Санкт-Петербургский Политехнический Университет Петра Великого

Институт прикладной математики и механики

Кафедра Телематики при ЦНИИ РТК

Лабораторная работа

по теме

«Деревья решений»

Преподаватель Уткин Л.В.

Студент гр.43607/2 Лисенкова А.А

Санкт-Петербург

2018 г.

# Постановка задачи

В ходе выполнения лабораторной работы необходимо выполнить следующие задачи:

1. Загрузить набор данных Glass из пакета “mlbench”. Построить дерево классификации для модели, задаваемой следующей формулой: **Type ~ .**, дайте интерпретацию полученным результатам. При рисовании дерева использовать параметр cex=0.7 для уменьшения размера текста на рисунке. Выявить, является ли построенное дерево избыточным. Выполнить все операции оптимизации дерева.

Определить, к какому типу стекла относится экземпляр с характеристиками

RI =1.516 Na =11.7 Mg =1.01 Al =1.19 Si =72.59 K=0.43 Ca =11.44 Ba =0.02 Fe =0.1

1. Загрузить набор данных spam7 из пакета DAAG. Построить дерево классификации для модели, задаваемой следующей формулой: **yesno ~.**, дать интерпретацию полученным результатам. Запустить процедуру “**cost-complexity prunning**” с выбором параметра **k** по умолчанию, **method = ’misclass’**, вывести полученную последовательность деревьев. Выявить, какое из полученных деревьев является оптимальным.
2. Загрузить набор данных nsw74psid1 из пакета DAAG. Построить регрессионное дерево для модели, задаваемой следующей формулой: **re78 ~.**.
3. Загрузить набор данных Lenses Data Set из файла Lenses.txt:
4. 3 класса (последний столбец): 1 : пациенту следует носить жесткие контактные линзы, 2 : пациенту следует носить мягкие контактные линзы, 3 : пациенту не следует носить контактные линзы.
5. Признаки (категориальные):
   1. возраст пациента: (1) молодой, (2) предстарческая дальнозоркость, (3) старческая дальнозоркость
   2. состояние зрения: (1) близорукий, (2) дальнозоркий
   3. астигматизм: (1) нет, (2) да
   4. состояние слезы: (1) сокращенная, (2) нормальная
6. Построить дерево решений. Какие линзы надо носить при предстарческой дальнозоркости, близорукости, при наличии астигматизма и сокращенной слезы?
7. Для построения классификатора используйте заранее сгенерированные обучающие и тестовые выборки, хранящиеся в файлах svmdata4.txt, svmdata4test.txt.
8. Разработать классификатор на основе дерева решений для данных Титаник (Titanic dataset).

# Ход работы

**1 задание**

Для обучающего множества Glass построено дерево решений, изображённое на Рис.1. Также, было выявлено, что приведённый пример принадлежит второму классу с вероятностью 0,875.

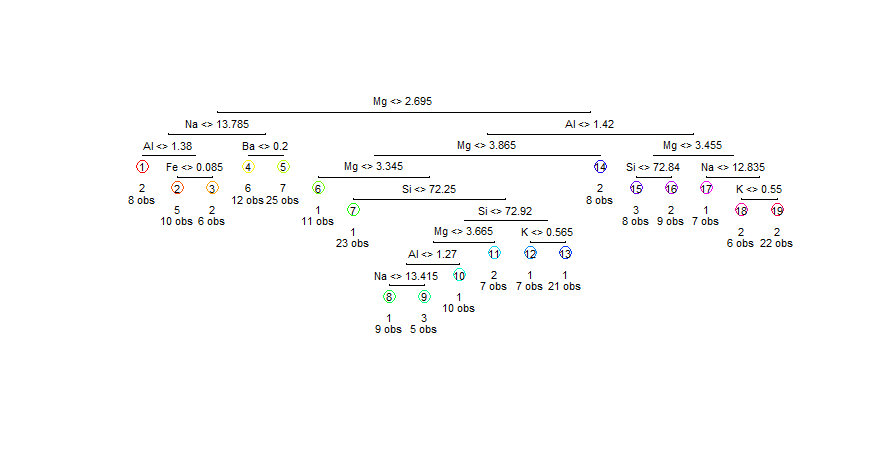


Рис. 1. Дерево решений для обучающего множества Glass

После применения функции snip(), которая удалила 2 вершины, дерево приняло вид, который можно посмотреть на Рис.2. Следует отметить, что результатом функции prune() также будет полученное дерево.

