Guide for Module 1.3 exercises

Data Preparation, Summarization and Visualization

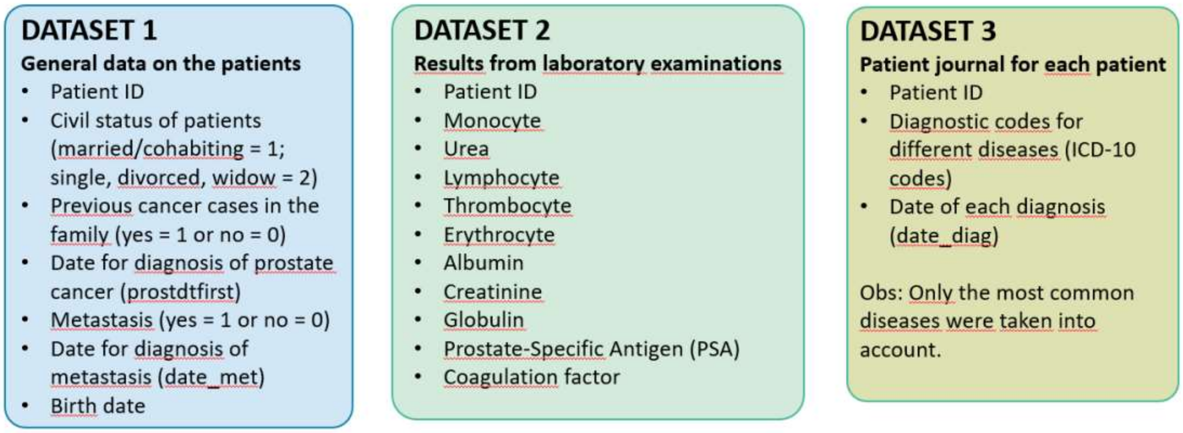
Study description:

Urologists from the Urology department at Odense University Hospital have selected 1159 patients that were diagnosed with prostate cancer (PC) at their first visit with them.

Some of those patients were also diagnosed with metastasis (metastasis = 1) and others do not present metastasis (metastasis = 0). Metastatic cases were detected by the appropriate method (e.g. CT scans, MRI).

The urologists have requested a bunch of laboratory examinations from these patients (measured in blood and urine), as well as retrieved their medical history (i.e. patient journals).

They got in contact with us, the AI team, to help them to identify ways of predicting which patients will present metastasis and which will not. This could be further used as a prediction tool in an earlier stage, and would also avoid overtreatment and invasive examinations.



DATASET 1

General data on the patients

- Patient ID
- Civil status of patients (married/cohabiting = 1; single, divorced, widow = 2)
- Previous cancer cases in the family (yes = 1 or no = 0)
- Date for diagnosis of prostate cancer (prostdtfirst)
- Metastasis (yes = 1 or no = 0)
- Date for diagnosis of metastasis (date_met)
- Birth date

DATASET 2

Results from laboratory examinations

- Patient ID
- Monocyte
- Urea
- Lymphocyte
- Thrombocyte
- Erythrocyte
- Albumin
- Creatinine
- Globulin
- Prostate-Specific Antigen (PSA)
- Coagulation factor

DATASET 3

Patient journal for each patient

- Patient ID
- Diagnostic codes for different diseases (ICD-10 codes)
- Date of each diagnosis (date_diag)

Obs: Only the most common diseases were taken into account.

Can PSA (Prostate-Specific Antigen) be used as a biomarker to detect metastatic prostate cancer?

Previous studies say the following:

“Based on results from some small, single centre studies published in the beginning of the 1990s, PSA levels above 100 ng/mL have been used as a proxy for metastatic prostate cancer” Thomsen et al., 2020

“PSA < 20 ng/ml have high predictive value in ruling out skeletal metastasis” Kamaleshwaran et al., 2012

TASK 1 – Running simple logistic regression models (30 Min)

The clinicians hypothesize that urinary retention is a relevant variable to predict whether a patient will or not present metastasis.

- Make a simple model, where you want to predict metastasis only based on the variable that indicates whether a patient has or not urinary retention
  - Tip: you can use the glm function in R with family = ”binomial”
  - Check and discuss the results from the model by using the summary function

  Example:

  ```
  your.model.name <- glm(metastasis ~ urin_ret, data=df, family="binomial")

  summary(your.model.name)
  ```

- What is the estimated effect of urinary retention on the probability of having metastatic prostate cancer? How would you interpret this effect?
- Based on this very simple model, calculate the probability of metastasis for a man with urinary retention, and calculate the probability of metastasis for a man without urinary retention.

  ```
  predict(model, data.frame(urinary.retention = ""), type="response")
  ```

  < in urinary.retention, you can select whether 0 or 1>

TASK 2 – Running multiple logistic regression models (30 min)

Now it is time for more complex models:

- First: make a model including all variables you have, except for the principal components you have calculated in Module 1.2 and some other variables that do not make sense to have in the model (e.g. ID, prostdtfirst, date_met, birth_date). You can use the same functions as in the previous task.
  - Does the model fulfill all criteria for a logistic regression model? To check collinearity between the variables, you can calculate the Variance Inflation Factor (VIF) using the car package in R:

  ```
  library(car)

  vif(your.model.name)
  ```

- Second: make a model where you replace all biochemistry variables (e.g. albumine, PSA...) by the principal components you calculated in Module 1.2.
  - Does the model fulfill all criteria for a logistic regression model?
- Discuss the models with your peers (we will discuss it together soon!). What are the advantages and disadvantages of using principal components in the model?

