

# ELEC 378 – Spring 2023

## Homework 7

**Due:** Friday March 3, 5PM

### 1 Reading Assignment

Read this rant on multiple regression: <https://goo.gl/Cq7YGP>

Please write a paragraph summarizing the article, tying it into the class content, and then summarizing your learnings.

### 2 Linear Regression, Cubic Style (Fall 2020 Exam Problem)

Given a set of training data  $\{x_i \in \mathbb{R}, y_i \in \mathbb{R}\}_{i=1}^n$ , you wish to learn a predictor for the label  $y_i$  given an input  $x_i$ .

- a) After plotting the data, you have a hunch that the labels are approximately a cubic function of the inputs. Write down the general form of a cubic function.
- b) Plot an example of a set of  $> 10$  data points that visually conform to this kind of model.
- c) Explain carefully how you would set up and solve this prediction problem, including assembling an appropriate data matrix  $\mathbf{X}$  (write out the entries of  $\mathbf{X}$  in detail) and writing the equations you would solve for the optimal least-squares predictor.
- d) Explain some of the things that could go awry in your solution in the previous part and how you would fix them.
- e) Explain how you would deal with a cubic dependence of the label on the input data when the data are now vectors  $\mathbf{x}_i \in \mathbb{R}^p, p > 1$ . What issues might arise in this case?

### 3 Predicting House Sale Price with Linear Regression II

Download `train.csv` from the following Kaggle competition:

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

- a) Demonstrate that the provided features are ill-conditioned, i.e., that the data matrix has singular values close to zero.
- b) Use ridge regression to predict home sale prices from the provided features for various values of  $\lambda$ . How does this method compare to the unregularized linear regression attempted last week?
- c) Use Lasso regression to predict home sale prices from the provided features for various values of  $\lambda$ . How does this method compare to the unregularized linear regression attempted last week?
- d) Compare the performance of ridge and Lasso on this dataset. Which method is more accurate? Which method makes it easier to determine features that are important in assessing a home's value?

### 4 $K$ -Nearest Neighbors Image Classification

Download and familiarize yourself with the MNIST handwritten digits dataset:

[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_digits.html](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html)

- a) Randomly select 20% of the MNIST data points and their labels to be removed from the training data and set aside as test data.
- b) Using the 2-norm distance, classify the test data using a  $K$ -Nearest Neighbor classifier (implemented from scratch) for various values of  $K$ , reporting the misclassification rate in each case.
- c) Repeat the experiment in b) but using the 1-norm distance. Compare your results to the those in b) using the 2-norm distance.
- d) Repeat the experiment in b) but using the  $\infty$ -norm distance. Compare your results to the those in b) and c).

## **Submission Instructions**

Every student must submit their work in PDF format, providing intermediate and final results as well as any necessary code. Submit your homework on Gradescope.

## **Collaboration Policy**

Collaboration both inside and outside class is encouraged. You may talk to other students for general ideas and concepts, but individual write-ups must be done independently.

## **Plagiarism**

Plagiarism of any form will not be tolerated. You are expected to credit all sources explicitly.