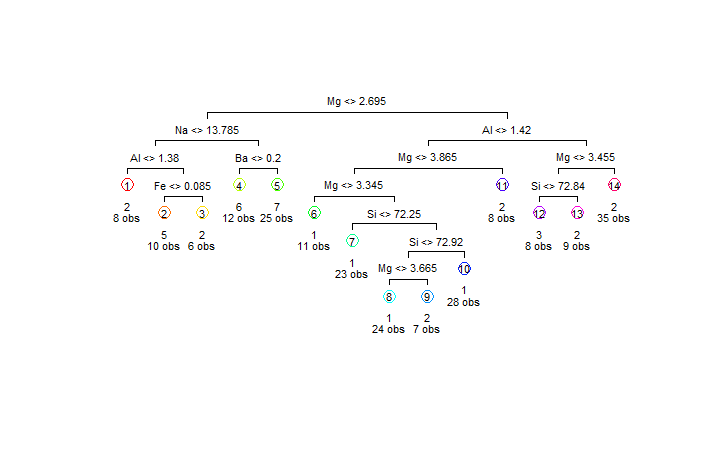


Рис. 2. Усеченное дерево решений для множества Glass

**2 задание**

Дерево классификации для модели, задаваемой формулой yesno**~.,** представлено на Рис.3.

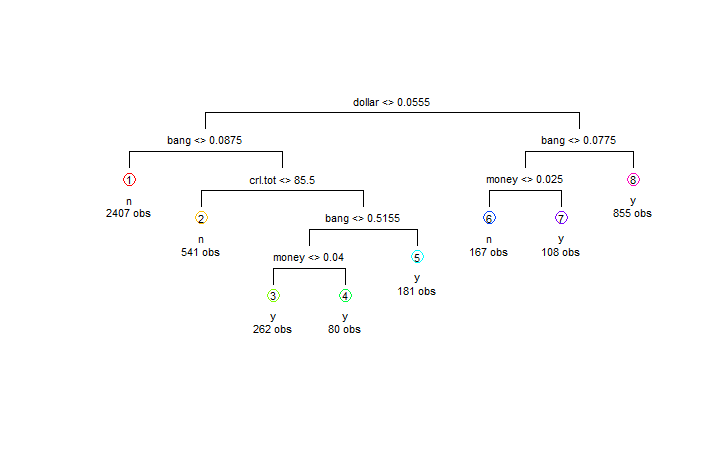


Рис. 3. Дерево решений для примера «Спам»

Применив функцию cv.tree(), и построив график функции, получаем изображение, представленное на Рис.4. Анализируя полученный график, можно сделать вывод, что наилучшим выбором для количества терминальных вершин будет значение от 5 до 7. Опытным путём выяснено, что значения 5 и 6 являются оптимальными для данной выборки.

