Assignment 1: Solutions

1 Probabilistic Modeling and Bayes' Rule (10 marks)

(a) Given the information in the problem, we have $P(M) = 0.01, P(T|M) = 0.95$ and $P(T|\overline{M}) = 0.05$.

1.

$$\begin{aligned} P(T) &= P(T \wedge M) + P(T \wedge (\overline{M}) \\ &= P(T|M)P(M) + P(T|\overline{M})P(\overline{M}) \\ &= 0.95 \times 0.01 + 0.05 \times 0.99 \\ &= 0.059 \end{aligned} \tag{1}$$

(b)

$$P(M|T) = \frac{P(T|M) \times P(M)}{P(T)} = \frac{0.95 \times 0.01}{0.059} \approx 0.16 \tag{2}$$

2.

$$P(raintomorrow|raintoday) = \frac{P(tomorrow \wedge today)}{P(today)} = \frac{0.25}{0.30} = \frac{5}{6}$$

3.

P (odd) = P(1) + P(3) + P(5) = 0.1 + 0.2 + 0 = 0.3.

This is worse than a fair die which has probability 0.5 to land on an odd number.

2 Weighted Squared Error (15 marks)

The weighted error function we wish to minimize is:

$$E_{\hat{D}}(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \alpha_n \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3}$$

We take its derivatives and set them to zero:

$$\nabla E_{\hat{D}}(\boldsymbol{w}) = -\sum_{n=1}^{N} \alpha_n \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n)^T \tag{4}$$

$$\nabla E(\boldsymbol{w}) = 0 \tag{5}$$

$$\Leftrightarrow \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n)^T = \sum_{n=1}^{N} \alpha_n \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \boldsymbol{\phi}(\boldsymbol{x}_n)^T \tag{6}$$

In order to write this sum using matrix notation, define the matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} \alpha_1 & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & 0 & 0 \\ 0 & \vdots & 0 & \alpha_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \alpha_n \end{pmatrix} \tag{7}$$

Now,

$$\nabla E(\boldsymbol{w}) = 0 \Leftrightarrow (\boldsymbol{A}\boldsymbol{t})^T\boldsymbol{\Phi} = \boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{\Phi} \tag{8}$$

Taking transpose of both sides:

$$\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{t} = \boldsymbol{\Phi}^T\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{w} \tag{9}$$

Noting $\boldsymbol{A}$ is symmetric:

$$\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{t} = \boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{w} \tag{10}$$

We can take the square root of $\boldsymbol{A}$ since each $\alpha_i > 0$:

$$\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{t} = \boldsymbol{\Phi}^T\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{w} \tag{11}$$

Reassociating:

$$\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{t} = (\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Phi})^T(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Phi})\boldsymbol{w} \tag{12}$$

The resulting product is invertible, same reasoning as with design matrix:

$$\boldsymbol{w} = \left[(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Phi})^T(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Phi})\right]^{-1}\boldsymbol{\Phi}^T\boldsymbol{A}\boldsymbol{t} \tag{13}$$

3 Training vs. Test Error (12 marks)

For the questions below, assume that error means RMS (root mean squared error).

1. (4 marks) *Suppose we perform unregularized regression on a dataset. Is the* **validation error** *always higher than the* **training error***? Explain in 1-2 sentences.*

   No. There are no guarantees on the relationship between training error and validation error. While it is often the case that training error is lower than validation error, it is always possible that due to a particular validation set choice, the validation data points lie perfectly on the learned curve, while some of the training data points do not.

2. (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 in 1-2 sentences.*

   Yes. First, note that RMS is a monotonically increasing function of squared error, the criterion used in training unregularized regression.

   Degree 10 polynomials contain degree 9 polynomials. Unregularized regression leads to the optimal solution (caveats possible here). At worst, training error should be the same using a degree 10 polynomial as that using a degree 9 polynomial (10th term would have 0 coefficient for example).

3. (4 marks) *Suppose we perform both* **regularized** *and* **unregularized** *regression on a dataset. Is the* **testing error** *with a degree 20 polynomial always lower using* **regularized** *regression compared to* **unregularized** *regression? Explain in 1-2 sentences.*

   No. Again, there are no guarantees on the test data. Regularization is a technique designed to combat overfitting. Often it is useful (leads to lower testing error) when used to train complex models on small datasets. However, there is no guarantee that it will necessarily produce better test results. Beyond this, there are issues in tuning the hyper-parameter (λ) used to balance regularization with squared error on a particular dataset.

