CAP-5615 Intro to Neural Networks Summer 2018
Homework 1
Christian Garbin

Question 1 [2 points: 0.25/each]

Please use your own language to briefly explain the following concepts (Must use your own language. No credit if descriptions are copied from external sources):

Machine Learning

Machine learning is the process of programming computers to use sample data or past experience to make sound and reasonable (even if not perfect) decisions about instances it hasn't yet "seen" (of that sample space). It's used when we are not able to directly write a program (a well-defined algorithm) to make the decisions.

Decision Trees

A Decision Tree is a set of rules extracted from a sample data set to make decisions about future instances in that space. The data set is composed of instances, each with a set of features (or attributes). The decision made by the tree can be categorical (classification - predict a class) or numerical (regression - predict a number).

Overfitting

Overfitting is the excessive partitioning of a sample set, resulting in decision trees that work well for the sample set, but increase the error rate for unseen instances (or instances in the test set).

Prepruning and Postpruning for decision tree learning

Pruning is removing nodes from a decision tree to resolve overfitting. Prepuning prevents adding nodes upfront, by stopping partitioning when the data set under analysis (for that branch of tree) is small. Postpruning methodically removes nodes from a tree and tests the results of the pruned tree against the unpruned tree, removing the nodes that don't contribute to accuracy.

Information Gain

Information Gain is the difference between the entropy of the entire sample set and the weighted entropy of the subsets defined by selecting an attribute of the sample set. A higher information gain indicates an attribute that is useful for differentiating between classes and therefore a good candidate to be used to partition a set.

Information Gain Ratio

Information Gain Ratio is a method to avoid information gain bias, which is caused by selecting an attribute that has a high Information Gain, but is biased or not accurate. The Information Gain Ratio applies a "penalty" to such attributes to reduce the probability that they take a key role in the decision tree. The "penalty" is defined in terms of Split Information - the higher this value, the more the attribute is "penalized".

Gini-Index

Gini Index is a measure if impurity of elements (instances) in a set, based on the relative frequency of each class in that set. The lower the Gini Index, the more pure the set is.

Explain why ID3 decision trees are said to be unstable, and why ID3 decision trees are said to be robust to errors:

ID3 decision trees are said to be unstable because a small change to the sample set may result in a large change in the tree. For example, adding a new element to the sample set may cause the root node to change.
The ID3 decision trees are robust to errors because they don't make implicit assumptions about statistics model of the sample set and can cope with missing information data (it will use whatever information is available to make calculate Information Gain to build the tree).

Question 2 [2 pts]

The following figure shows a toy dataset with two numeric attributes/features (x1 and x2) and nine types of instances (color coded with different shapes).
The sub-figure on the right panel shows a constructed decision tree from the toy dataset. Please explain

Roles of interior nodes vs leaf node of the decision trees [0.5 pt].

