Assignment 2: Solutions

1 Cross Validation (10 marks)

a) It depends on the data size. If you have maybe 100 examples in total, maybe 1000 examples in total, maybe after 10,000 examples. These sorts of ratios were perfectly reasonable rules of thumb. The goal of the validation set is that you're going to test different algorithms on it and see which algorithm works better. So the validation set just needs to be big enough for you to evaluate, say, two different algorithm choices or ten different algorithm choices and quickly decide which one is doing better. And you might not need a whole 20% of your data for that. So, for example, if you have a million training examples you might decide that just having 10,000 examples in your validation set is more than enough to evaluate which one or two algorithms does better. And in a similar vein, the main goal of your test set is, given your final classifier, to give you a pretty confident estimate of how well it's doing. And again, if you have a million examples maybe you might decide that 10,000 examples are more than enough in order to evaluate a single classifier and give you a good estimate of how well it's doing. So in this example where you have a million examples, if you need just 10,000 for your validation and 10,000 for your test, your ratio will be more like this 10,000 is 1% of 1 million so you'll have 98% train, 1% validation, 1% test.

b)

i) C

ii) A

iii) B

c) Solution: (iii). The model is underfitting, so it would be improved by making it more complex.

2 Regression (40 marks)

2.1 Getting started

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

   Niger, 313.7

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

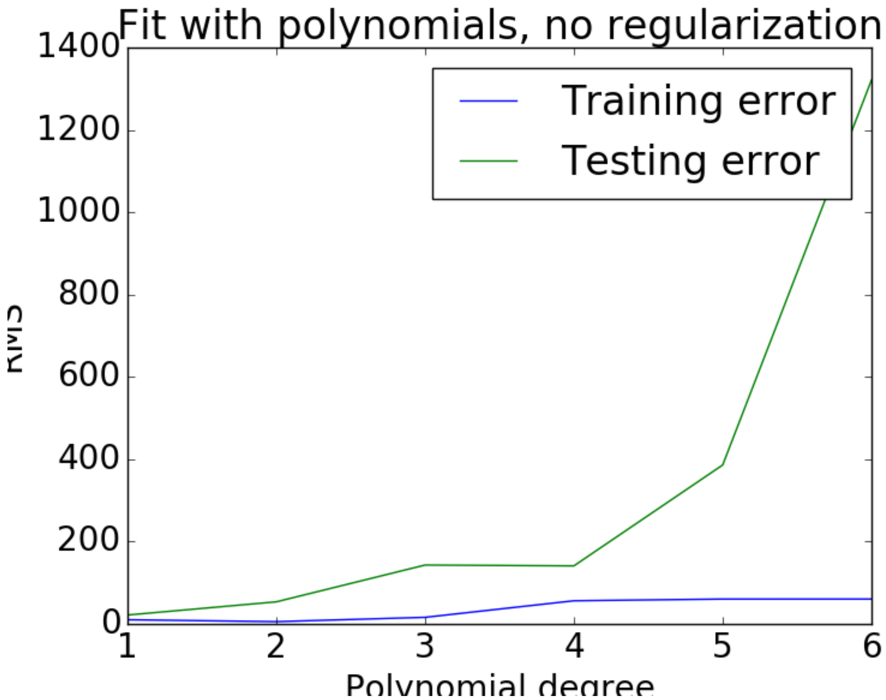
   Sierra Leone, 185.3

3. (2 marks) *Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function `assignment2.load_unicef_data()`?*

