**Machine learning II**
**Master Data Science**
**Winter Semester 2019/20**

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

**Prof. Tim Downie**

# Workshop 1
# Imputing missing values in R

### Exercise 1 Getting started

(a) Download the file `Wksp1MissingData.R` from Moodle, and save it to a sensible directory.

(b) Each workshop you should start and finish your session properly. Start R or RStudio, open a new script file and set your working directory to the same directory, in which you saved the R source file in part (a).

(c) Today you will be using the following libraries, which you will probably need to install. `mice`, `VIM`, and `NHANES`. Install these using

```
> install.packages("mice","VIM","NHANES")
```

(d) Throughout the Workshop add comments to the source code to make your work easier to read and regularly save your source code.

(e) Work through the example code step by step, reading the explanations and hints in this exercise sheet in parallel.
You will need to replace the missing code `???` for it to work

*Structure of the exercises*: in Exercise 2 you will load the tropical atmosphere ocean data, do some simple data exploration and fit a linear regression on the non-missing observations.

In Exercise 3, you will implement some of the imputation methods covered in Lecture 1 using the `tao` data set (in the `VIM` library). In Exercise 4 you will use the Gibbs sampling method in the `mice` library.

### Exercise 2 The tropical atmosphere ocean data

The `tao` data set is a small subsample of the Tropical Atmosphere Ocean (TAO) project data, containing daily measurements at 5 locations, for two years; one El-Niño year and one La-Niña year. The variables are: `Year`, `Latitude`, `Longitude`, `Sea.Surface.Temp`, `Air.Temp`, `Humidity`, `UWind` (East-West daily average) and `VWind` (North-South daily average).

(a) After loading the data, there are some commands to inspect the number missing of missing values in each variable, and the pattern of the non missing values. The graphics functions alter the margins in the graphics window and do not re-set them. The command `startMar<-par()$mar` saves the starting values for the margins so that you can return to the default settings later.

(b) How many observations have a missing value for *humidity*? How many observations have more than one missing value?

(c) Fit a linear model to predict the `Sea.Surface.Temp` using the "non-missing data"; specifically those rows with no missing values.

(d) What is the conspicuous problem with using univariate imputation methods on these missing values?

**Exercise 3 Univariate imputation**

(a) Define a function which completes the missing data using the *mean replacement* method. Inspect the imputed values and compare the linear model results with the model obtained using the "known" data.

(b) Repeat using the *Mean/Variance Simulation* method

(c) Repeat using *Direct Random Sampling*.

**Exercise 4 Multivariate imputation using Gibbs sampling**

(a) Using multivariate imputation, we can use the information from `Year`, `Sea.Surface.Temp` etc. to get more realistic imputed values. The function `mice()` calls the Gibbs sampling routine. `maxit=50` and `m=5` has been specified. This means that 50 full iterations of the Gibbs sample are run and the last 5 will be used for the imputation. This means that we get 5 versions of the imputed data.

(b) Only the imputed values are stored in `GibbsData`. To get a full dataset with known and imputed values use `complete()`.

(c) The `with()` function runs the lm() function 5 times, once for each of the imputed data sets.

N.B. `with()` is a generic function. Because GibbsData has class `"mids"` ('multiply imputed data set') the function `with.mids()` will be called. This runs the `lm()` function 5 times using each of the imputed data sets.

(d) To aggregate the results of several models, use the `pool()` function. Choose a final model.

**Exercise 5 Imputing missing data for the Diabetes data set**

In ML 1, Workshops 7 and 8 you used the data set `Diabetes` which contained just 3 variables. The data were obtained from the NHANES data (American National Health and Nutrition Examination surveys).

The data set `Diabetes2` used in today's lecture comes from the same source but uses 13 variables. In total there are $n{=}10\,000$ observations, but there are only $6492$ observations with no missing values. You will now use `mice()` to impute the missing values for these 13 variables.

(a) Load the data and store the 13 variables in the data frame `Diabetes`. Inspect the missing values.

(b) Run the Gibbs sampler on the `Diabetes2` data. Because this data set is much larger than in Exercise 3, the Gibbs sampler takes much longer to run. The total number of iterations is reduced to 10 which takes roughly 3 minutes to run. This is sufficient for a Workshop, in practice it is recommended to increase the burn-in period.

(c) Fit a logistic regression model to the non-missing rows, and to the first of the imputed data sets.

(d) Fit an aggregated logistic regression model. Remove the non significant variables to obtain the final logistic regression model.

(e) This last section of code uses the Gibbs sampled imputed values to fit a tree classifier to the data.