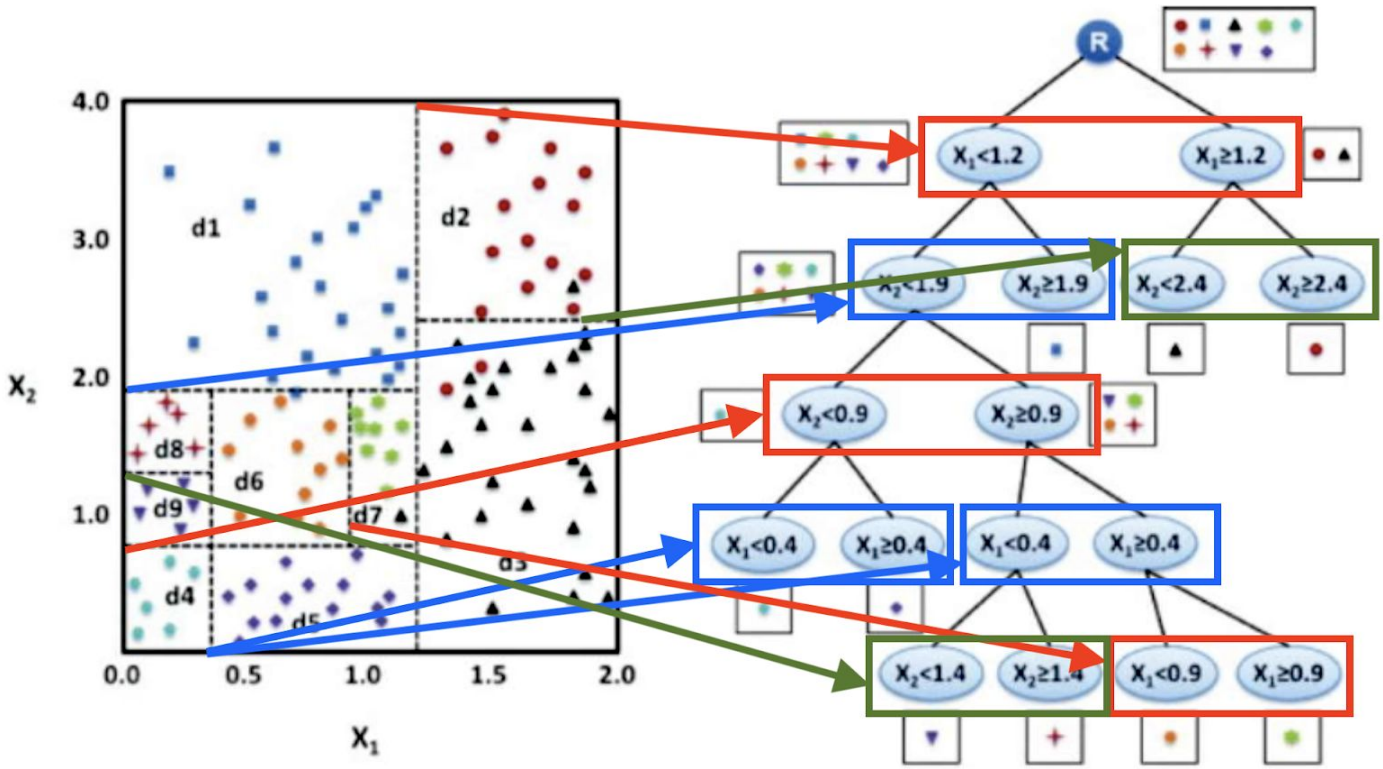
The internal nodes are the decision (or test) points. The leaf node represents the final classification outcome (the decision) reached after following the branches defined in the internal nodes.

Explain the process of building a decision tree though recursive partitioning of the feature space [1.0 pt].

- Select the attribute (or region) to split the sample set, based on purity (higher purity is desired, but must be balanced against bias)
- Decide if a node is a leaf node (i.e. decide if it's time to stop splitting) finding a balance between the accuracy of the tree and excessive partitioning that could lead to bias
- If it's a leaf node: assign a class to the leaf node, based on the instances of the sample set that map to this leaf node
- If it's not a leaf node: repeat the attribute selection process to split the sample set for this node of the tree

What are the meaning of the dashed lines on the left panel, and how does each dashed line correspond to the node on the right panel [0.5 pt].

The dashed lines are partitions of the feature space, classifying elements to construct the nodes of the decision tree. The partitions are used to increase the accuracy of the decision tree (with considerations to avoid overfitting).
The correspondence of dashed lines to nodes are indicated in the picture below.



Question 3 [2 pts]

In database showing in Table 1, please calculate the Entropy of the whole dataset (0.5 pt). Use information gain to determine which attribute has the highest Information Gain (1.5 pts) (List major steps)

ID	Outlook	Temperature	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Mild	Normal	Weak	No
14	Rain	Hot	High	Strong	Yes
15	Rain	Mild	High	Strong	No

Table 1

Entropy of the whole data set

$$Entropy(S) = -p_+ \log_2 p_+ - p_- log_2 p_- = -\frac{9}{15} log_2 \frac{9}{15} - \frac{6}{15} log_2 \frac{6}{15} = 0.971$$

Highest Information Gain

Attribute with highest Information Gain is *Outlook*.

Information Gain for each attribute shown in the following sections.

