Assignment 1: Solutions

1 Probabilistic Modeling (12 marks)

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)*

   Appendix B / Sec. 2.2 in PRML, on the multinomial distribution, is a useful read.

   The parameter is a vector that determines the probability of each party winning. For example, this could be represented as a 4-dimensional vector $\boldsymbol{\mu} \in \Re^4$, where μ_i is the probability that party i wins.

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

   $\boldsymbol{\mu} = (1/4, 1/4, 1/4, 1/4)$

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.*

   $\boldsymbol{\mu} = (0, 0, 1, 0)$

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?*

   Recall that the posterior is proportional to the likelihood times the prior. In this case, the poll results would be data, and one could compute the likelihood of the data.

   If I truly believe the Green Party has perfectly rigged the election, the observed data are inconsistent with this. If I had a model of polls that allowed a party (e.g. the NDP) to win the poll even though they have no chance of winning the election, then this would be fine, and my posterior probability on the parameters μ would remain a Dirac delta function on $\boldsymbol{\mu} = (0, 0, 0, 1)$.

2 Regularized Least-Squares Linear Regression (15 marks)

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

The regularized error function we wish to minimize is:

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2} ||\boldsymbol{w}||^2 \tag{1}$$

We take its derivatives and set them to zero:

$$\nabla E(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n)^T + \lambda \boldsymbol{w}^T \tag{2}$$

$$\begin{aligned}
\nabla E(\boldsymbol{w}) &= 0 && (3)\\
\Leftrightarrow \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\boldsymbol{x}_n)^T &= \left[\sum_{n=1}^{N} \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\boldsymbol{\phi}(\boldsymbol{x}_n)^T\right] + \lambda \boldsymbol{w}^T && (4)\\
\boldsymbol{t}^T\boldsymbol{\Phi} &= \boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{w}^T && (5)\\
\boldsymbol{t}^T\boldsymbol{\Phi} &= \boldsymbol{w}^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I}) && (6)\\
\boldsymbol{\Phi}^T\boldsymbol{t} &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^T\boldsymbol{w} && (7)\\
\boldsymbol{\Phi}^T\boldsymbol{t} &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})\boldsymbol{w} && (8)\\
\boldsymbol{w} &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} && (9)
\end{aligned}$$

3 Training vs. Test Error (12 marks)

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.*

   Yes. 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.

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.*

   No. As discussed in lecture, more complex models can over-fit to a given training dataset, and therefore it is possible that a more complex model (degree 10 polynomial) could have higher testing error compared to a simpler model (degree 9 polynomial).

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

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

4 Regression (40 marks)

4.1 Getting started

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

   Niger, 313.7.

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

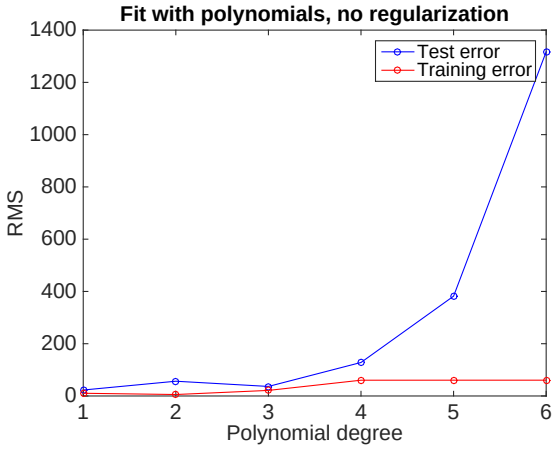
   Sierra Leone, 185.3.

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

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

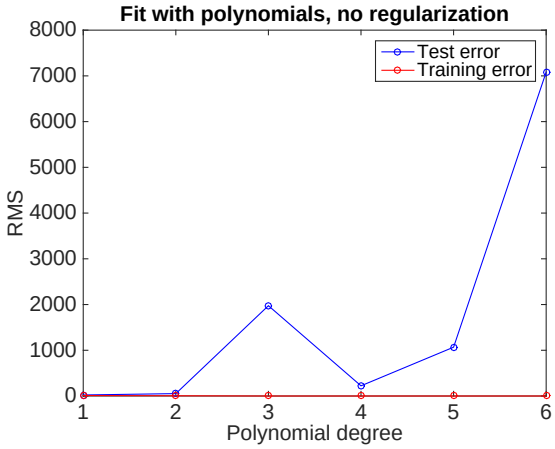
4.2 Polynomial Regression

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

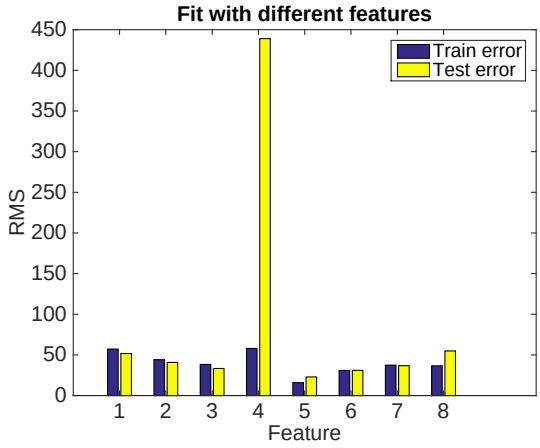


   Note that training error actually increases with larger degree. This is due to numerical instabilities, due to large ranges in the values of inputs.

