Predicting power load

Task:

The task is a nonlinear regression: To predict the power load for Puget Sound Power & Light Co. 24 hours in advance, at 8 in the morning, when the current day is a working day and tomorrow is a working day (to make the problem a little bit easier, because things look a little different when tomorrow is a holiday or when the current day is a holiday).

To solve the problem, you get observations for the period January 1985 – October 1990 (in all seasons).

To test your system, I have witheld data from the winter months (November – March) of 1990/1991 and 1991/1992. You will be given the input data corresponding to those months, without information about the correct output, and asked to provide predictions for them.

(It is possible to "cheat" on this task but you are expected to not do this.)

The 15 input variables are:

1. The current power load (MW)
2. Average power load over the last 24 hours (MW)
3. Average power load over the last week (MW)
4. Peak power load during the last 24 hours (MW)
5. Peak power load during the last week (MW)
6. Forecasted temperature 24 hours ahead (Farenheit)
7. The current temperature (Farenheit)
8. Average temperature last 24 hours (Farenheit)
9. Average temperature last week (Farenheit)
10. Variance of the temperature over the last 24 hours (Farenheit^2)
11. Variance of the temperature over the last week (Farenheit^2)
12. Average forecasted temperature for the next 24 hours (Farenheit)
13. Variance of forecasted temperature for the next 24 hours (Farenheit^2)
14. $\cos(2\pi \cdot \text{daynum})$ where daynum = (the number of the day in the year)/365
15. $\sin(2\pi \cdot \text{daynum})$

These inputs are the result of quite a lot of variable selection so you can assume that these variables should all be used.

The output that you shall predict is the load 24 hours ahead.

Data

You are given the file *PowerTrainData.mat*, which contains the following matrices: *powerTrainInput* (15×844), *powerTrainOutput* (1×844), *powerTrainDate* (1×844), and *powerTestInput* (15×115). The *powerTrainDate* is the date for the training observations (in MATLAB datenum format).

Steps and subgoals

1. Get acquainted with the data. Plot the data and try to get a feel for the possible relationships between input and output.
2. Construct a standard linear model and estimate the generalization performance.
3. Construct a linear ridge regression model. Use cross-validation to estimate the ridge regression parameter λ. Estimate the generalization performance.
4. Construct a nearest neighbor regression model. Estimate the generalization performance.
5. Construct a standard multilayer perceptron (MLP) model and estimate the generalization performance.
6. Construct a committee with 10 different multilayer perceptrons (no weight decay) and estimate the generalization performance.
7. Produce the predictions for the test inputs and mail to me (for your best linear model, nearest neighbor, single MLP model and the MLP committee), together with your estimates for how well you will do on the test samples.

   Note: These predictions must be mailed to me no later than 48 hours before your oral presentation of your project.
8. Write a report.

Report and presentation of results

You will present the results from your project in two ways: (1) A written report where the main conlusions are presented together with figures and tables supporting your conclusions. (2) An oral presentation, of about 20 minutes, to your course colleagues.

The report should be about 10 pages, including figures and tables, and should contain the elementary report constituents:

- Introduction (brief presentation of problem, 1 page)
- Methodology (brief listing of methods, 1 page)

- Data (presentation of your data set with important observations, 1-2 pages)
- Results (4-5 pages)
- Discussion (your results and comparison to other researchers' results, 1 page)

The report writing should not take much more than one full day, since you are two persons sharing the work.

When you are finished with your report, and it has been accepted, then you should produce a postscript file with it, and pack it together with your dataset and other important parts of your project (like MATLAB M-files). The idea being that someone else could unpack it and repeat the main steps in your analysis without rewriting everything.