Homework 7 STAT/CS 387: Data Science II
Statistical Learning

Instructions

Provide a typed (not handwritten) write-up that addresses the problems given below. Be sure to show all your work when answering these problems. Please use one of the provided write-up templates. All graphics such as plots should be computer-generated, appropriately labeled including axes labels for plots, and be included in the write-up at their appropriate locations and with figure captions.

This homework includes a dataset covering features of housing markets and neighborhoods in the 1970s. Please familiarize yourself with the data to begin the project.

Coding

The scikit-learn Python module is appropriate for this assignment, and quite helpful due to all the helper and utility functions it provides (*find them and use them!*). You are free to use other code bases, however, such as R. Be sure to include all your code as part of your submission. Learning to use these tools is an important aspect of this assignment.

Grading

You will be graded on how thoroughly and accurately you address the below questions.

To submit

Prepare your write-up as a PDF named `HW07_write-up_[NETID].pdf` (Ex: `HW07_write-up_jbagrow.pdf`). The write-up should address all of the assignment in a clearly organized way. Place all of your (well-organized and readable) code and other work into a directory called `work`. Place `work` and your write-up PDF inside a directory called `HW07_[NETID]` (Ex: `HW07_jbagrow`). Submit this directory as a *zipped file* to Blackboard.

If you are having problems, please let me know. Extra office hours are available by appointment!

- This is a very open-ended assignment and you will need to "reimagine" or re-interpret basic questions to make specific progress. What is *really* being asked, and what questions can *really* be answered with these data? No maps for these territories.

Part 1. **Data exploration**

Use summary statistics, plots, the dataset documentation, and anything else you can think of to identify as many meaningful features, structures, or relationships within the data set. What is in the data? What can we learn about the data that is not obvious? What parts matter (if any) and what parts do not matter (if any)?

Part 2. **Predictive modeling**

One column in the dataset is the **crime rate**. Use statistical learning to develop models to predict this quantity as a function of the other columns in the data. In other words, the response or target is crime and the features or predictors are everything else.

(1) Construct at least one model using a ***linear*** learning method and at least one model using a ***nonlinear*** learning method. Provide a detailed description of all methods used for background and reproducibility. Justify your choice of methods. Compare and contrast the linear and nonlinear models. Beyond prediction, what do we learn from these models? Discuss feature selection/engineering, overfitting, and at least two other important aspects of predictive models as applied to these data.

(2) Using ***cross-validation***, estimate the predictive (out-of-sample) accuracies of your models (cross-validation may also be necessary for the previous step). What is your true predictive accuracy compared with your estimate?

Part 3. **The ethics of prediction**

Machine learning has taken the world by storm. Predictive models are now being applied to problems of social importance. Indeed, in this homework you are studying how to predict future crime levels within a major American city. In general and in specific, what are the ethical concerns, if any, of such predictive methods? Please include in your write-up a short but formal essay (at least 2–3 substantive pages) with bibliography discussing this question.