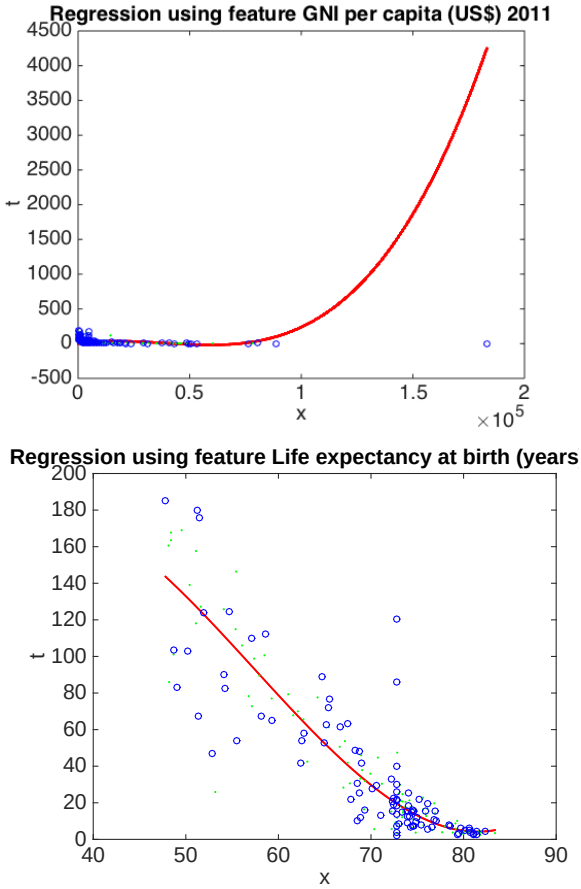
   Results with normalized data are below.

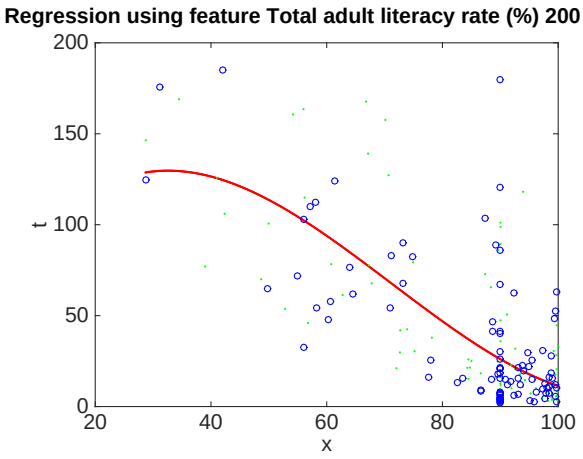


2. Single feature regression. Summary bar chart.



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

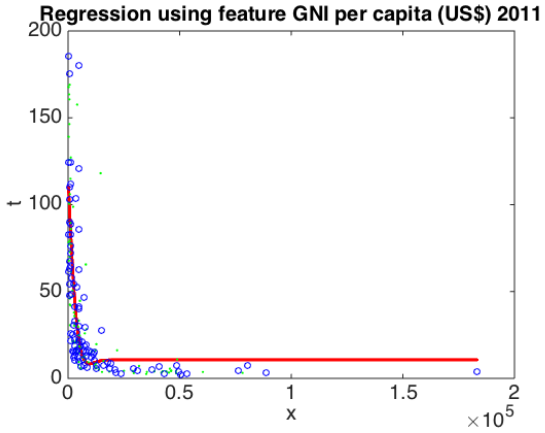




The training error for feature 11 (GNI per capita) is very high. To see what happened, produce plots of the training data points, learned polynomial, and test data points. The code `visualize_1d.m` may be useful.

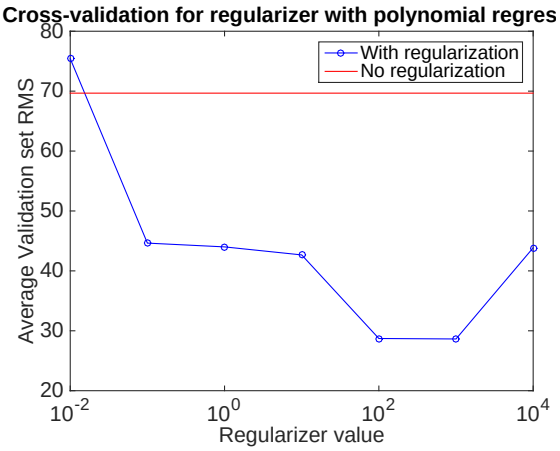
In your report, include plots of the fits for degree 3 polynomials for features 8 (Life expectancy), 11, 12 (Life expectancy), 13 (literacy).

4.3 Sigmoid Basis Functions



Training error is 27.4302, testing error is 30.6816.

4.4 Regularized Polynomial Regression



The value of cross-validation error for $\lambda = 1000$ is lowest, at 28.64.