1 Requirement

1. Pick up ONE (or more if you like) favorite problem *below* or *from the datasets in textbook* to attack. If you would like to work on a different problem outside the candidates we proposed, please email me about your proposal. Brave hearts for explorations will be encouraged!

2. Team work: we encourage you to form small team, up to THREE persons per group, to work on the same problem. Each team just submit ONE report, *with a clear remark on each person's contribution.* The report can be in the format of either Python (Jupyter) Notebooks with a detailed documentation, a *poster* (you can download the templates here)

   `https://www.latextemplates.com/cat/conference-posters`

   or a *technical report within 8 pages*, e.g. NIPS conference style

   `https://nips.cc/Conferences/2016/PaperInformation/StyleFiles`

3. In the report, show your proposed scientific questions to explore and main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted through email as a .zip file, GitHub link, or as an appendix if it is not large. There is no restriction on the programming languages to use, but R or Python are recommended.

4. Submit your report by email no later than the deadline (Beijing time 23:59, Oct. 16, 2020), to the following address (datahomework2020@163.com) with Title: Project 1. Late submissions may consume grades.

2 Regression: Animal Species Sleeping Hours

The following dataset contains $n = 62$ species with several features including the average sleeping hours per day (`sleep`). Some values are missing (`NA`).

`https://github.com/liuhaixias1/data_mining/blob/master/sleep1.csv`

Explore the question that *what might affect the sleep that an animal needs.* In this explorative study, you probably need to deal with

- remove or fill-in missing values;
- design your models, e.g. multiple linear regression;
- mixed-type of features: real-valued features (e.g. body weight (`body`) and life time (`life`)) and discrete-valued (categorical) features (e.g. predation (`predation`) and danger level (`danger`));
- estimation of prediction/test error by cross-validation, e.g. in MSE, and choose your favourite model;

- quantification of uncertainty in your model estimates, e.g. error bar for sleeping hour prediction by bootstrap.

3 Bi-Classification: Switch unsafe wells

The following data set contains decision of switching unsafe wells for arsenic pollution in Bangladesh.

`https://github.com/liuhaixias1/data_mining/blob/master/wells.csv`

The predictor `arsenic` described the measured amount of arsenic pollution and the `distance` is how far is the well from the nearest living area. The response is a binary decision variable on switching-off the well (`TRUE/FALSE`). You may explore the models on prediction of switching unsafe wells given various features about the situation. For example,

- logistic regression with your chosen predictors, such as real-valued features (`arsenic`, `unsafe`, and `distance` etc.) and categorical features (`education`);
- fit your models with z-values, p-values;
- estimate the misclassification error, confusion matrix (type I and type II errors);
- compute the ROC curve and Area-Under-Curve to evaluate your model;
- choose your favourite model by cross-validation;
- quantify the uncertainty of your model, e.g. by bootstrap.

4 Multi-classification: Hand-written Digits

The following website about the Elements of Statistical Learning contains a subset of hand-written digit MNIST dataset, which contains 7,291 training examples and 2007 test examples, each example being 16-by-16 256 grayscale images. There are ten classes with id from 0 to 9.

`https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.info.txt`

Training (1.7M): `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.train.gz`

Test (429K): `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.test.gz`

Explore the dataset with your classifiers, such as LDA, QDA, logistic regression with various models.