

Assignment 7

This assignment accounts for 4% of the assessment for the Data Mining unit.

You are required to analyse a (real) data set concerned with hyperactive thyroid function. Apply **two** of the techniques covered in lectures (your choice which ones), in order to produce predictive models for hyperactive thyroids. The data is provided as a Microsoft Excel file `hyperthyroid.xls` which can be downloaded from the unit website.

The data consists of 3772 records with 24 variables, split as follows:

Binomial input variables

- sex:
- on thyroxine:
- query on thyroxine:
- on antithyroid medication:
- sick:
- pregnant:
- thyroid surgery:
- I131 treatment:
- query hypothyroid:
- query hyperthyroid:
- lithium:
- goitre:
- tumor:
- hypopituitary:
- psych:

Categorical input variable

- referral source:

Continuous input variables

- age:
- TSH:
- T3:
- TT4:
- T4U:
- FTI:
- TBG:

Binomial output variable

- hyperthyroid classification:

There are many missing values, denoted by question marks “?”. How you deal with these is important. Some suggestions are:

- Ignore records with a missing value
- Ignore variables with a missing value
- Substitute the mode, mean or median for the missing value
- Introduce a new category of missing value
- Use 0 for missing values
- Use 0 for missing values and introduce an auxiliary binary variable indicating variable present or not

Your analysis should consider the following points:

- Choice of test and training data;
- A consideration of sensitivity vs. specificity (e.g. using an ROC curve);
- The relative performance of the two techniques considered;
- Model interpretability.

Assignment 8

This assignment accounts for 2% of the assessment for the Data Mining unit.

Download the data set [stars4.xls](#) from the website. Using hierarchical clustering, cluster stars based on their $\log(\text{temperature})$ and $\log(\text{luminosity})$. Try a couple of distance metrics (such as the Euclidean and Manhattan (or taxi-cab) metrics) and a few agglomeration/linkage methods (such as single linkage or complete linkage).

See if you can reproduce the groups shown in the Hertzsprung-Russell diagram. You will find the R-commands `hclust`, `dist` and `rect.hclust` useful. Note that luminosity and temperature are related by $L \approx c_0 T^4$, that is, $\log L = c_1 + 4 \log T$. The constant c_1 is what is used to classify stars, and the scaling factor 4 is important when it comes to forming clusters.