HW assignment 1

- This assignment will count towards 8 percent of your final grade.
- It is to be done individually
- Due: Sept. 25th

Tasks:

Your goal is to develop data-mining models for predicting the median house prices in the city of Boston based on the data file BostonHousing.XLSX.

The file contains information on over 500 census tracts in Boston, where for each tract 14 variables are recorded. The last column (CAT.MEDV) was derived from the MEDV (standing for median Value), such that it obtains the value 1 if MEDV > 30 and 0 otherwise. Your goal is to develop a suitable data mining model to predict the CAT.MEDV of a tract, given the information in the first 13 columns (see column description below).

Please submit your code, as well as a short (no longer than 1 page) summary of your work.

1) Develop a prediction model using the decision tree algorithm. (2 points)
2) Use KNN prediction using different values of k. Choose the most appropriate value of K. (2 points)
3) Choose the best model among the decision tree and k-NN using a suitable metric. (2 points)

2) Now use the best model you have determined to be superior to predict whether the following house is above median or below (CATV.Median), also in the newBostonhouse.xlsx file (2 points)

CRIM 0.2
ZN 0
INDUS 7

CHAS 0
NOX 0.538
RM 6
AGE 62
DIS 4.7
RAD 4
TAX 307
PTRATIO 21
B 360
LSTAT 10

Housing Values in Suburbs of Boston – Variable meaning

The medv variable is the target variable.

Data description

The Boston data frame has 506 rows and 14 columns.

This data frame contains the following columns:

crim
per capita crime rate by town.
zn
proportion of residential land zoned for lots over 25,000 sq.ft.
indus
proportion of non-retail business acres per town.
chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox
nitrogen oxides concentration (parts per 10 million).
rm
average number of rooms per dwelling.
age
proportion of owner-occupied units built prior to 1940.
dis
weighted mean of distances to five Boston employment centres.
rad
index of accessibility to radial highways.
tax
full-value property-tax rate per $10,000.
ptratio
pupil-teacher ratio by town.
black
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
lstat
lower status of the population (percent).
medv
median value of owner-occupied homes in $100