ECE5268 – Theory of Neural Networks (Spring 2020)
Major Project #1

Dr. Georgios C. Anagnostopoulos

Saturday 8th February, 2020

1 Objectives

The objective of Major Project (MP) I is to assess students' understanding in a variety of aspects pertaining to linear regression.

Before putting together your work, carefully review the preparation guidelines (Section 3) and submission instructions (Section 4) that are provided.

2 Assignments

Task 1. [30 total points]
Task 1 of this MP consists of a series of essay-type questions. For each part/question provide a succinct, but complete, response in the form of a paragraph.

(a) **[10 points]** Explain what *model over-fitting* stands for. How does over-fitting manifest itself? You may assume that that we are talking about regression models.

(b) **[10 points]** What is meant by *validation procedure*? What is it useful for? Provide a brief example of circumstances, where a validation procedure would be useful, if employed (explain in what sense).

(c) **[10 points]** What is is meant by *model regularization*, what is its purpose and what kinds of regularization approaches are you familiar with?

Task 2. [30 total points]
This task pertains to linear regression models; the following parts are independent of each other.

(a) **[10 points]** Show that, for a linear regression problem with design matrix $\mathbf{X}$, the training Mean Squared Error (MSE) of a linear model with parameters $\mathbf{w}$ is given as

$$\mathrm{MSE}_{\mathrm{train}}(\mathbf{w}) = \frac{1}{N}\left\|\mathbf{w}-\mathbf{w}^*\right\|_{\mathbf{R}}^2 + \frac{1}{N}\mathbf{y}^T\left(\mathbf{I}_N - \mathbf{X}\mathbf{X}^\dagger\right)\mathbf{y} \tag{1}$$

where $\mathbf{w}^*$ are the optimal parameter values, $\mathbf{R} \triangleq \mathbf{X}^T\mathbf{X}$ is invertible and $\mathbf{X}^\dagger = \mathbf{R}^{-1}\mathbf{X}$.

(b) **[10 points]** Show that for a linear regression problem featuring an intercept parameter, the optimal hyper-plane passes through the point $(\bar{\mathbf{x}}, \bar{y})$, where

$$\bar{\mathbf{x}} \triangleq \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{2}$$

$$\bar{y} \triangleq \frac{1}{N}\mathbf{1}_N^T\mathbf{y} = \frac{1}{N}\sum_{n=1}^{N}y_n \tag{3}$$

(c) **[10 points]** As we know, *Ridge Regression* considers the ℓ_2^2-regularized MSE given as

$$\mathrm{RMSE}_{\mathrm{train}}(\widetilde{\mathbf{w}}|\mu) \triangleq \frac{1}{N}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\mathbf{w}}\right\|_2^2 + \frac{\mu}{N}\|\mathbf{w}\|_2^2 \tag{4}$$

where $\mu \geq 0$ is the penalty parameter; here, it is divided by N for algebraic convenience. Also, notice that the model uses an intercept term. Given μ, find the minimizer $\widetilde{\mathbf{w}}^*(\mu)$ and the optimal value $\mathrm{RMSE}_{\mathrm{train}}(\widetilde{\mathbf{w}}^*(\mu)|\mu)$.

Task 3. [40 total points]

This task calls for the prediction of house sale prices using linear regression based on a limited collection of their characteristics.

For this task, you are not allowed to use any regression/validation libraries; develop your own code for addressing this task.

The relevant dataset, whose data are stored in `housing_dataset.csv`, is a modified, reduced and cleaned-up version of the House Prices dataset hosted at Kaggle, an online platform for predictive modeling and analytics competitions. In turn, this dataset was constructed from the *Ames Housing dataset* introduced in [1], which pertains to sale prices of residential properties in Ames, Iowa from 2006 to 2010. A description of the Kaggle-hosted dataset can be accessed at this location.

The dataset provided here contains a total of 1195 samples and contains the following 11 columns (variables):

Column Name	Description
SalePrice	Property's sale price
LotFrontage	Length of street connected to property
LotArea	Lot area
OverallQual	Rating of the overall material and finish of the house
MasVnrArea	Masonry veneer area
YearBuilt	Normalized construction year
BsmtUnfSF	Unfinished basement area
YearRemodAdd	Normalized remodeling year
TotalBsmtSF	Total basement area
BsmtFinSF1	From the official description: "Type 1 finished square feet"
1stFlrSF	First floor area

All columns are numeric and their values have been appropriately normalized to lie in $[0, 1]$. For our linear regression model, the first column variable, namely *SalePrice*, will serve as the model's output, while the remaining 10 variables will serve as input features.

(a) **[10 points]** Create scatter plots depicting the output variable versus each input feature and comment on the appropriateness of considering these features in a linear regression model.

(b) **[15 points]** Let's suppose that we are entertaining three modeling options: a model that considers

(i) all 10 features as inputs

(ii) all features except *OverallQual* as inputs

(iii) *YearBuilt* squared in addition to all 10 original features

Perform 5-fold cross-validation using a training set to determine, which of the aforementioned options seems to be our best bet (the champion model) for modeling the given data via linear regression. For the purpose of cross-validation, use the first 1000 samples of the dataset as your training set. Then, report on your champion model, *i.e.*, which one it is and how much better it is compared to the other two options.

(c) **[15 points]** First, split your available data into two sets:

- training set, which consists of the first 1000 samples in the dataset
- test set, which consists of the remaining (last) 195 samples in the dataset

Next, fit a linear regression model with the features used by the champion model that you identified in part (b), plot the fitted residual plot (fitted residuals vs. fitted predictions)[1] for this model and compute the fitted model's test MSE (MSE as evaluated on the test data). Finally, based on the residual plot and the test MSE, comment on whether the fitted linear regression model seems to be appropriate for this dataset and how well it generalizes.

