Getting started hints - for students

This section provides a few brief hints for the student in how to begin thinking about analysing the data.

Project 1 - Regression Analysis

The problem is essentially a regression problem, where the response variable is `price` and should be regressed against the other variables. Before fitting the model you need to carefuly delete any outliers and judge the validity of assumptions of the model Variable selection and transformation of variables should be examined before proposing the final model. For the advanced chapter you can explore models beyond linear models and least squares fit. For example you can explore Lasso, Ridge regression and other advanced machine learning techniques. You can also make sure the assumptions of linear models are satisfied.

Reading material

- "The element of statistical learning"; J. Friedman et.al., Springer, pages- 79-91, 2008.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).
- Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2002). Least angle regression. Published in Annals of Statistics 2003

Project 2 - Classification

In context of model based classification the problem is essentially a `glm` problem with binary response `asking` price. You can also take a model free approach and treat it as a pure classification problem and try out classification techniques such as k-nearest neighbour, Random Forest, Neural networks and other machine learning techniques. If you need to choose any tuning paramaters for the model free approaches you should provide a complet description of your approach. Variable selection approaches can also be tried. For the advanced chapter you can explore several tools for classification approaches and choose the best one base on the 100 test samples. For the advanced chapter you can also explore Bayesian GLM. GLM lasso

Reading material

- Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. (2009). "A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models." The Annals of Applied Statistics 2 (4): 1360–1383. http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf

- Jerome Friedman, Trevor Hastie and Rob Tibshirani. (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010.

Project 3 - Clusterng

The problem is essentially a clustering problem. You need to evaluate several clustering method, choose the right variables and provide an interpretation for the clusters that you have obtained. Additionally you need to judge the number of clusters. You can start with the basic clustering techniques such as k-means and a range of hierarchical clustering. You can also use model based approaches, such as normal or non-normal mixture models using available packages in R. For the advanced chapter you can explore how to cluster data that have variables that are a mix of discrete, continuous and categorical variables. You can also explore Modal clustering

Reading material

- Li, J., Ray, S. and Lindsay, B.G. (2007) A nonparametric statistical approach to clustering via mode identification. Journal of Machine Learning Research: Proceedings Track, 8, pp. 1687-1723.
- Hennig, C. and Liao, T. F. (2013), How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics), 62: 309-369. doi:10.1111/j.1467-9876.2012.01066.x
- Adrian E Raftery & Nema Dean (2006) Variable Selection for Model-Based Clustering, Journal of the American Statistical Association, 101:473, 168-178, DOI: 10.1198/016214506000000113