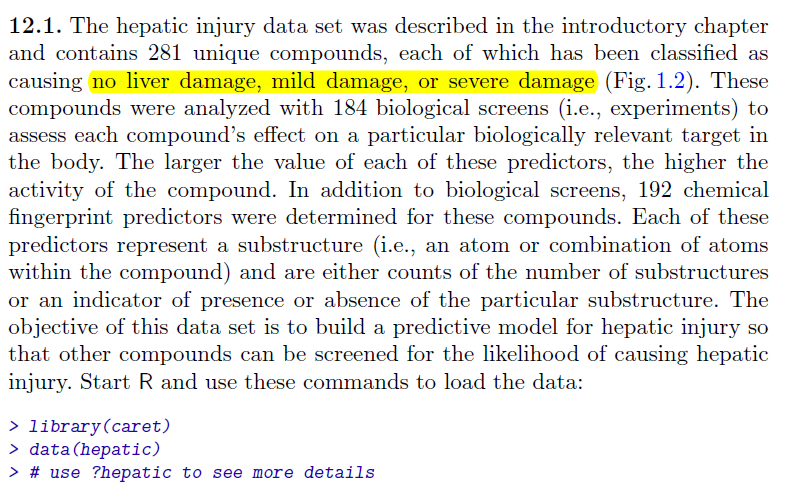
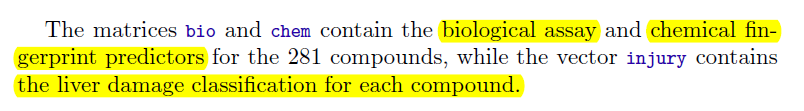
**精算與大數據專題 (期中練習)**

**The data is located in “AppliedPredictiveModeling” package.**



**Data description: [ data(hepatic) ]**



**injury : Y ; bio & chem : X**

1. Data tidying:

- Bind “Severe” & “Mild” into “1”, “None” into “0”

- Combine data to a large X matrix

2. Data preprocessing (Optional)

3. Explore the data (EDA), and modify any new variable. (Optional)

4. Split data to Training (train + validation) & Prediction set. (8:2)

5. Variable selection

Use LASSO to extract important variable.

6. Now, we build all the model and compare them.

1. CART

2. Conditional Inference Tree

3. Random forest

4. XGBoost

5. Elastic net regression

(Tune the alpha & lambda)

* 1. Select different alpha, and run cv.glmnet and get optimal lambda.
* 2. Manually set a grid of alpha and lambda.

(See whether both yield same result)

Calculate the accuracy for each model, using the prediction set. Which model yields the best performance?

* The homework should include “Coding + result screenshot”, “Accuracy comparison” & The following table.
* For Question 5 (Elastic net regression)

Method 1 :

|  |  |  |  |
| --- | --- | --- | --- |
| Lambda / Alpha | Lambda 1 | Lambda 2 | … |
| Alpha 1 | Evaluation index  (e.g. AUC) | Evaluation index  (e.g. AUC) |  |
| Alpha 2 | Evaluation index  (e.g. AUC) | Evaluation index  (e.g. AUC) |  |
| … |  |  |  |
| … |  |  |  |

Method 2 : (Same table as above)