 - (e) Find all rules with a single premise (e.g. `religion = Marxist`) that predict saltires, crosses or existence of sun or stars on the flag. Order all rules according to confidence and support. Decide which of those rules are justified, i.e., they are not random coincidences (**1p**).
 - (f) Find all justified rules consisting up to two premises for predicting at least three visual attributes on the flag (e.g. saltires, crosses, existence of stripes). Order the rules in the order of confidence and interpret results (**1p**).
 - (g) Use the set of prioritised rules to predict visual attributes. Create the confusion matrix (2×2 matrix containing true positives, false positives, false negatives, true negatives). Compute precision and recall. Do these values change if you change the positive class. For

instance, you can consider the existence of stripes a positive feature or negative feature based on your preference. **(1p)**

2. It is often too difficult or expensive to gather the entire dataset. Hence, statistical summaries are often computed based on relatively small sample. The aim of this exercise is to study whether strategy is reasonable. Let \mathcal{X} be the set of consecutive numbers $\{1, 2, \dots, 10000\}$. Study what happens if we compute minimum and mean on based on a small random sample \mathcal{S} instead of computing it over the entire dataset \mathcal{X} .

- (a) Write a function that draws randomly n elements from the set \mathcal{X} and outputs corresponding minimum and mean values $\min(\mathcal{S})$ and $\text{mean}(\mathcal{S})$ on that sample. Note that some elements can be sampled several times. You can use function `sample` for sampling **(1p)**.
- (b) Study how these estimates depend on the sample size n . For that repeat the sampling procedure 1000 times for a fixed value of n and compute the average and standard deviation for $\min(\mathcal{S})$ and $\text{mean}(\mathcal{S})$. Study what happens with these values if $n = 1, 10, 100, 1000, 10000$. For that draw a scatter plot where n is on the x -axis and the average value together with error bars is on the y -axis. Interpret results **(2p)**.
- (c) Study how do these result change if the sampling procedure is without replacement. Repeat the same experiment as in the previous subtask and do the same visualisation. Interpret results **(1p)**.

3. One often needs to simulate real-life processes in order to make justified decisions. Let us consider the following artificial problem. You need to take a 100,000 € loan for 20 years. You are can choose an offer with fixed rate 2.5% interest and a offer with floating rate. In both cases, you need to pay 5,000 € per year plus the interest rate for the remaining sum.

Long terms observations have shown that the floating rate can be modelled as follows. In normal circumstances, the rate is drawn form the normal distribution with parameters $\mu = 2$ and $\sigma = 0.5$. In a crisis, central banks lower the basis of interest rates and thus the rate is drawn form the normal distribution with parameters $\mu = 0$ and $\sigma = 0.5$. In both cases the rate cannot go below zero, that is, if the rate is below zero, it is raised to zero. A crisis occurs if the rate is below 1% in the previous time point.

- (a) Write a simple for loop that models interest rate fluctuations for the following 20 years. Draw the corresponding graph **(1p)**.
- (b) To get grip what actually occurs, simulate the interest rate fluctuations 10,000 times. For that rewrite the for loop so that the interest rate is stored in $10,000 \times 20$ matrix and there are no if statements in the code. Compute the total amount of interest for each simulation and visualise the resulting distribution with histogram **(1p)**.
- (c) Find out how interest you would need to pay for fixed rate and estimate the probability that the floating rate interest would have been a better option **(1p)**.

- (d) Tabulate the probability that a floating rate is a better option for various fixed interest rates and specify what would be a fair fixed rate offer. Illustrate the discussion with an appropriate graph (**1p**).
4. A probability $\Pr[\mathcal{A}]$ of an event \mathcal{A} can be interpreted as follows. *If we conduct enough independent trials, then the fraction of trials where the event \mathcal{A} occurs is roughly $\Pr[\mathcal{A}]$. This claim can be also reversed. Given enough independent trials we can estimate the probability of any event.* The aim of this exercise is to test this claim in practice. GNU R has a set of prebuilt distributions for which probabilities can be computed analytically. In particular, the function `rnorm` allows us to draw elements from a normal distribution $\mathcal{N}(0, 1)$.
- (a) Draw 10000 samples x_1, \dots, x_{10000} from the normal distribution $\mathcal{N}(0, 1)$ and estimate what is the fraction $f(t)$ of $x_i \leq t$. Tabulate the results for $t = -3, -2.5, \dots, 2.5, 3$ and draw the corresponding graph. You can add more points if you wish (**1p**).
- (b) GNU R also has an analytical function `pnorm` to estimate the probability $F(t) = \Pr[x \leftarrow \mathcal{N}(0, 1) : x \leq t]$. Tabulate the results for $t = -3, 2.5, \dots, 2.5, 3$ and draw the corresponding graph. Compare the results. Does the claim hold in practice (**1p**)?
- (c) Note that the knowledge of $F(t)$ is enough to compute the probability that a sampled element x is lies in the interval $(a, b]$:

$$\Pr[x \leftarrow \mathcal{N}(0, 1) : a < x \leq b] = F(b) - F(a) .$$

Use this inequality to predict the shape of a histogram with breaks positioned at $-3.5, 2.5, \dots, 2.5, 3.5$. Compare the prediction with the true histogram (**1p**).