CSC 423 - Data Analysis Project

Jasmine Dumas

October 26, 2015

1. Get the data from source UCI

data <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data", header = FALSE, sep = ",")  
head(data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11  
## 1 1000025 5 1 1 1 2 1 3 1 1 2  
## 2 1002945 5 4 4 5 7 10 3 2 1 2  
## 3 1015425 3 1 1 1 2 2 3 1 1 2  
## 4 1016277 6 8 8 1 3 4 3 7 1 2  
## 5 1017023 4 1 1 3 2 1 3 1 1 2  
## 6 1017122 8 10 10 8 7 10 9 7 1 4

1. Clean the data: Add column headers to the data & remove an missing values if any

Attribute Information: (class attribute has been moved to last column)

|  |  |  |
| --- | --- | --- |
| # | Attribute | Domain |
| 1. | Sample code number | id number |
| 2. | Clump Thickness | 1 - 10 |
| 3. | Uniformity of Cell Size | 1 - 10 |
| 4. | Uniformity of Cell Shape | 1 - 10 |
| 5. | Marginal Adhesion | 1 - 10 |
| 6. | Single Epithelial Cell Size | 1 - 10 |
| 7. | Bare Nuclei | 1 - 10 |
| 8. | Bland Chromatin | 1 - 10 |
| 9. | Normal Nucleoli | 1 - 10 |
| 10. | Mitoses | 1 - 10 |
| 11. | Class: | (2 for benign, 4 for malignant) |

colnames(data) <- c("Sample\_code\_number", "Clump\_Thickness", "Cell\_Size", "Cell\_Shape", "Marginal\_Adhesion", "Single\_Epithelial\_Cell\_Size", "Bare\_Nuclei", "Bland\_Chromatin", "Normal\_Nucleoli", "Mitoses", "Class" )