Assignment 1: Regression

Due October 9 at 11:59pm
109 marks total

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment

---

1 Probabilistic Modeling (12 marks)

In lecture we went over an example of modeling coin tossing – estimating a parameter that is the probability the coin comes up heads.

Consider instead the problem of modeling the outcome of the Canadian Federal election. To simplify matters, assume one party will win a majority (i.e. either the NDP, Liberals, Conservatives, or Green Party wins).

Check through question 1

1. (4 marks) What is the type of distribution that describes this situation? What are the parameters μ of this distribution? (See PRML Appendix B)

2. (2 marks) What would be the value of the parameters μ for an election where the outcome is an equal chance of each party winning?

3. (2 marks) What would be the value of the parameters μ for an election that is completely "rigged"? E.g. the party currently in power is definitely going to win.

4. (4 marks) Suppose my prior is that the Green Party has completely rigged the election. Assume I see a set of polls where the NDP has the largest share of the vote in each poll. What would be my posterior probability on the parameters μ?

2 Regularized Least-Squares Linear Regression (15 marks)

Show that the minimizer for least-squares linear regression with L_2 regularization is $\boldsymbol{w} = \left(\lambda \boldsymbol{I} + \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{t}$.

3 Training vs. Test Error (12 marks)

Double check this

1. (4 marks) Suppose we perform unregularized regression on a dataset. Is the **training error** with a degree 10 polynomial always lower than or equal to that using a degree 9 polynomial? Explain.

2. (4 marks) Suppose we perform unregularized regression on a dataset. Is the **testing error** with a degree 10 polynomial always lower than or equal to that using a degree 9 polynomial? Explain.

3. (4 marks) Suppose we perform unregularized regression on a dataset. Is the **training error** always lower than the **testing error**? Explain.

4 Regression (70 marks)

In this question you will train models for regression and analyze a dataset. Start by downloading the code and dataset from the website.

The dataset is created from data provided by UNICEF's State of the World's Children 2013 report: http://www.unicef.org/sowc2013/statistics.html

Child mortality rates (number of children who die before age 5, per 1000 live births) for 195 countries, and a set of other indicators are included.

4.1 Getting started

Run the provided function `loadUnicefData.m` to load the dataset and names of countries / features.

Answer the following questions about the data. Include these answers in your report.

1. (2 marks) Which country had the highest child mortality rate in 1990? What was the rate?
2. (2 marks) Which country had the highest child mortality rate in 2011? What was the rate?
3. (2 marks) Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function `loadUnicefData.m`?

For the rest of this question use the following data and splits for train/test and cross-validation.

- **Target value**: column 2 (Under-5 mortality rate (U5MR) 2011).
- **Input features**: columns 8-40.
- **Training data**: countries 1-100 (Afghanistan to Luxembourg).
- **Testing data**: countries 101-195 (Madagascar to Zimbabwe).
- **Cross-validation**: subdivide training data into folds with countries 1-10 (Afghanistan to Austria), 11-20 (Azerbaijan to Bhutan), ... I.e. train on countries 11-100, validate on 1-10; train on 1-10 and 21-100, validate on 11-20, ...

4.2 Polynomial Regression

Implement linear basis function regression with polynomial basis functions. Use only monomials of a single variable (x_1, x_1^2, x_2^2) and no cross-terms ($x_1 \cdot x_2$). You may find the provided function `designMatrix.m` useful.

Perform the following experiments:

1. (20 marks) Create a MATLAB script `polynomial_regression.m` for the following.

   Fit a polynomial basis function regression (unregularized) for degree 1 to degree 6 polynomials. Plot training error and test error (in RMS error) versus polynomial degree.

??

Put this plot in your report, along with a brief comment about what is "wrong" in your report.

Normalize the input features before using them (not the targets, just the inputs $\boldsymbol{x}$). Use `normalizeData.m`.

Run the code again, and put this new plot in your report.

2. (20 marks) Create a MATLAB script `polynomial_regression_1d.m` for the following.

   Perform regression using just a single input feature.

   Try features 8-15 (Total population - Low birthweight). For each (un-normalized) feature fit a degree 3 polynomial (unregularized).

   Plot training error and test error (in RMS error) for each of the 8 features. This should be a bar chart (use `bar([train_err test_err])`).

   Put this bar chart in your report.

   The training error for feature 11 (GNI per capita) is very high. To see what happened, produce plots of the training data points, learned polynomial, and test data points. The code `visualize_1d.m` may be useful.

   In your report, include plots of the fits for degree 3 polynomials for features 11 (GNI), 12 (Life expectancy), 13 (literacy).

4.3 Sigmoid Basis Functions

1. (10 marks) Create a MATLAB script `sigmoid_regression.m` for the following.

   Implement regression using sigmoid basis functions for a single input feature. Use two sigmoid basis functions, with $\mu = 100, 10000$ and $s = 2000$. Include a bias term.

   Fit this regression model using feature 11 (GNI per capita).

   In your report, include a plot of the fit for feature 11 (GNI).

   In your report, include the training and testing error for this regression model.

4.4 Regularized Polynomial Regression

1. (20 marks) Create a MATLAB script `polynomial_regression_reg.m` for the following.

   Implement L_2-regularized regression. Fit a degree 2 polynomial using $\lambda = \{0, .01, .1, 1, 10, 10^2, 10^3, 10^4\}$. Use 10-fold cross-validation to decide on the best value for λ. Produce a plot of average validation set error versus λ. Use a `semilogx` plot, putting λ on a log scale[1].

   Put this plot in your report, and note which λ value you would choose from the cross-validation.

[1]The unregularized result will not appear on this scale. You can either add it as a separate horizontal line as a baseline, or report this number separately.

Submitting Your Assignment

The assignment must be submitted online at `https://courses.cs.sfu.ca`. In order to simplify grading, you must adhere to the following structure.

You must submit two files:

1. You must create an assignment report in **PDF format**, called `report.pdf`. This report must contain the solutions to questions 1-3 as well as the figures / explanations requested for 4.

2. You must submit a .zip file of all your code, called `code.zip`. This must contain a single directory called `code` (no sub-directories, no leading path names), in which all of your files must appear[2]. There must be the 4 scripts with the specific names referred to in Question 4, as well as a common codebase you create and name.

   As a check, if one runs

   ```
   unzip code.zip
   cd code
   matlab
   polynomial_regression_1d
   ```

   the script produces the plots in your report from the relevant question.

[2]This includes the data files and others which are provided as part of the assignment.