4 Regression (40 marks)

4.1 Getting started

1. (2 marks) *Which country had the lowest child mortality rate in 1990? What was the rate?*

   Iceland : 6.3

2. (2 marks) *Which country had the lowest child mortality rate in 2011? What was the rate?*

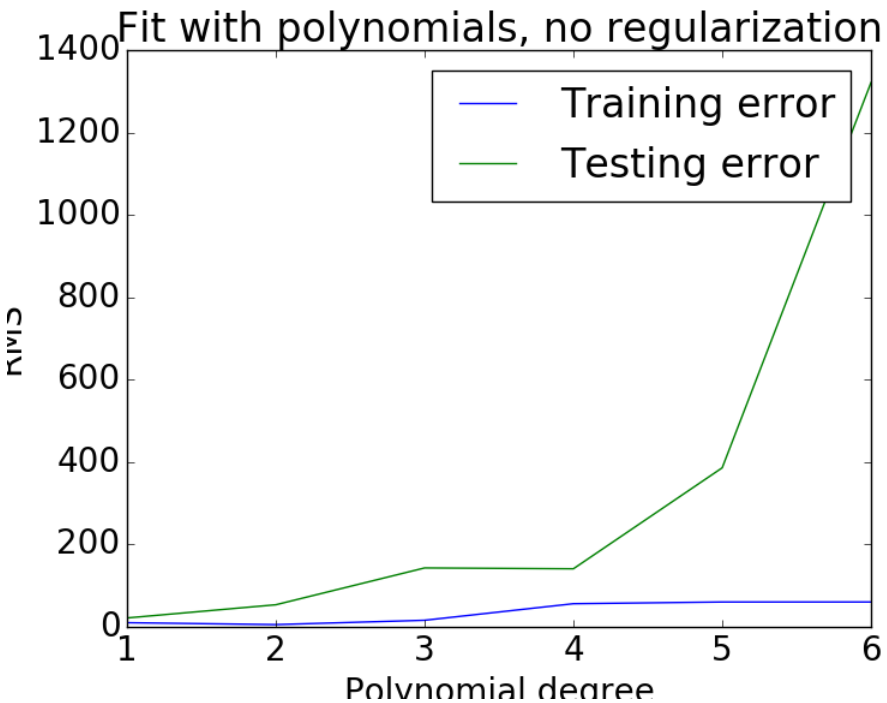
   San Marino : 1.8

3. (2 marks) *Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function* `assignment1.load_unicef_data()`*?*

   The mean value of the other countries' values is used for this feature.