Outlook

$S_{sunny} = \{D1^-, D2^-, D8^-, D9^+, D11^+\}$
$S_{overcast} = \{D3^+, D7^+, D12^+, D13^-\}$
$S_{rain} = \{D4^+, D5^+, D6^-, D10^+, D14^+, D15^-\}$

$Entropy(S_{sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$
$Entropy(S_{overcast}) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{4}{5} = 0.811$
$Entropy(S_{rain}) = -\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6} = 0.918$

$Conditionalentropy = \frac{5}{15} * 0.971 + \frac{4}{15} * 0.811 + \frac{6}{15} * 0.918 = 0.907$
$InformationGain = 0.971 - 0.907 = 0.064$

Temperature

$S_{hot} = \{D1^-, D2^-, D3^+, D14^+\}$
$S_{mild} = \{D4^+, D8^-, D10^+, D11^+, D12^+, D13^-, D15^-\}$
$S_{cool} = \{D5^+, D6^-, D7^+, D9^+\}$

$Entropy(S_{hot}) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$
$Entropy(S_{mild}) = -\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7} = 0.985$
$Entropy(S_{cool}) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.811$

$Conditionalentropy = \frac{4}{15} * 1 + \frac{7}{15} * 0.985 + \frac{4}{15} * 0.811 = 0.943$
$InformationGain = 0.971 - 0.943 = 0.028$

Humidity

$S_{high} = \{D1^-, D2^-, D3^+, D4^+, D8^-, D12^+, D14^+, D15^-\}$
$S_{normal} = \{D5^+, D6^-, D7^+, D9^+, D10^+, D11^+, D13^-\}$

$Entropy(S_{high}) = -\frac{4}{8}log_2\frac{4}{8} - \frac{4}{8}log_2\frac{4}{8} = 1$
$Entropy(S_{normal}) = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} = 0.863$

$Conditionalentropy = \frac{8}{15} * 1 + \frac{7}{15} * 0.863 = 0.936$
$InformationGain = 0.971 - 0.936 = 0.035$

Wind

$$S_{weak} = \{D1^-, D3^+, D4^+, D5^+, D8^-, D9^+, D10^+, D13^-\}$$
$$S_{strong} = \{D2^-, D6^-, D7^+, D11^+, D12^+, D14^+, D15^-\}$$

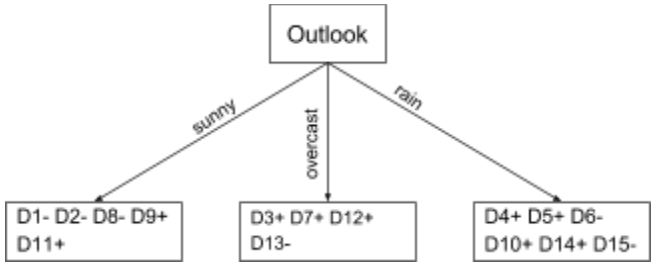
$$Entropy(S_{weak}) = -\tfrac{5}{8}log_2\tfrac{5}{8} - \tfrac{3}{8}log_2\tfrac{3}{8} = 0.954$$
$$Entropy(S_{strong}) = -\tfrac{4}{7}log_2\tfrac{4}{7} - \tfrac{3}{7}log_2\tfrac{3}{7} = 0.985$$

$$Conditionalentropy = \tfrac{8}{15} * 0.954 + \tfrac{7}{15} * 0.985 = 0.968$$
$$InformationGain = 0.971 - 0.968 = 0.003$$

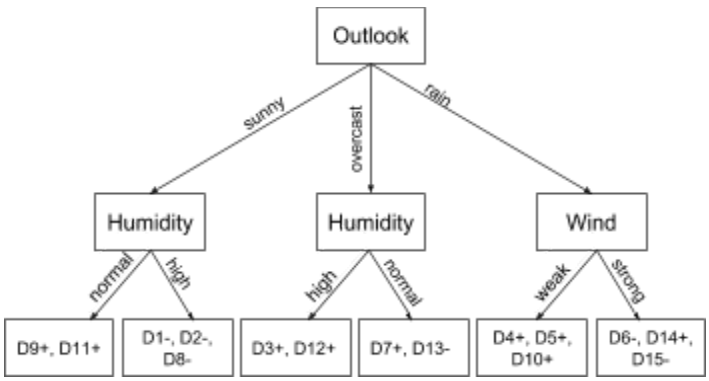
Question 4 [3 pts]

