Homework 9

AMATH 352, Fall 2022

Due on Dec 12, 2022 at midnight.

Directions, Reminders and Policies

Read these instructions carefully:

- **You are required to upload a PDF report to Canvas along with a zip of your code.**
- The report should be a maximum of **5 pages long** with references included. Minimum font size 10pts and margins of at least 1inch on A4 or standard letter size paper.
- Do not include your code in the report. Simply create a zip file of your main scripts and functions, without figures or data sets included, and upload the zip file to Canvas.
- Your report should be formatted as follows:
  - Title/author: Title of report, your name and email address. This is not meant to be a separate title page.
  - Sec. 1. Introduction and overview of the problem.
  - Sec. 2. Theoretical background and description of algorithms.
  - Sec. 3. Computational Results
  - Sec. 4. Summary and Conclusions
  - References
- I suggest you use LaTeX(Overleaf is a great option) to prepare your reports. A template is provided on Canvas under the Syllabus tab. You are also welcome to use Microsoft Word or any other software that properly typesets mathematical equations.
- I encourage collaborations, however, everything that is handed in (both your report and your code) should be your work.
- Your homework will be graded based on how completely you solved it as well as neatness and little things like: did you label your graphs and include figure captions. **The homework is worth 10 points. 5 points will be given for the overall layout, correctness and neatness of the report, and 5 additional points will be for specific technical things and computational results that the TAs will look for in the report itself.**

Problem Description

Your goal in this HW is to fit various types of models to a data set of fuel consumption (mpg) of various cars as a function of different attributes such as engine displacement, horsepower and weight. You will investigate the quality of your models and compare whether higher complexity results in better model performance and relate your observations to numerical stability and conditioning of the problem at hand.

- Python users: Download the file `mpg_train_test.npz` from Canvas and load it using `numpy.load`.
- MATLAB users: Download the file `mpg_train_test.mat` from Canvas and load it using the `load` command.
- The file `auto-mpg.xls` is also included which contains the raw data set. You do not need this file for the HW and it is only provided for those of you who are curious about it.

The provided files contain a training data set $X_{\text{train}}, Y_{\text{train}}$ where X_{train} is a matrix of size 254×3 and Y_{train} is a vector of size 254. The first column of X_{train} contains engine displacements, second column contains horsepower, while the third column contains the curb weight of the vehicles. The vector Y_{train} contains the mpgs (fuel consuptions). The training data contains 254 instances. The test data set $X_{\text{test}}, Y_{\text{test}}$ are analogous to the training data set but contain 138 instances.

1. Each data point in the training set has three features (displacement, horsepower, weight) which constitute the rows of the matrix X_{train}. We let $\underline{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ denote the set of features of a car and let $y(\underline{x}) \in \mathbb{R}$ denote the mpg of that car. Your first task is to find a function $y(\underline{x})$ that is only a function of one of the features. More precisely consider functions of the form

   (a) (mpg as a function of displacement) $y^{(1,2)}(x_1) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
   (b) (mpg as a function of horsepower) $y^{(2,2)}(x_2) = \theta_0 + \theta_1 x_2 + \theta_2 x_2^2$
   (c) (mpg as a function of weight) $y^{(3,2)}(x_3) = \theta_0 + \theta_1 x_3 + \theta_2 x_3^2$

   In other words, in each model you want to predict the mpg y as a function of one of the features only. For each of the above models formulate an appropriate least squares problem using the training data set to find the parameters θ_j that define the models (note, you will find a different $\underline{\theta}$ vector for each model). In a table report the relative test error of each of your models

   $$\text{relative error}(i) = \frac{1}{\|Y_{\text{test}}\|_2^2} \sum_{j=1}^{138} \left| y^{(i)}((X_{\text{test}})_{ji}) - (Y_{\text{test}})_j \right|^2,$$

   Also for each model present a 2D scatter plot of x_i-vs-Y of the test data overlaid with a line plot of your model $y^{(i)}(x_i)$. Discuss your findings and report which feature is the better predictor of the mpg. *Hint: you can use Vandermonde matrices to easily fomulate and solve your least squares problems.*

2. Now we consider higher order analogues of the models from part 1. More precisely, consider K-th degree models of the form

   $$y^{(i,K)}(x_i) = \theta_0 + \sum_{k=1}^{K} \theta_k x_i^k,$$

   for $i = 1, 2, 3$. Repeat the experiments of part 1 by replacing the quadratic models in (a)–(c) with their K-th degree counter parts for $K = 4, 8, 12$. Present a table containing the relative errors of each model and choice of K as well as the requisite scatter and line plots showing the behavior of each model. Is there a point of diminishing returns in terms of the degree K? Discuss your results. *Hint 1: In your scatter plots you can plot each model $y^{(i)}$ with three different values of K on the same plot to save space. Don't forget to generate a legend. Hint 2: Take a look at the condition number of the matrices in your least squares problem.*

3. Next we consider models that depend on multiple features at a time. Let

   (a) $y^{(12)}(\underline{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$
   (b) $y^{(13)}(\underline{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_3 + \theta_3 x_1 x_3 + \theta_4 x_1^2 + \theta_5 x_3^2$
   (c) $y^{(23)}(\underline{x}) = \theta_0 + \theta_1 x_2 + \theta_2 x_3 + \theta_3 x_2 x_3 + \theta_4 x_2^2 + \theta_5 x_3^2$
   (d) $y^{(123)}(\underline{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_3^2$

   Formulate the appropriate least squares problem to find the $\underline{\theta}$ vector for these models using the training data. Compute the relative errors on the test set and compare it to your previous models. Does adding features improve the performance of the model?

4. Discuss your findings and pick the model that performs best among the many models that your trained.