

# ISTA 421/521 – Homework 2

**Due: Tuesday, September 24, 10pm**

16 points total

Ian Montgomery

Undergraduate

## Instructions

In this assignment you are required to write 3 scripts in python. They will be submitted as 3 separate files, although you are free to copy chunks of code from one script to the next as desired. The names for the script files are specified in problems 1, 5 and 8, below.

Included in the homework 2 release are two sample scripts (`fitpoly_incomplete.py` and `cv_demo_incomplete.py`) and two data files (`womens100.csv`, and `synthdata.csv`). The sample scripts are provided for your convenience – you may use any part of the code in those scripts for your submissions. Note that neither will run just as provided: you must fill in the calculation of  $\mathbf{w}$ . The data files are provided in `.csv` (comma separated values) format. The script `fitpoly_incomplete.py` shows how to load the files.

All problems except problem 1 require that you provide some “written” answer (in some cases also figures), so you will also submit a `.pdf` of your written answers. (You can use  $\text{\LaTeX}$  or any other system (including handwritten; plots, of course, must be program-generated) as long as the final version is in PDF.)

**The final submission will include (minimally) the three scripts and a PDF version of your written part of the assignment. You are required to create either a `.zip` or tarball (`.tar.gz` / `.tgz`) archive of all of the files for your submission and submit the archive to the d2l dropbox by the date/time deadline above.**

NOTE: Problems 6 and 7 are required for Graduate students only; Undergraduates may complete them for extra credit equal to the point value.

(FCMA refers to the course text: Rogers and Girolami (2012), *A First Course in Machine Learning*. For general notes on using  $\text{\LaTeX}$  to typeset math, see: <http://en.wikibooks.org/wiki/LaTeX/Mathematics>)

1. [2 points] Adapted from **Exercise 1.2** of FCMA p.35:

Write a Python script that can find the parameters  $\mathbf{w}$  for an arbitrary dataset of  $x_n, t_n$  pairs. You will use this script to answer problem 2, which only requires fitting a simple line to the data (i.e., you only need to fit parameters  $w_0$  and  $w_1$ ); however, in problems 5 and 8 you will need to fit higher-order polynomial models (e.g.,  $t = w_0 + w_1x + w_2x^2 + \dots$ ), so make sure your script is generalized to handle higher-order polynomials. You may use the provided script `fitpoly_incomplete.py` to achieve this, or write your own from scratch. Whichever you choose, your script must have the ability to plot the data with the fitted model.

You will submit your script as a stand-alone file called `fitpoly.py`.

2. [1 point] Adapted from **Exercise 1.6** of FCMA p.35:

Table 1.3 (p.13) of FCMA lists the women's 100m gold medal Olympic results – this data is provided in the file `womens100.csv` in the data folder. Using your script from problem 1, find the 1st-order polynomial model (i.e., a line with parameters  $w_0$  and  $w_1$ ) that minimizes the squared loss of this data. Report the model here as a linear equation and also include a figure that includes a plot of your model with the data.

**Solution.**  $y = 40.9241546 - 0.0150718122x$

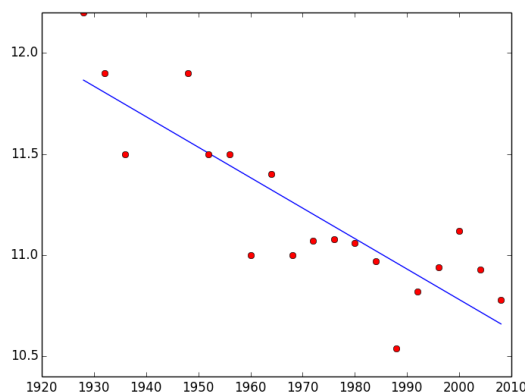


Figure 1: Women's 100m times and best-fit linear model

3. [1 point] Adapted from **Exercise 1.7** of FCMA p.35:

Using the model obtained in the previous exercise, predict the women's winning time at the 2012 and 2016 Olympic games; report the values here. What is the squared error of your model's prediction for Shelly-Ann Fraser-Price's gold medal time of 10.75 seconds in last year's 2012 Olympics women's 100m race?