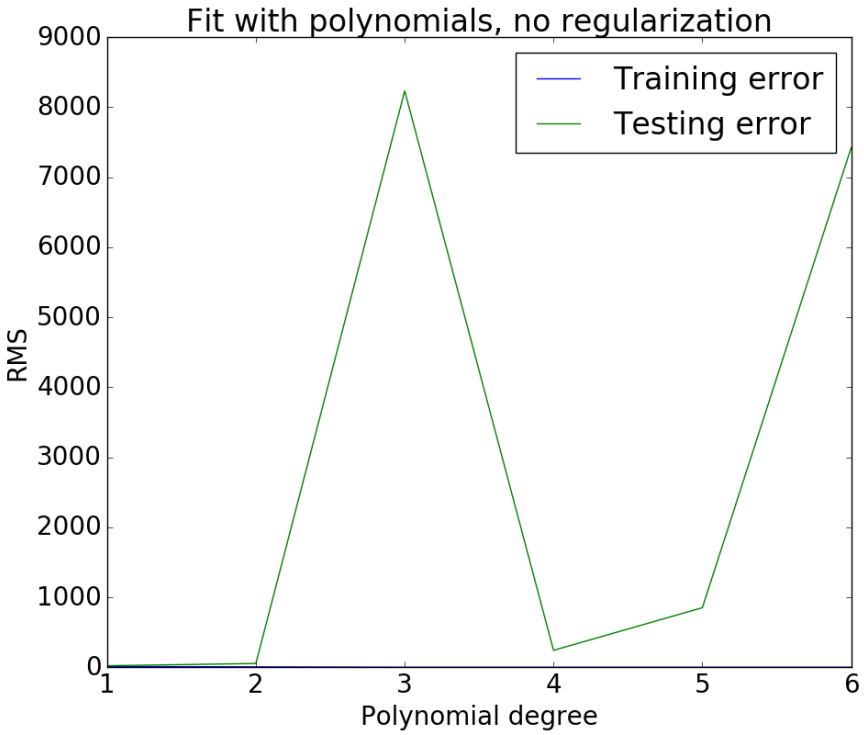
4.2 Polynomial Regression

1. Un-normalized data results in the following training and test errors.

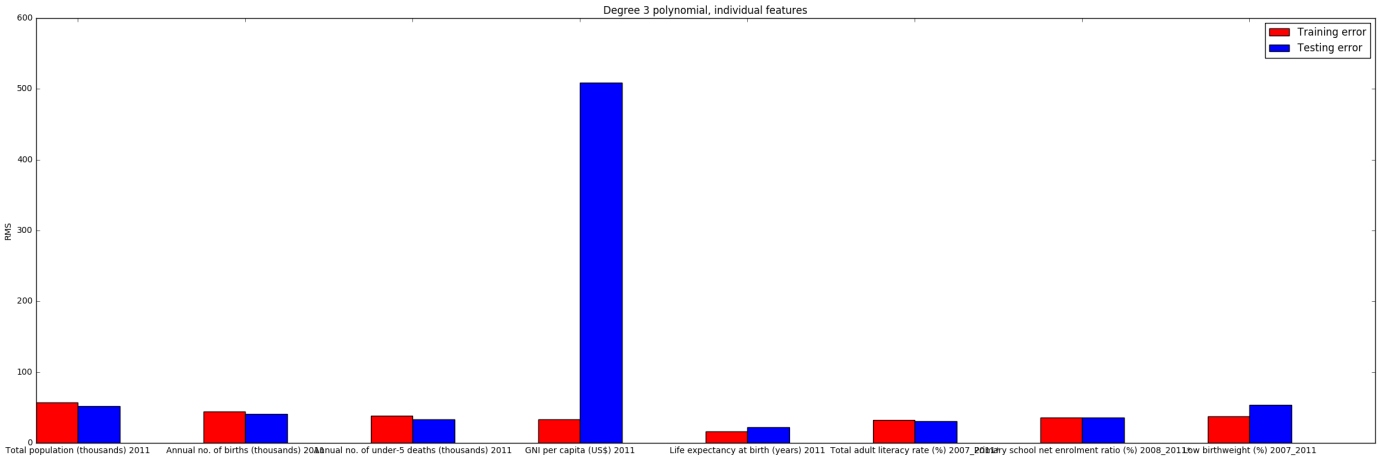


Note that training error actually increases with larger degree. This is due to numerical instabilities, due to large ranges in the values of inputs.

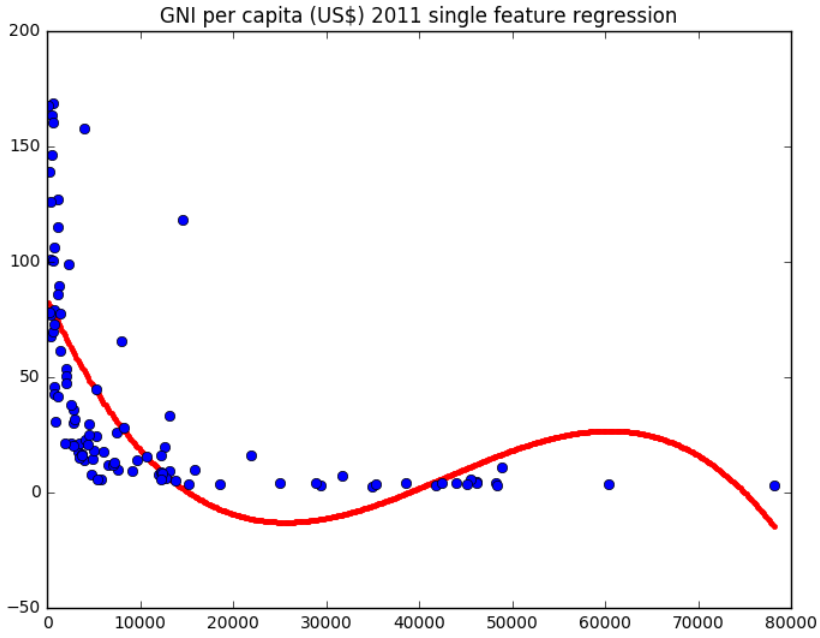
Results with normalized data are below.

