Department of Statistics

STATS 784: Data Mining

Assignment 1 2017

Question 1

Consult the references in Lecture 1 and identify and describe two applications of data mining technology. Write about a page for each one.

Question 2

The dataset for this question contains house sale prices for King County, Washington State, which includes Seattle. It includes homes sold between May 2014 and May 2015.
The data set contains information on 21613 houses, with variables price (the sales price of the house) and 18 other variables, plus an ID for each house.
The variables are

Variable	Definition	Type
id	House ID	Numeric
date	Date house was sold	String
price:	Price is prediction target	Numeric
bedrooms:	Number of Bedrooms/House	Numeric
bathrooms:	Number of bathrooms/bedrooms	Numeric
sqft_living:	square footage of the home	Numeric
sqft_lot:	square footage of the lot	Numeric
floors:	Total floors (levels) in house	Numeric
waterfront:	House which has a view to a waterfront	Numeric
view:	Has been viewed	Numeric
condition:	How good the condition is (Overall)	Numeric
grade:	overall grade given to the housing unit, based on King County grading system: Numeric	
sqft_above:	square footage of house apart from basement	Numeric
sqft_basement:	square footage of the basement	Numeric
yr_built:	Built Year	Numeric
yr_renovated:	Year when house was renovated	Numeric
zipcode:	Zip code	Numeric
lat:	Latitude coordinate	Numeric
long:	Longitude coordinate	Numeric
sqft_living15:	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area	Numeric
sqft_lot15:	lotSize area in 2015(implies-- some renovations)	Numeric

Using linear regression, construct a formula for predicting the price of a house from the other variables. Calculate estimates of prediction error for your predictor using the R330 functions **`cross.val`** and **`err.boot`**, as well as

the functions **`crossval`** and **`bootpred`** in the bootstrap package, and the function **`train`** in the caret package.

How accurate do you think your estimates are?

Note: the R330 function **`cross.val`** produces only printed output, it does not return a value. Code for a modified version returning a value for the estimate and its standard error is in the code handout for Lecture 2 on the web page.

Points to note: Although some variables are numeric, it makes no sense to include them in the regression as numeric variables (e.g. zip). Also, it may be advantageous to transform some variables to make them more symmetric.

Email answers to me by Friday August 4.