In the dataset showing in Table 1, please manually construct a decision tree by using Information Gain Ratio as the attribute selection criteria (List major steps of the tree constructions [2 pts], and report the final decision tree [1 pt])

1) *Outlook* selected as the root node, based on information gain (question 3)
2) This partitions the sample set into three subsets, one for each value of *Outlook*



3) The Information Gain for each of the nodes result in selecting these attributes for the next nodes:
   3.1) Sunny: Humidity, with a conditional entropy of 0 (zero)
   3.2) Overcast: all attributes have the same conditional entropy (0.5), picked Humidity
   3.3) Rain: Wind, with a conditional entropy of (0.459)



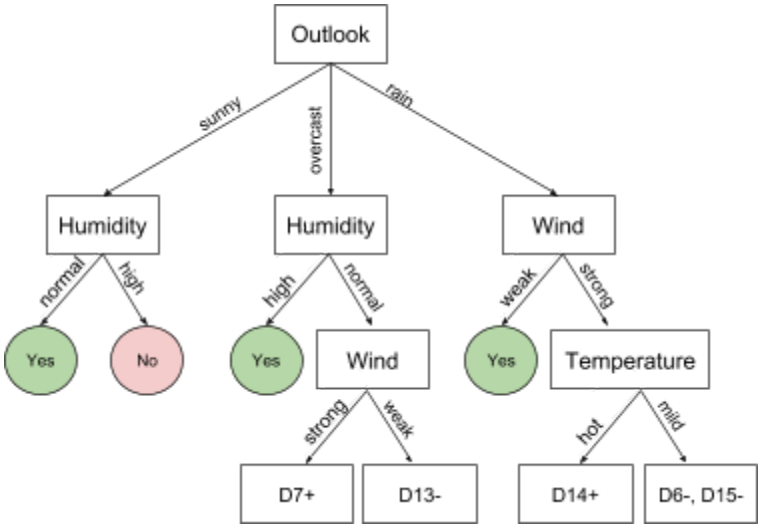
4) Sunny branch: Humidity splits the sample set into instances of the same class. No further split is needed for this branch.

5) Overcast → High: all instances are of the same class, no further split needed.

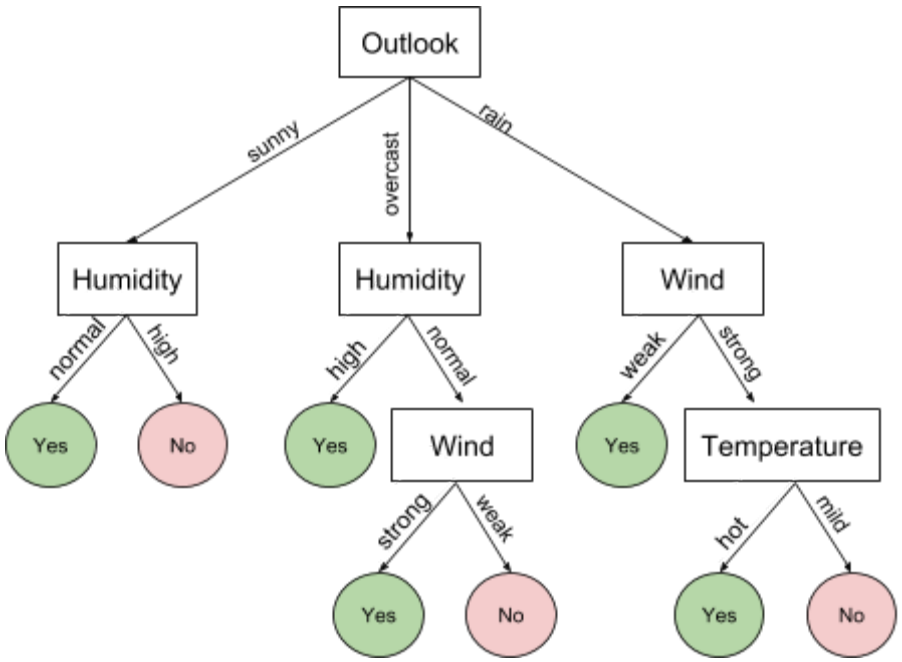
6) Overcast → Normal: *Temperature* and *Wind* same the same entropy (zero). Picked *Wind*.