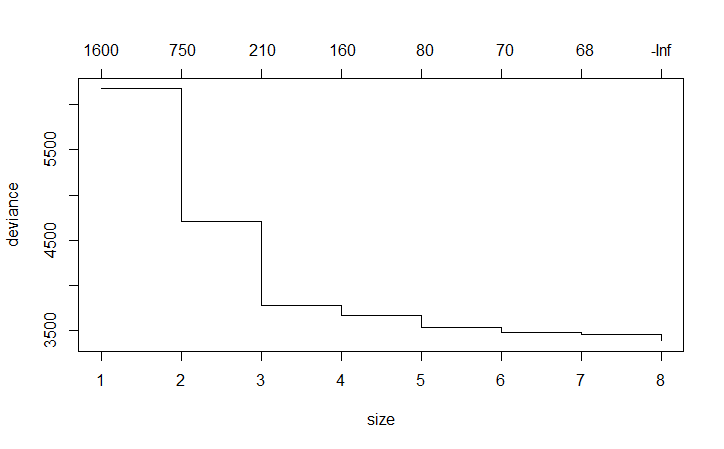


Рис. 4. График зависимости числа ошибок классификации от размера дерева

Таким образом, преобразованное дерево выглядит следующим образом (см. Рис. 5).

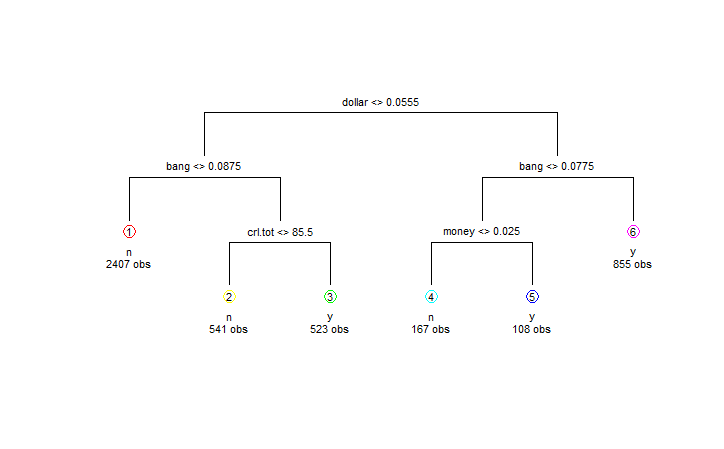


Рис. 5. Преобразованное дерево решений для примера «Спам»

**3 задание**

Для загруженного набора данных построено дерево решений, изображенное на Рис.6.

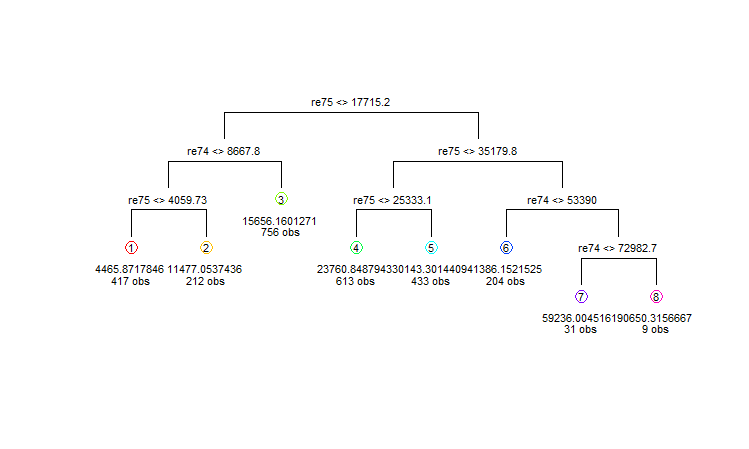


Рис. 6. Дерево решений для примера «nsw75psid1»

Полученное дерево обладает следующими характеристиками:

Regression tree:

tree(formula = re78 ~ ., data = nsw74psid1)

Variables actually used in tree construction:

[1] "re75" "re74"

Number of terminal nodes: 8

Residual mean deviance: 106700000 = 2.847e+11 / 2667

Distribution of residuals:

Min. 1st Qu. Median Mean 3rd Qu. Max.

-59240.0 -4466.0 -615.2 0.0 4423.0 105500.0

**4 задание**

Для построения дерева решений (см. Рис.7) использован модифицированный набор данных: метки класса переписаны из integer в string. Также, заданный набор признаков был определён с помощью построенного классификатора как 3 класс.

