UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Group Assignment

The file ames.xls is a dataset describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. It was compiled by Dean De Cock, and documented in an article appeared in the Journal of Statistics Education, Volume 19, Number, 3 (2011).

The data set contains 1460 observations on 80 variables. The last column is the Sale Price, which is the output variable we aim at predicting. There are 34 quantitative house characteristics relating to the dimension of the house and its age. There are also a few categorical variables associated with this data set with a varying number of categories. The full description of the variables is in the file data_description.pdf.

1 Reading the data and preliminary operations

1.1 Remove Missing Data

Some of the variables contain a large number of missing data. The following statements import the data and delete the columns with missing values, returning a complete dataset. In R You should copy and paste them as they are.

```
ames = read.csv("ames.csv", stringsAsFactors = F)
# drop variables with missing obs
ames = ames[sapply(ames, function(x) !any(is.na(x)))]
sapply(ames, function(x) sum(is.na(x)))
```

1.2 Training and Test samples

As your training sample, select randomly $N = 1,000$ observations from the original dataset, using the following statements (copy and paste in your R script). The `seed` for the sample selection (which is the argument of the `set.seed()` function), must be set equal to the eight digits number making up the registration number (matricola) of the representative member of the group.

```
Ntot = nrow(ames)
set.seed(28021999); # set seed here by inserting your matricola (here 28021999)
N = 1000;
s = sample(1:Ntot, N);
ames.train = ames[s,];  # training sample of size N
ames.test = ames[-s,];  # test sample of size Ntot-N
summary(ames.train);
attach(ames.train)
```

2 Analytical tasks

You should address the following points:

1. Perform graphical and correlation analyses to explore the main features, associations and dependencies in the dataset. Be aware of the scale of measurement of the variables (qualitative vs quantitative).

   Useful functions and examples (list not exhaustive):

   ```
   boxplot(SalePrice~ factor(HouseStyle)) # conditional boxplots
   cmeans = tapply(SalePrice,HouseStyle,mean) # conditional means
   X =  data.frame(SalePrice, x1stFlrSF,x2ndFlrSF, TotalBsmtSF, GrLivArea, YearBuilt)
   pairs(X)
   cor(X)
   ```

2. Consider the problem of predicting the output variable `SalePrice` using the following candididate input variables: `x1stFlrSF,x2ndFlrSF, LotArea, FullBath, PoolArea, GarageCars, TotRmsAbvGrd, KitchenAbvGr, GrLivArea, BedroomAbvGr, YearRemodAdd, OverallCond, OverallQual, Street, Alley, Age`, where `Age = 2010-YearBuilt`.

   - Estimate the full model (containing all the input variables) and comment about the results. What effects are not significant at the 5% level?
   - Perform variable selection using forward stepwise selection, and discuss the results. Compare the model selected according to the minimum Mallow's C_p criterion with the full model specification. Any interesting results?
   - Would the model selected by forward stepwise change if you use BIC as your estimate of the test error?
   - Estimate the test error of the two rival specifications using the test set `ames.test`. Which one delivers the smallest test error? To get the out-of-sample predictions use the function `predict(model, newdata = ames.test)`, where `model` is the specification under investigation.
   - Extract the first 3 principal components from the quantitative input variables listed above (i.e., drop the qualitative variables), after a standardization. Are they interpretable? What R^2 do you get when you regress `SalePrice` on these components?

3. Consider the nominal variable taking value 1 if `OverallCond` is from `above average` to `excellent`. This is constructed as `Y = I(OverallCond>5)`, which returns a binary variable with 2 categories, `TRUE` and `FALSE`. By estimating a logistic regression model, assess the role of the dimension of the house, measured by `Size = x1stFlrSF+ x2ndFlrSF+TotalBsmtSF`, the age of the house, constructed as `Age = 2010-YearBuilt`, and the house characteristics `HeatingQC`, `CentralAir`, `KitchenQual`, `Fireplaces`, `PavedDrive`, in explaining `Y`.

   Comment on the results and on the goodness of fit as measured by the deviance and the missclassification rate. Finally, evaluate the missclassification rate in the test sample.

4. Estimate the conditional mean function of `SalesPrice` given `Age` in the training sample by local polynomial regression using the Nadaraya-Watson estimator. Estimate the optimal bandwidth of the NW estimator by adapting the code provided during the course (Inflation.R). Compare the fit with that of a smoothing spline with smoothness parameter selected by cross-validation. What is the estimated λ? What is the corresponding number of degrees of freedom?

3 Deliverables

This assignment can be carried in groups of maximum 4 students. It is due by 11:00 p.m. of October 30. The representative member of the group should send a zip file containing:

1. the report in pdf format. The filename should contain all the surnames of the members of the group (Surname1_..._Surname4.pdf).
2. The script files (Surname1_..._Surname4.R).

4 Getting started

To get the work started, recall the following steps:

1. Download the file ames.csv in a working directory that you can trace back.
2. Open RStudio
3. Set the working directory from the menu Session → Set Working Directory → Choose Directory
4. Open a new script that contains your R statements. From the menu File → New → R Script
5. Copy and paste the following lines:

```
ames = read.csv("ames.csv", stringsAsFactors = F)
# drop variables with missing obs
ames = ames[sapply(ames, function(x) !any(is.na(x)))]
```

6. Enjoy working with the data.
7. Remember to save the script.