**Solution.**

Predictions for 2012 and 2016:

2012: 10.599s MSE = .0226

2016: 10.539s

4. [1 point] Adapted from **Exercise 1.9** of FCMA p.36:

Use your python script from problem 1 to load the data stored in the file `synthdata.csv` (in the data folder). Fit a 3rd order polynomial function –  $f(x; \mathbf{w}) = w_0 + w_1x + w_2x^2 + w_3x^3$  – to this data. Present

and describe the parameters you obtain fitting to this data. Also, plot the data and your linear-fit model and include the plot in your answer.

**Solution.** parameters:

$$\begin{vmatrix} 2.92292634 \\ 7.07188097 \\ -0.18778007 \\ -1.45278081 \end{vmatrix}$$

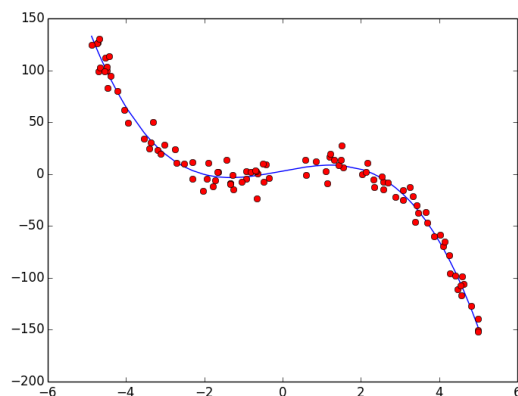


Figure 2: 3rd order polynomial fit to synthetic data

5. [3 points] Write a script that implements 1K-fold cross-validation to choose the polynomial order (between 1 and 5) with the best predictive error for the `synthdata.csv`. The provided script `cv_demo_incomplete.py` implements the synthetic data experiment described in the book; you are welcome to use and adapt any part of this code you like; keep in mind that this script won't successfully execute until you add the general (matrix form) normal equation calculation on line 80. Note also that in the synthetic data experiment in `cv_demo_incomplete.py`, 1000 test data points are generated in addition to the 100 data points used for 10-fold cross-validation; for problem 5, you won't have this independent test set, only the data from `synthdata.csv` on which to perform K-fold cross-validation. Run your script with 10-fold cross-validation and Leave-One-Out-CV (LOOCV). Which model order do the two cross-validation methods predict as the best order for predictive accuracy? Do the two different cross-validation runs agree? Report the best-fit model parameters for the best model order according to LOOCV, and plot this model with the data. Include a plot of the CV-loss and training loss for the 5 different polynomial model orders for both the 10-fold cross-validation and for LOOCV.

You will submit your script as a stand-alone file called `cv.py`

**Solution.**

parameters:

$$\begin{vmatrix} 2,905971271 \\ 7.06475915 \\ -0.18753937 \\ -1.45223967 \end{vmatrix}$$

Associated MSE:

First order: 915.86

Second order: 957.34

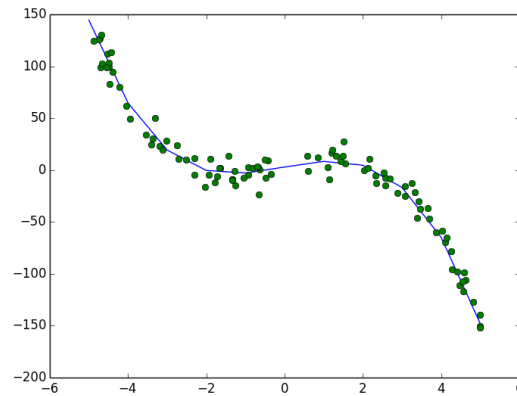


Figure 3: Perhaps it is a difference in algorithm, but I found that the 3rd order LOOCV has the best fit

Third order: 74.45

Fourth order: 79.03

Fifth order: 101.12

parameters:

$$\begin{pmatrix} 3.026727 \cdot 7 \\ 7.17823219 \\ -.229764832 \\ -1.45756401 \\ 0.00220409208 \end{pmatrix}$$

Associated MSE:

First order: 1271.68

Second order: 847.11

Third order: 84.61

Fourth order: 84.31

Fifth order: 87.69

8. [4 points] Variant of **Exercise 1.12** from FCMA p.36

Write a new Python script that uses  $K$ -fold cross-validation to find the value of  $\lambda$  that gives the (approximate) best predictive performance on the synthetic data (`synthdata.csv`) using **regularized** least squares with a **5th order polynomial** model. Report here the lambda you identified, the best fit linear model using that lambda (as a linear equation) and include a plot of the MSE loss as a function of  $\lambda$  that shows the loss curve with a minimum, and a plot of the best-fit model with the data.