7) Rain → Weak: all instances of the same class, no further split needed.

8) Rain → Strong: *Temperature* has the lowest conditional entropy (zero).



9) After this split all nodes are of the same class, so no further spit is needed and the tree is completed.



Question 5 [3 pts]

In the dataset showing in Table 1, please use Gini Index to calculate the correlation between each of the four attributes (outlook, temperature, humidity, wind) to the Class label, respectively [2 pts]. Please rank and select the most important attribute to build the root node of the decision tree [1 pt].

Gini Index for Outlook

$$Gini(Overcast) = 1 - ((\frac{1}{4})^2 + (\frac{3}{4})^2) = 0.375$$
$$Gini(Rain) = 1 - ((\frac{2}{6})^2 + (\frac{4}{6})^2) = 0.444$$
$$Gini(Sunny) = 1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2) = 0.480$$
$$Gini(Outlook) = \frac{4}{15}Gini(Overcast) + \frac{6}{15}Gini(Rain) + \frac{5}{15}Gini(Sunny) = 0.438$$

Gini Index for Temperature

$$Gini(Cool) = 1 - ((\frac{1}{4})^2 + (\frac{3}{4})^2) = 0.375$$
$$Gini(Hot) = 1 - ((\frac{2}{4})^2 + (\frac{2}{4})^2) = 0.500$$
$$Gini(Mild) = 1 - ((\frac{3}{7})^2 + (\frac{4}{7})^2) = 0.490$$
$$Gini(Temperature) = \frac{4}{15}Gini(Cool) + \frac{4}{15}Gini(Hot) + \frac{7}{15}Gini(Mild) = 0.462$$

Gini Index for Humidity

$$Gini(High) = 1 - ((\frac{4}{8})^2 + (\frac{4}{8})^2) = 0.500$$
$$Gini(Normal) = 1 - ((\frac{2}{7})^2 + (\frac{5}{7})^2) = 0.408$$
$$Gini(Humidty) = \frac{8}{15}Gini(High) + \frac{7}{15}Gini(Normal) = 0.457$$

Gini Index for Wind

$$Gini(Strong) = 1 - ((\frac{3}{7})^2 + (\frac{4}{7})^2) = 0.490$$
$$Gini(Weak) = 1 - ((\frac{3}{8})^2 + (\frac{5}{8})^2) = 0.469$$
$$Gini(Wind) = \frac{7}{15}Gini(Strong) + \frac{8}{15}Gini(Weak) = 0.479$$

Rank and selection of the Gini Indices

$$Gini(Outlook) < Gini(Humidity) < Gini(Temperature) < Gini(Wind)$$

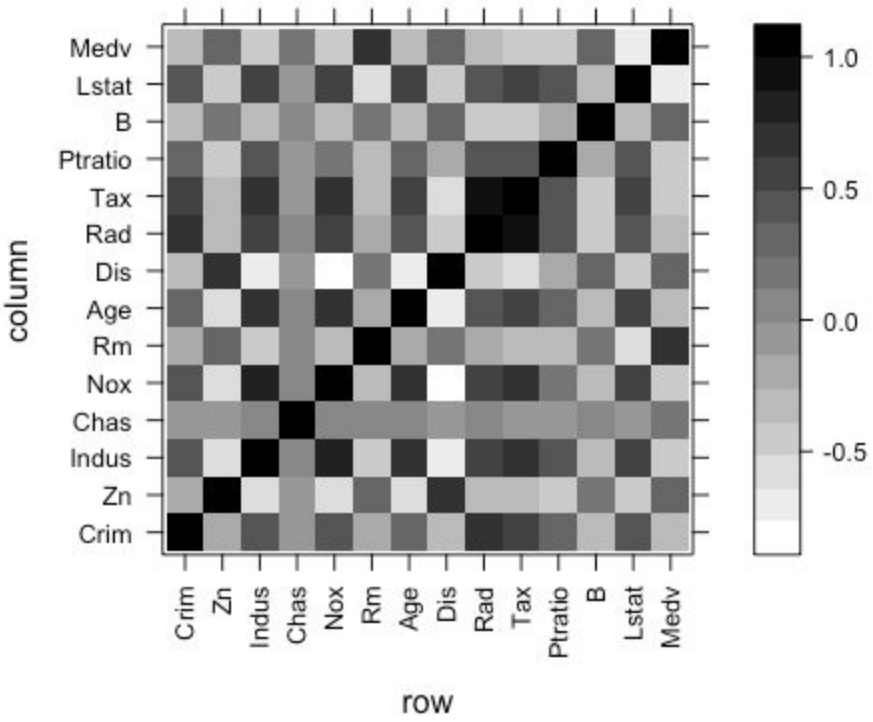
Outlook is the most important attribute, and thus the root node for the decision tree.

