Santa Clara University
Electrical Engineering Department

ELEN 520 Introduction to Machine Learning
Project

1. Objectives

(1) Designing a machine learning project on the PYNQ board.
(2) Fit the data to at least to different models.
(3) Use validation techniques LOOCV and K-Fold validation.
(4) Test for collinearity.
(5) Use the principal components approach.

2. Project

Design a machine learning project using the PYNQ board and at least one sensor and two predictor variables. You will collect data and fit it to two models, then use the techniques discussed in class to identify problems with the model and correct them. Select any two types of model that can relate the response variable to the predictor variables. Use cross-validation to estimate the test error associated with a given learning method. The condition of strong linear relationships among predictor variables is referred to as collinearity. Determine if collinearity is present in the model and what can be done to resolve the difficulties associated with collinearity. Use the principal components and ridge regression methods to analyze the data.

3. Resources

Here are some resources that you can refer to:

ML projects with PYNQ board:
http://www.pynq.io/ml.html

Examples with the PYNQ board interfacing to camera using USB and HDMI-IN
USB:
https://github.com/Xilinx/PYNQ/blob/master/boards/Pynq-Z1/base/notebooks/video/opencv_face_detect_webcam.ipynb
https://github.com/Xilinx/PYNQ/blob/master/pynq/notebooks/common/usb_webcam.ipynb

HDMI:
https://github.com/Xilinx/PYNQ-ComputerVision/blob/master/boards/Pynq-Z1/notebooks/filter2d.ipynb

4. Exercises

1. Devise a hypothesis and state the relevant null and alternative hypothesis.
2. Design an experiment to test your hypothesis with predictors X_1, X_2, ..,X_n and response Y. Set up your experiment using two or more sensors with the PYNQ board. Collect data from the sensors and store in a CSV file. Decide the number of observations and duration over which the data will be collected.
3. Fit the data to two different models.
4. Compute the LOOCV errors from fitting the two models.
5. Obtain the estimate for the test error using K-fold cross validation. What value of K did you choose and why? How does this test error compare with the LOOCV error?
6. Create the following diagnostic plots of the linear regression fit:
    (a) Influence plot
    (b) Pairwise plot
    (c) Studentized residual vs predicted response
    (d) QQ plot for residuals

Are there any problems with the fit? Are there any unusually large outliers? Are there any observations with unusually high leverage? Is there a non-linear association between any of the predictors and response? Is heteroscedasticity present in the model?

7. Compare the correlation matrix of the predictor variables and the corresponding scatter plot matrix. Do you see any evidence of collinearity?
8. Compute the corresponding principal components, their sample variances and the condition number. How many different sets of collinearity exist in the data? What variables are involved in each set?
9. Based on the number of PCs you choose to retain, obtain the PC estimates of the coefficients.
10. Using the ridge method, construct the ridge trace. What value of k to you recommend for estimation of parameters. Compute the ridge estimates of the regression coefficients using this value of k.

5. Submission and Demo (100 points)

Submit a report with programs and outputs for exercises 1 to 10. Give a demonstration of your program to the class. Discuss if there were any problems with the data collected for the experiment and model assumptions. How did you address these problems? Did you detect collinearity? How many PCs did you choose to retain? Which model would you recommend to describe the relationship between the response and the predictor variables?