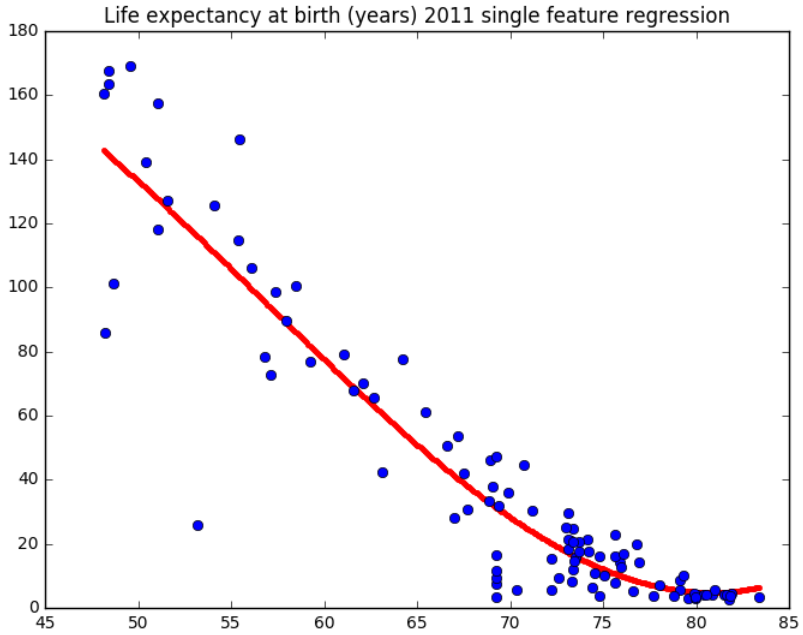


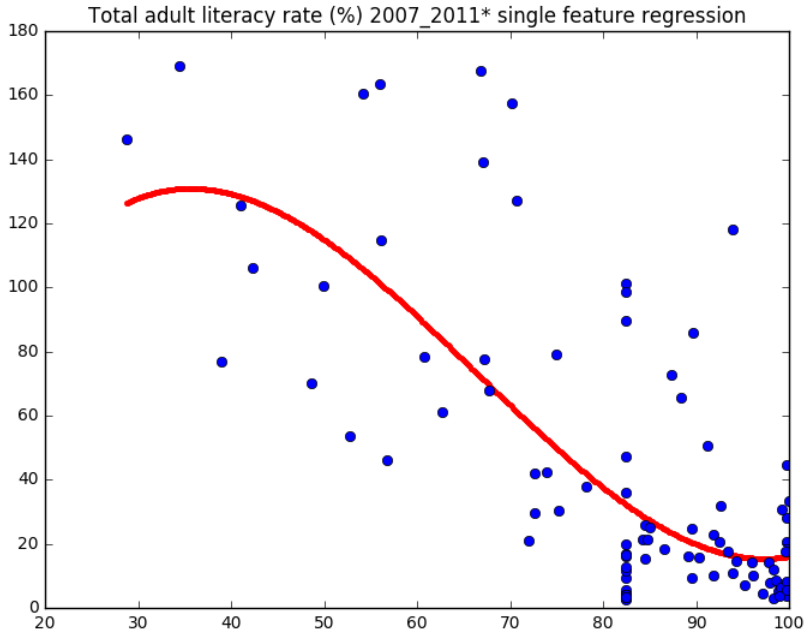
2. Single feature regression. Summary bar chart.