Question 6 [2 pts]

Please download housing.header.txt dataset from Canvas, and use R to implement tasks below (a brief description of this dataset is available from the following URL: https://archive.ics.uci.edu/ml/datasets/housing)

Report the pairwise correlation between every two variables (either as a matrix or as a level plot) [0.5 pt].

R code is in file cap5615-homework1-question6-part1.R



Please explain which variable is mostly positively correlated to Medv (medium house value), and which variable is mostly negatively correlated to Medv. [0.5 pt]

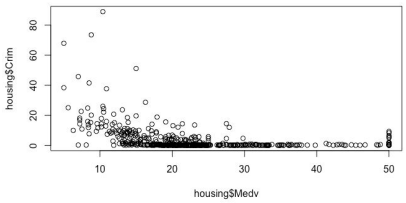
The most positively correlated variable is Rm, the darkest cell in the plot (in the Medv row)

The most negatively correlated variable is Lstat, the lightest cell in the plot (in the Medv row)

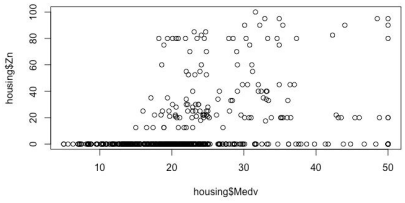
Draw scatterplots to show relationship between each attribute and Medv, respectively [0.5 pt].

