Part 1

Attached is a small data set (90 systems, 3 properties, and 25 features). Use this data (data1.csv) to answer the following.

There are three properties in the data. We are trying to identify a system which has large value for Property 1, large value for property 2, and a value of 1 for Property 3 (ie. Property 3 is whether the system can exist or not, with 1 being yes and 0 being no). The meaning of high for Properties 1 and 2 is relative – so can be compared to the other values in the data but does not have a specific cut-off. Keep in mind the discussion in class, where your definition of large should not just be above some arbitrary middle value.

The Properties are unknown for the first 10 systems (shown as blank for the first 10 rows). The question is, do any of these 10 systems meet the requirement just defined? Justify your answer, as to why or why not, and provide your final predicted value for each of the three properties for all 10 of the first systems. You should also provide a figure for each property corresponding to your final model (for example, a plot of predicted vs. actual).

Describe all steps you take in building the model (description should be brief, but should be detailed enough that your answer can be reproduced). For each step, why did you do it what was gained by doing that step? If you do not mention a step, it will be assumed that you did not do it. For each property, you should make your prediction with three (3) techniques covered in class. You should use only three techniques per property (you can use different techniques for different properties), and are limited only to those techniques discussed in class. Describe why you selected these techniques, and if you think this selection is sufficient for having confidence in your answer (ie. have you explored enough different assumptions in the techniques?).

The following are some additional questions to answer:
(i) How confident are you in your prediction, and why? Would your confidence in the prediction increase if you had data for more systems (ie. more rows in the data) and why?

(ii) If we wanted to begin to assess the uncertainty in your models, what is an approach we could follow, based only on what was discussed in class, to estimate the uncertainty, whether qualitative or quantitative?

(iii) Is there any issue with this data in terms of an unbalanced data set? Whether or not there is, how would you recommend addressing issues related with an unbalanced data?

Part 2

A data set “data2.csv” is provided, where the data is continuous with the property is profit and the values are the stock value of the company each day. From this, answer the following:

(i) predict the missing property values (those denoted by a ‘?’). You need to justify your predictions. Justifications for your predictions should include comparison between three different approaches and model accuracy and robustness. Also, describe the steps that you employed prior to the regression step, and why, including which features you included and why. You should provide enough information that your prediction can be reproduced – do not assume that it is obvious the steps you followed.

(ii) How confident are you in the predicted values and why? Also indicate how generalizable each model is, and how sensitive it is to the parameters used in the regression model.