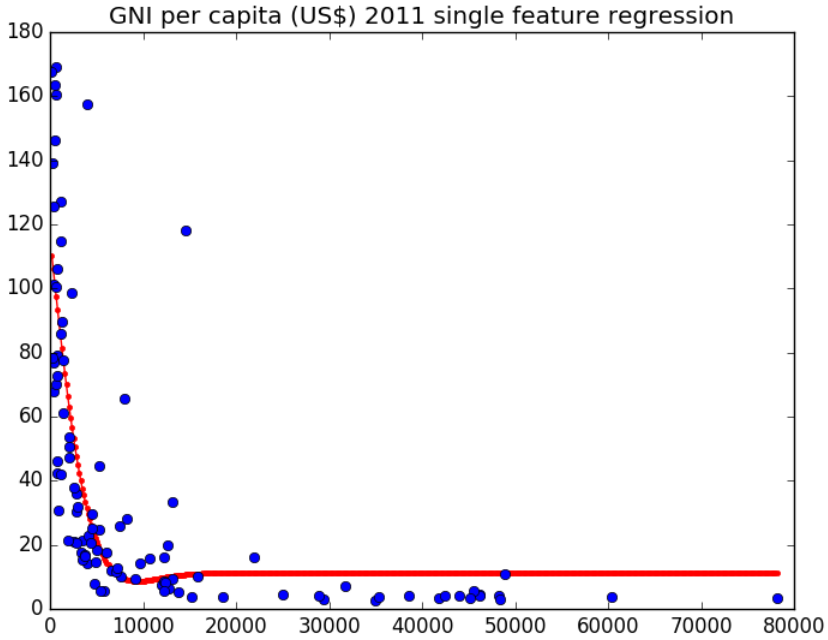
Fits for different features. Note the problems caused by outliers with large values of GNI.




Life expectancy at birth (years) 2011 single feature regression
180
160
140
120
100
80
60
40
20
0
45
50
55
60
65
70
75
80
85

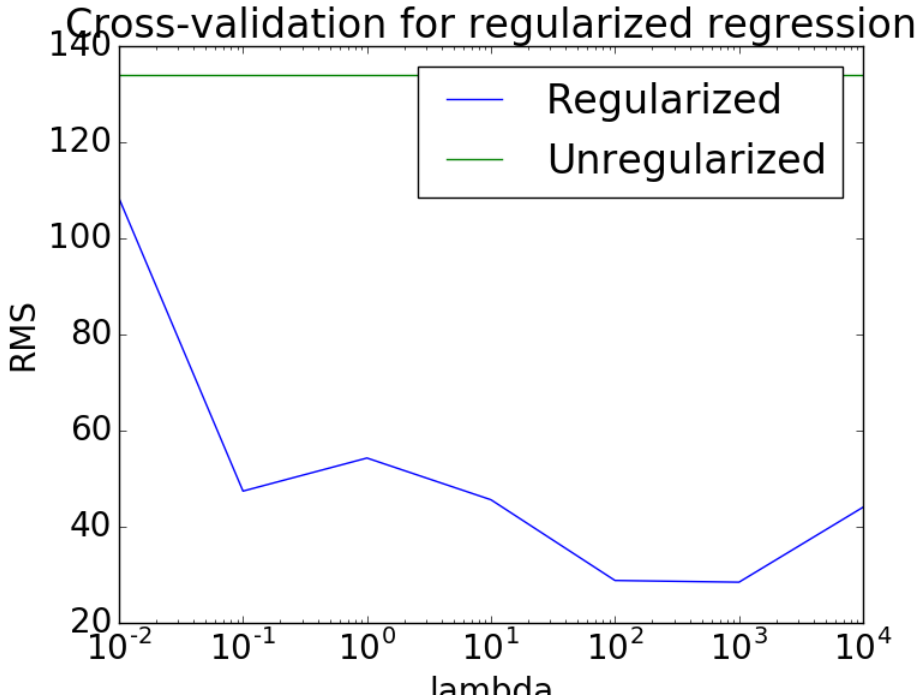

Total adult literacy rate (%) 2007_2011* single feature regression
180
160
140
120
100
80
60
40
20
0
20
30
40
50
60
70
80
90
100

4.3 Sigmoid Basis Functions



Training error is 28.46, testing error is 33.81.

4.4 Regularized Polynomial Regression



The value of cross-validation error for $\lambda = 1000$ is lowest, at 28.46.