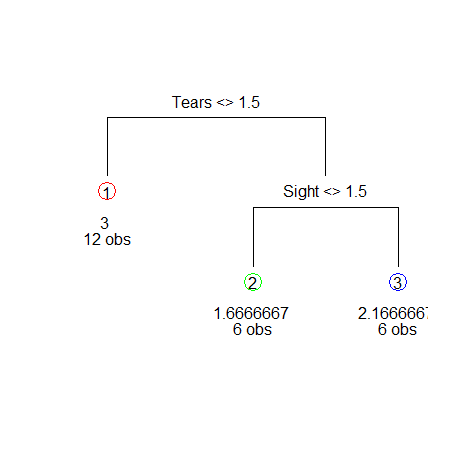


Рис. 7. Дерево решений для примера «Lenses»

**5 задание**

Для построения классификатора были использованы заранее сгенерированные наборы данных. Построенное дерево решений можно увидеть на Рис.8. Кроме этого, найдена величина ошибочной классификации, которая составила 0,1.

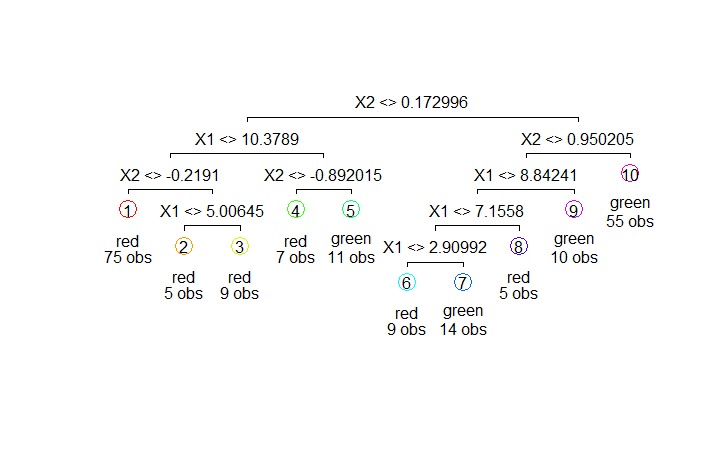


Рис. 8. Дерево решений для примера «SVM»

**6 задание**

Для построения классификатора были выполнены следующие действия:

1. Загрузка данных в соответствующие структуры.
2. Удаление первого признака из исходного набора данных, т.к. он не несет никакой информационной нагрузки.
3. Удаление признаков: Имя, Билет, Кабина – т.к. при построении дерева количество уровней ограничено значением 32.
4. Построение дерева решений на полученной обучающей выборке.
5. На основе тестирующей выборки, поиск значения величины ошибочной классификации.

По результатам проведённого исследования, было построено дерево решений (см. Рис. 9), а также найдена величина ошибочной классификации, которая составила 0,33.

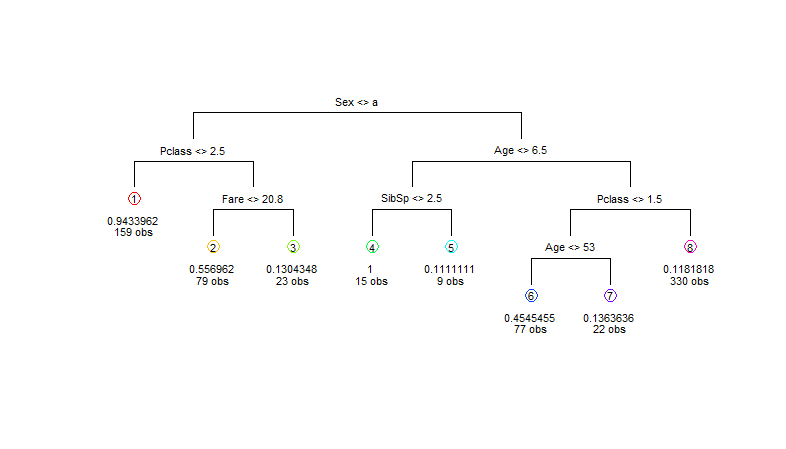


Рис. 9. Дерево решений для примера «Титаник»