You will submit the script as a stand-alone file called `regularize.py`.

**Solution.** <Solution goes here>

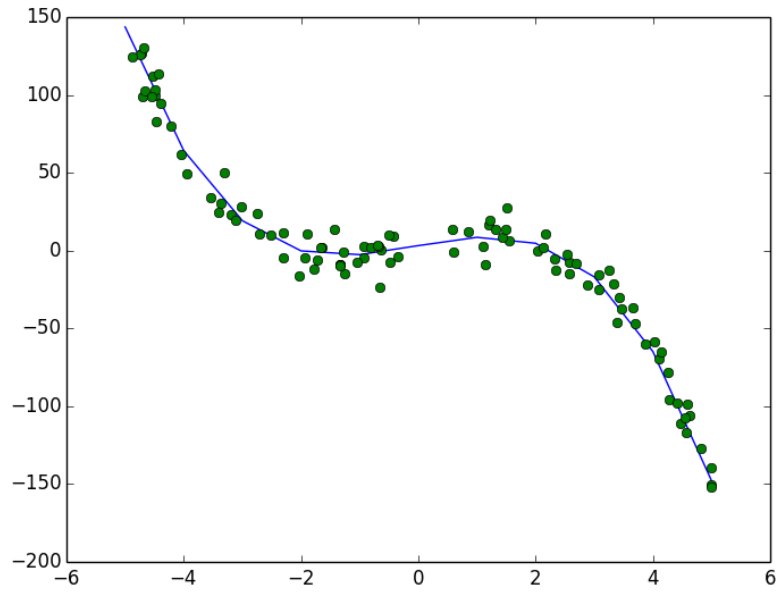


Figure 4: 10-fold CV: CV and training loss across polynomial model orders.

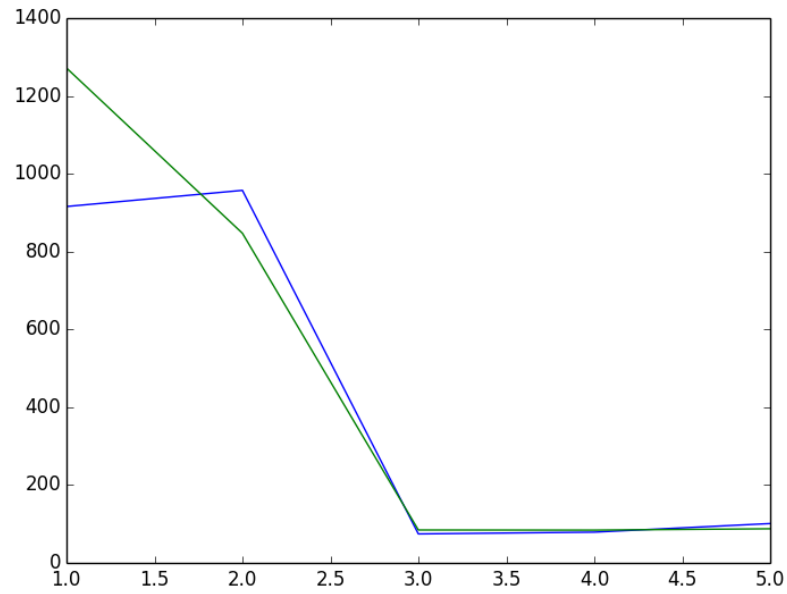


Figure 5: LOOCV: CV and training loss across polynomial model orders, blue represents LOOCV, and green is K-10 CV.

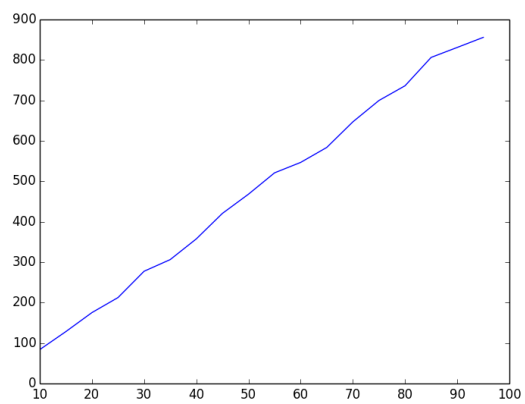


Figure 6: MSE Loss as a function of  $\lambda$ .