TASK 3 – Improving the models (15 min)

- Select the variables you want to keep in your model by using one of the variables' selection method we discussed in class. You can do it manually, or you can also use e.g. the package MASS, as shown below:

```
library(MASS)

step.model <- stepAIC(your.model.name, direction = "", trace = TRUE)
```

< in direction, you can select whether you want backward or forward selection (or both)>

- Calculate the predicted probabilities that each patient has or does not have metastasis.
  - You can do it "manually" by using the equation from your developed model or, in R, you can use the following:

    ```
    probabilities <- predict(your.model.name, new_data = df, type = "response")
    ```

- Based on the predicted probabilities, and using a cutoff of 0.5 to estimate the predicted response for metastasis (if pred_prob < 0.5 ∀ metastasis = 0 and if pred_prob >= 0.5 - ∀ metastasis = 1), fill out the confusion matrix below and calculate the predictive performance of the model (i.e. accuracy, sensitivity, and specificity).

OBSERVED RESPONSE	**PREDICTED RESPONSE**	
	Metastasis=1	**Metastasis=0**
Metastasis=1		
Metastasis=0		

One example for a R code to create a dataset that contains the predicted probabilities, and the observed and predicted responses:

```
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")

predicted.classes <- as.factor(predicted.classes)

pred.results <- data.frame(observed.classes, probabilities, predicted.classes)
```

To fill out the table, you can either count how many true positives, true negatives, false positives and false negatives there are, as you did in Module 1.1 or you can use the package "caret" in R:

```
install.packages("caret") #Only first time the package is used

library(caret)

confusionMatrix(data = predicted.classes, observed.classes)
```

TASK 4 – Validating and evaluating the new model (45 min)

- Perform a ten-fold cross validation method to evaluate the model performance.
  - Tip for R users: the package "caret" (functions trainControl and train)

```
train.control <- trainControl(method = "cv", number = 10, savePredictions=TRUE) #savePredictions = TRUE is important, so we can check the model's accuracy in each folder

set.seed(123) #set the seed of the pseudo-random so that the same result can be reproduced. Any number can be used.

# Train the model

model <- train(metastasis ~ variables you selected, data=df, method = "glm", family = "binomial", trControl = train.control)

# Summarize the results

print(model)
```

- What is the averaged accuracy found in the previous step? What was the accuracy obtained for each fold? Were they consistent?

  - If using the package ”caret”, remember to set savePredictions=TRUE.

  In R, an example code is:

```
#Inspecting the results for each folder

pred_fold <- model$pred

pred_fold$equal <- ifelse(pred_fold$pred == pred_fold$obs, 1,0)

#Calculating the accuracy per folder

library(dplyr)

eachfold <- pred_fold %>% group_by(Resample) %>% summarise_at(vars(equal),list(Accuracy = mean))

eachfold
```

- Evaluate the model’s predictive performance by calculating the model’s sensitivity, specificity, and accuracy and using

```
pred_fold <- model$pred
```

TASK 5 – Closing remarks

- Compare the predictive performance of your final model with the predictive performance of the first approach we introduced in the beginning of this module (i.e. simply looking at the PSA levels and classifying PC according to a PSA cutoff). Discuss the differences in both approaches, keeping also in mind the ”healthcare system point of view” and the applicability of the model.

- Discuss the results you obtained and make an analysis of the variables kept in your final model (and the ones that were removed from the model). Were you expecting these results? Do you think the significance and estimates for each of the predictors make sense, considering previous studies?

- Imagine you were contacted by prostate cancer clinicians, who would like to predict which of the patients with prostate cancer will develop metastasis. Together with them, you will design a study from scratch, by collecting data from patients with and without metastasis. How would you do this study? What would you do differently in comparison with what we have done in this module (e.g. changes in the study design, different variables to collect, changes in the statistical approach (e.g. classification method), etc)?

For the report, make sure to include (at least) the following:

- Describe your final predictive model and elaborate on your decisions along the way to decide on this specific model (i.e. explain why your model has those specific variables).

- Show the results of the model's predictive performance before and after cross-validation, including the confusion matrix you found in TASK 3.

- Reply and elaborate on all questions from the closing remarks (i.e. TASK 5). Remember to discuss it with your peers.

Remember to always save the script!