

WEEK 4 HOMEWORK

INSTRUCTIONS

- Every learner should submit his/her own homework solutions. However, you are allowed to discuss the homework with each other (in fact, I encourage you to form groups and/or use the forums) – but everyone must submit his/her own solution; you may not copy someone else's solution.
- The homework will be peer-graded.
- The homework grading scale reflects the fact that the primary purpose of homework is learning:

Rating	Meaning	Point value (out of 100)
4	All correct (perhaps except a few details) <u>with</u> a deeper solution than expected	100
3	Most or all correct	90
2	Not correct, but a reasonable attempt	75
1	Not correct, insufficient effort	50
0	Not submitted	0

Question 1

Using the same crime data set as in Homework 3 Question 4, apply Principal Component Analysis and then create a regression model using the first 4 principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Homework 3 Question 4. You can use the R function `prcomp` for PCA. (Note that to first scale the data, you can include `scale. = TRUE` to scale as part of the PCA function.)

Question 2

Using the same crime data set as in Homework 3 Question 4, find the best model you can using (a) a regression tree model, and (b) a random forest model. In R, you can use the `tree` package or the `rpart` package, and the `randomForest` package. For each model, describe one or two qualitative takeaways you get from analyzing the results (i.e., don't just stop when you have a good model, but interpret it too).

Question 3

Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.

Question 4

1. Using the GermanCredit data set at <http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/> (description at <http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>), use logistic regression to find a good predictive model for whether credit applicants are good credit risks or not. Show your model (factors used and their coefficients), the software output, and the quality of fit. You can use the glm function in R. To get a logistic regression (logit) model on data where the response is either zero or one, use family=binomial(link="logit") in your glm function call.
2. Because the model gives a result between 0 and 1, it requires setting a threshold probability to separate between "good" and "bad" answers. In this data set, they estimate that incorrectly identifying a bad customer as good, is 5 times worse than incorrectly classifying a good customer as bad. Determine a good threshold probability based on your model.