Problem Set 1
Due Monday, April 16, 2017 at 11:55pm

How to Submit

Create one .zip file (**not** .rar or something else) of your code and written answers and submit it via `ilearn.ucr.edu`. Your zip file should contain `plotdata.m`, `ridgells.m`, `llserr.m`, and a file of your written answers for problems 1, 3, and 4c. Please submit your written answers in a pdf or ascii text file, not an MS Word document.

Each file should include at the top (in comments if necessary)

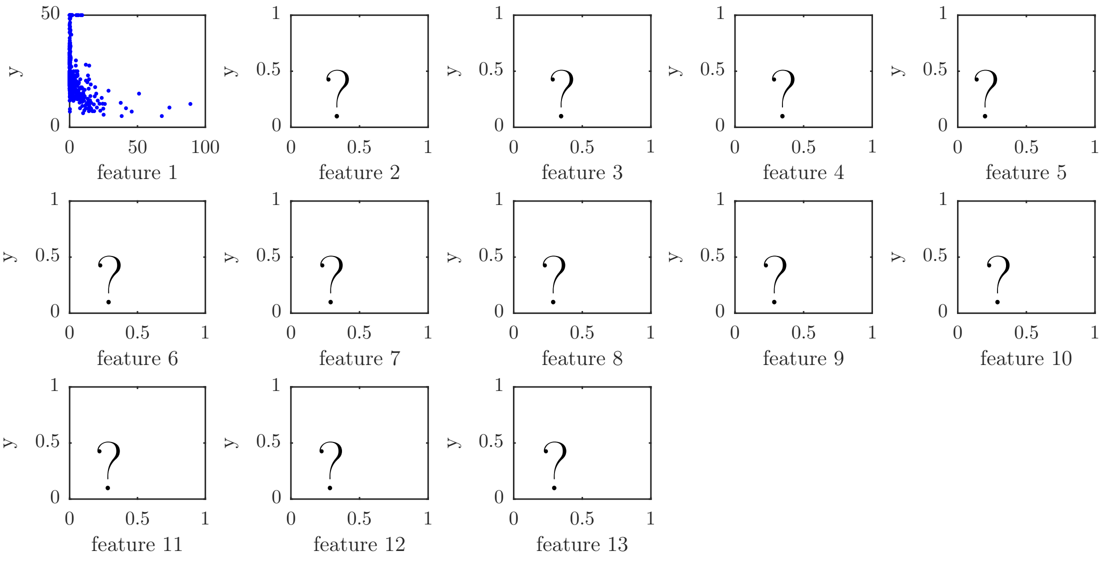
- Your name & UCR student ID number
- The date
- The course (CS 171) & assignment number (PS 1)

Problem 1. [5 pts] During your regular medical check-up, your physician orders a regular blood test (that is, a test she orders for everyone having an annual check-up) to check for foobarinosis, a disease that was only discovered since your last check-up. This test has a false-positive rate of 1% (that is, if you don't have the disease, there is a 1/100 chance that the test will come back positive) and a false-negative rate of 0.2% (that is, if you do have the disease, there is a 2/1000 chance that the test will come back negative). The disease is present in 1 out of 4,000 people.

Your blood test comes back positive. What is the probability you have this disease? (Show your calculations)

Problem 2. [5 pts] Write a Matlab function `plotdata(fname)` that takes as input a filename (as a string), and plots (in a single figure) each feature (column) in that data file against the last column (the y value) as a scatter plot. Each feature should be its own subplot (see the command `subplot`), four to a row.

For instance, when run on the supplied `housetrain.data`, you would get the following (except without the question marks and with the real data in the last five axes).



Problem 3. [5 pts]

Run your function `plotdata` on the `housetrain.data`. What does this plot tell you about the data and prediction in this dataset. Look at the file `housing.names` for information about the features to help your interpretation of the plots.

Problem 4. [10 pts]

The data in `housetrain.data` and `housetest.data` are training and testing data for the task of predicting the value of a house.

part a.

Implement the function `ridgells`, without using similar functions from the Matlab toolboxes. This function should take three arguments, `X`, `Y`, `lambda`, and return `w` and `b`. `w` and `b` are to be the ridge regression fit to the data represented by `X` and `Y` using regularization constant `lambda`. `X` is an m-by-n matrix, where m is the number of examples and n is the number of features. `Y` is an m-by-1 vector of the desired targets. And, `lambda` is a scalar. `w` is a n-by-1 vector of weights for the linear regressor and `b` is the corresponding scalar offset.

part b.

Implement the function `llserr` that computes the mean squared error. It should return a single scalar value (the average squared error) and should take four arguments: `X`, a data matrix of features as above, `Y`, a vector of target output as above, `w` a weight vector as above, and `b` an offset as above. Using `w` and `b`, it should predict from each row in `X` the target, compare to the actual target (in `Y`) and find the squared error. It should report the average squared error over all such rows.

part c.

Run the supplied function `plotacc` that uses the `ridgells` and `llserr` functions you wrote above. What does this plot tell you about ridge regression? Read the code to understand what it is plotting. Think carefully about how you would use ridge regression and what the two curves represent.