R code is in file cap5615-homework1-question6-part3.R

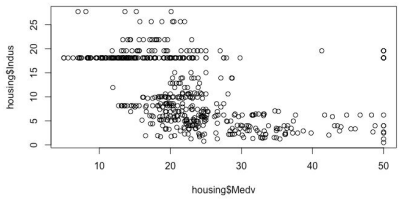
Crim



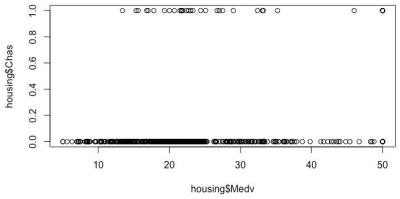
Zn



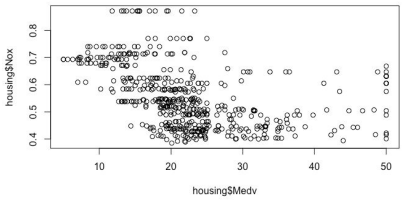
Indus



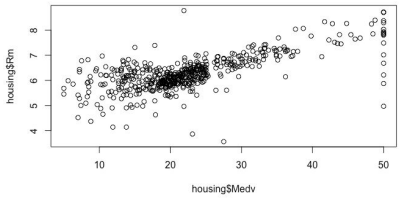
Chas



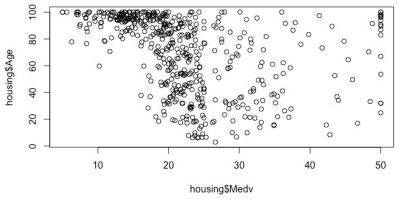
Nox



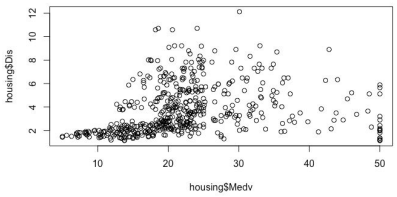
Rm



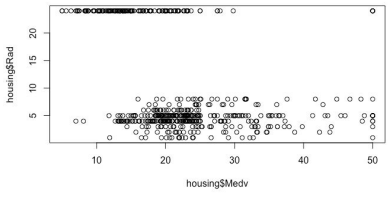
Age



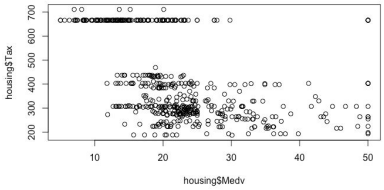
Dis



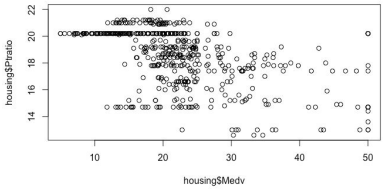
Rad



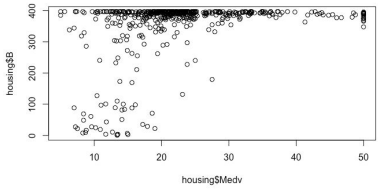
Tax



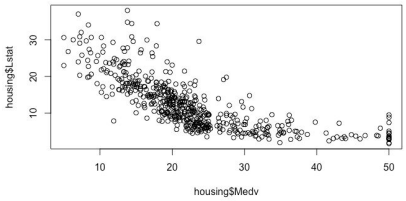
Ptratio



B



Lstat

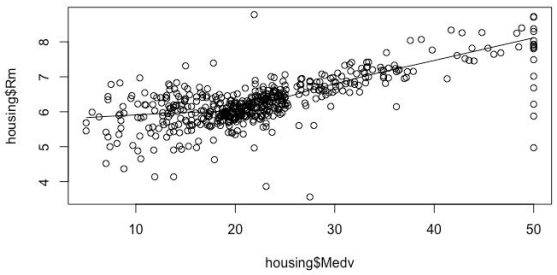


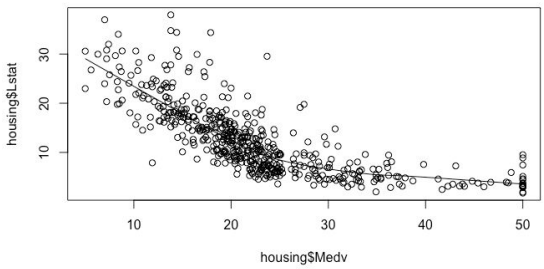
Explain how to use scatterplots to find attributes which are positively correlated, negatively correlated, or independent of Medv, respectively [0.5 pt]

R code is in file cap5615-homework1-question6-part4.R

The correlation can be found from the distribution of the points in the scatterplot. Points clustered around one of the diagonals of the plot indicate a correlation.

For example, Rm shows a positive correlation, Lstat shows a negative correlation, while the other graphs don't show a correlation. The correlation can be more evident with a fitting curve (examples below use the basic `lowess` and `plot` functions in R - more sophisticated packages exist that can illustrate the relationship with richer details).





Question 7 [2 pts]

Please download housing.header.binary.txt dataset from Canvas, and use R to implement tasks below (This dataset is the same as housing.header.txt, except that its Medv value is binarized with Medv value of the house greater than 200k being 1, or 0 otherwise.)