References

[1] Dean De Cock. Ames, Iowa: Alternative to the Boston Housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3):1–14, 2011. URL: https://ww2.amstat.org/publications/jse/v19n3/decock.pdf.

[1] In other words, predicted errors vs. predicted outputs.

3 Preparation Guidelines

Below are some general guidelines that you are asked to follow, when compiling a Major Project report. These guidelines greatly facilitate your work's assessment by a grader and, at the same time, aim at helping you sidestep some major pitfalls that would prevent you from receiving the maximum credit for your work.

- **Task Statements:** Before attempting to address a particular task, ensure that you completely understand what is asked from you to perform and/or to produce. When in doubt, ask your instructional staff for clarifications! Also, make sure you did not omit your response to any of the parts that you have attempted. Finally, make sure that it is crystal clear, which response corresponds to which task/part.

- **Derivations & Proofs:** If you provide handwritten derivations and/or proofs, make sure you use your best handwriting. Each derivation should have a logical and organized flow, so that it is easy to follow and verify.

- **Code & Data:** The code that you author should be as well-organized as possible and amply commented. This is very useful for assessing your work, as well as for you, while you are debugging/or modifying it, or when you have to go back to it in the near future. **Caution:** You are not allowed to use any code that you have not produced without having/obtaining explicit prior permission, in which case the source(s) you have obtained this code from must be clearly indicated via comments inside your code as well in your report. You are deemed to be plagiarizing, if you fail to do so, which may have dire consequences. Finally, if a task asks you to generate data, keep them organized in a separate folder and document, *e.g.*, in a text file, the specifics of how they were generated.

- **Figures, Plots & Tables:** Plots should have their axes labeled and, if featuring several visual elements such as curves or types of points on the same graph, an appropriate legend should be used. Whether figures or tables, each one of these elements should feature a caption with sufficient information on what is being displayed and how were these results obtained (*e.g.*, under what experimental conditions or settings, etc.). You should ask yourself the question: if someone comes across it, will they understand what is being depicted? Apart from a concise description, major, relevant conclusions stemming from the display should also be included in the caption text.

- **Observations, Comments & Conclusions:** When stating observations about a particular result, do not stop at the obvious that anyone can notice (*e.g.*, "*... we see that the curve is increasing.*"). Instead, assess whether the result is expected, either by theory or intuition (*e.g.*, "*... This is as expected, because X is the integral of ...*"), or, if it is unexpected, offer a convincing reasoning behind it (*e.g.*, "*... We expected a decreasing curve ... All points to that I must have not been calculating X correctly ...*"). The latter is more preferable (*i.e.*, expect partial credit) than stopping at the obvious, which happens to be wrong (*i.e.*, do not expect partial credit). Next, descriptions and comments on results should be sufficient. Be concise, but complete. Finally, conclusions that you draw must be well-justified; vacuous conclusions will be swiftly discounted.

4 Submission Instructions

Kindly adhere to the conventions and submission instructions outlined below. Deviations from what is described here may cause unnecessary delays, costly oversights and immense frustrations related to the assessment of your hard work.

First, store all your MP deliverables in a folder named `lastname_mpX`, where `lastname` should be your last name and `X` should be the number of the MP, like 1, 2, etc. The folder name should be all lower case. For example, Anagnostopoulos' folder for MP 1 would be named: `anagnostopoulos_mp1`.

Secondly, your `lastname_mpX` folder should have the following contents:

- A signed & dated copy of the Work Origination Certification page in Adobe PDF format. You can either scan such a page after you complete, date and sign it, or do so electronically, as long as your signature is not typed. If this page is missing from your report, your MP work will not be considered for assessment (grading) and will be assigned a default total score of 0/100.
- Your report in Adobe PDF format using the file naming convention `lastname_mpX.pdf`. The report needs to address all tasks in sequence as given, even tasks that are not attempted. For tasks and parts that require you to show analytical work (*e.g.*, a derivation/proof), you are not obliged to typeset it. While it would be nice to do so, such effort may turn out to be quite time-consuming. Instead, you can scan your work into images, as long as it is legible and well organized with a clear logical flow. Also, please make sure that you indicate which task/part your work corresponds to. When scanning your hand-written work, use a relatively low-resolution (DPI) setting, so your resulting PDF document does not become too big in size, which may prevent you from uploading your work to Canvas.
- A subfolder named `src`, which contains all the code you wrote for this MP.

When you are done with producing your responses to this MP, compress your folder called `lastname_mpX` into a single ZIP archive named `lastname_mpX.zip` and upload it to Canvas by the specified deadline.

WORK ORIGINATION CERTIFICATION

By submitting this document, I, _____, the author of this deliverable, certify that

1. I have reviewed and understood the section regarding Academic Honesty of the current version of FIT's Student Handbook's Standards And Policies, which are available at https://policy.fit.edu/policy/9267 and which discusses academic dishonesty (plagiarism, cheating, miscellaneous misconduct, etc.)
2. The content of this Major Project report reflects my personal work and, in cases it is not, the source(s) of the relevant material has/have been appropriately acknowledged after it has been first approved by the course's instructional staff.
3. In preparing and compiling all this report material, I have not collaborated with anyone and I have not received any type of help from anyone but the course's instructional staff.

Signature _____ Date _____