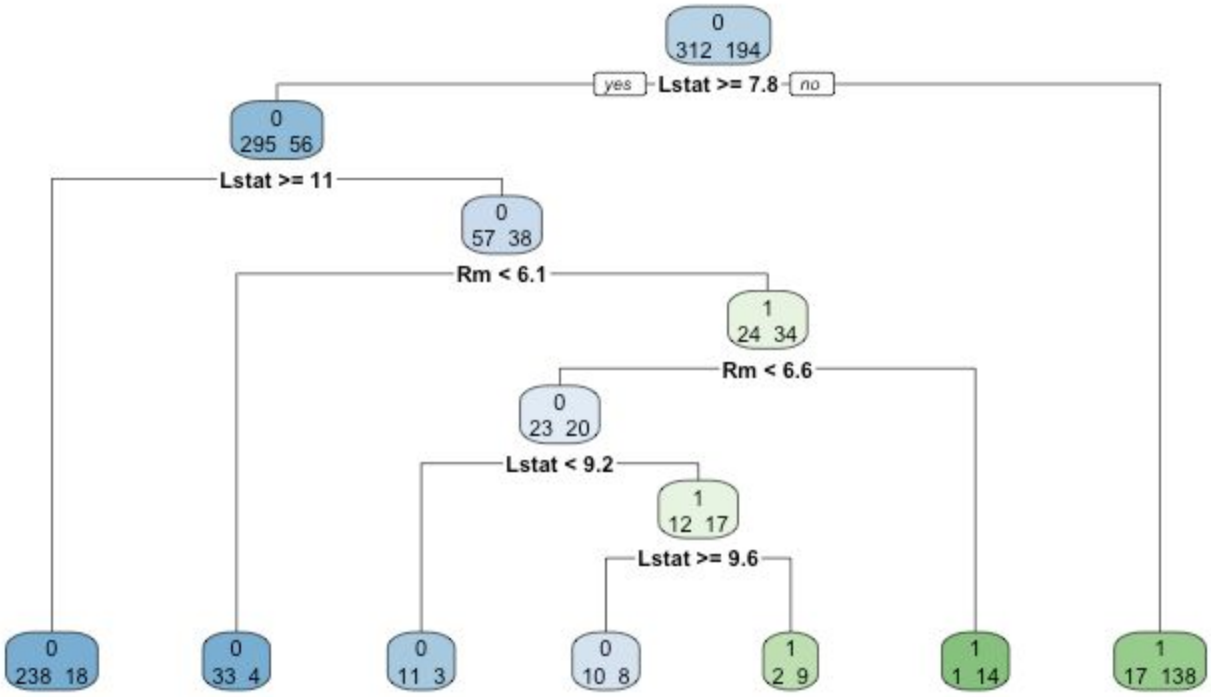
Use the most positively correlated and the most negatively correlated attributes from Question 6 to train an RPART decision tree (0.5 pt) to predict whether a house value Medv is 1 or 0 (i.e., whether the house value is greater than 200k or not). Visualize the tree, and explain the structure of the tree, and the meaning of each node (including leaf nodes) (1 pt).

R code is in file cap5615-homework1-question7.R

The structure of the tree:

- The root node partitions on the Lstat attribute, with value 7.8.
- All samples with an Lstat value less than than 7.8 are classified as "1", on the right branch of the root node.
- All samples with an Lstat value equal to or greater than 7.8 go to the left branches of the root node, where more tests are executed.
- The internal nodes on the left side add more tests based on Lstat and Rm to classify the samples.

- From those nodes we have these decisions in the tree, where each test is an internal node of the left-side subtree (Lstat < 7.8):
    - Lstat < 11 = 0
    - (Lstat >= 11) ^ (Rm < 6.1) = 0
    - (Lstat >= 11) ^ (Rm >= 6.1) ^ (Rm >= 6.6) = 1
    - (Lstat >= 11) ^ (Rm >= 6.1) ^ (Rm < 6.6) ^ (Lstat < 9.2) = 0
    - (Lstat >= 11) ^ (Rm >= 6.1) ^ (Rm < 6.6) ^ (Lstat >= 9.2) ^ (LStat < 9.6) = 0
    - (Lstat >= 11) ^ (Rm >= 6.1) ^ (Rm < 6.6) ^ (Lstat >= 9.2) ^ (LStat >= 9.6) = 1
- Each leaf node shows the predicted class (top value) and how many instances of the training data are correctly and incorrectly classified under that branch of the tree (bottom values). The number on the left side are the one for the "0" class and the one on the right is for the "1" class.



Save the tree as a PS file (0.5 pt), and include the PS file in your final homework submission.

Please see file medv-tree.ps