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

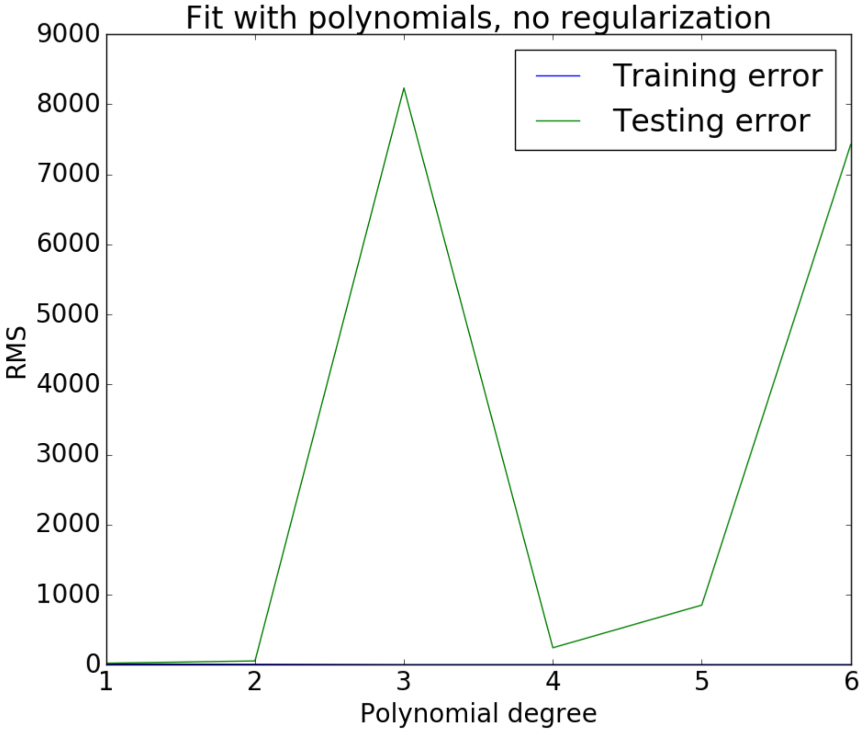
2.2 Polynomial Regression

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

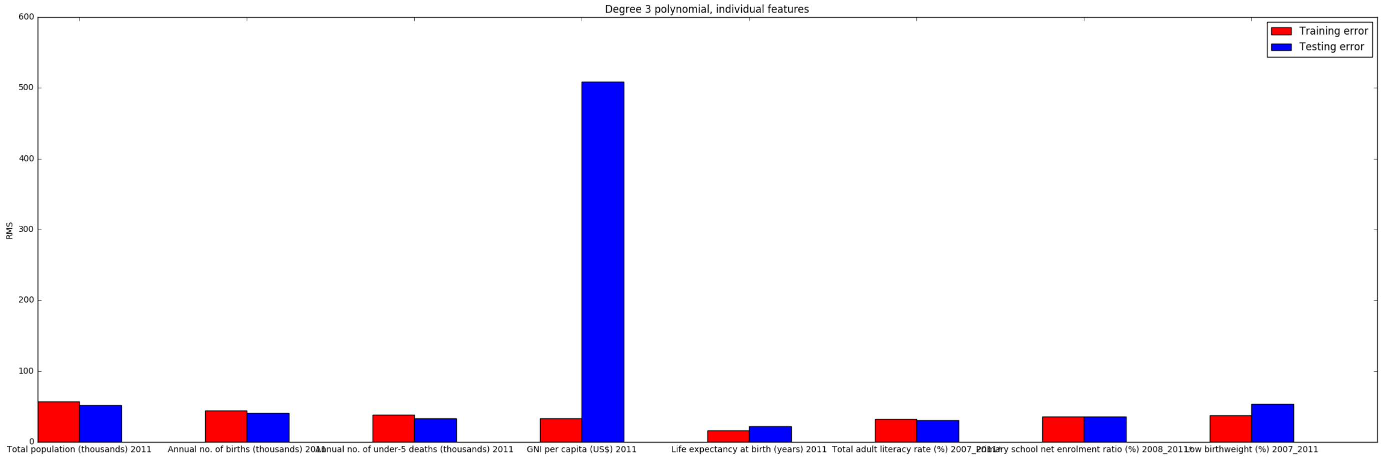


Note that training error actually increases with larger degree. This is due to numerical instabilities, due to large ranges in the values of inputs.

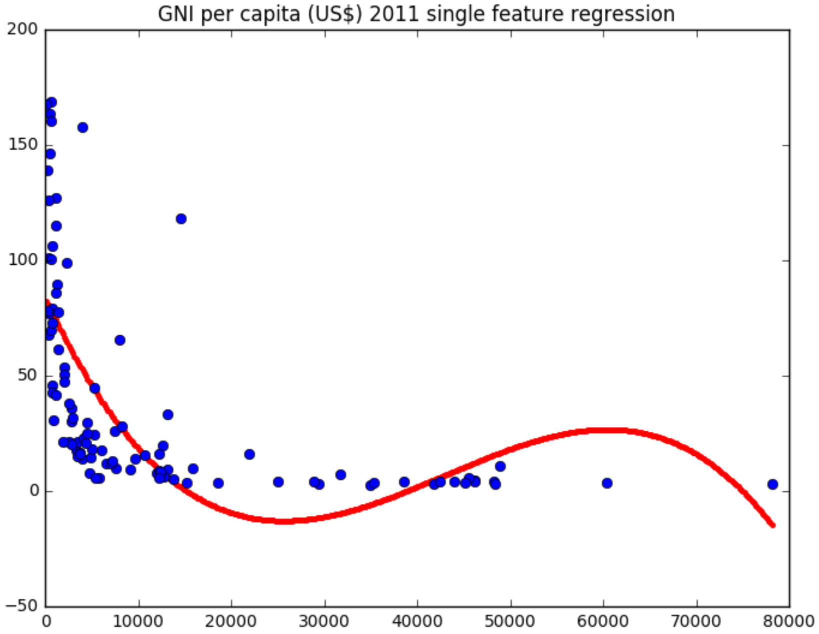
Results with normalized data are below.

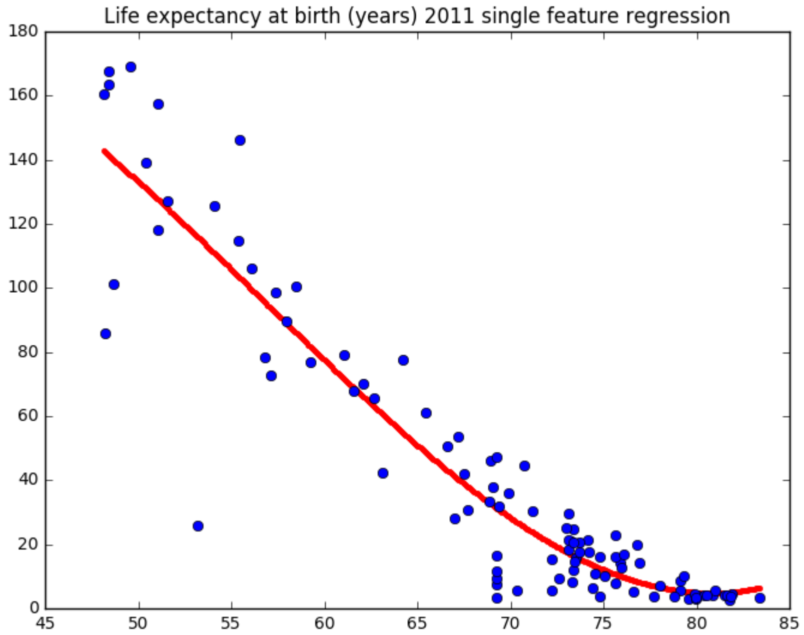


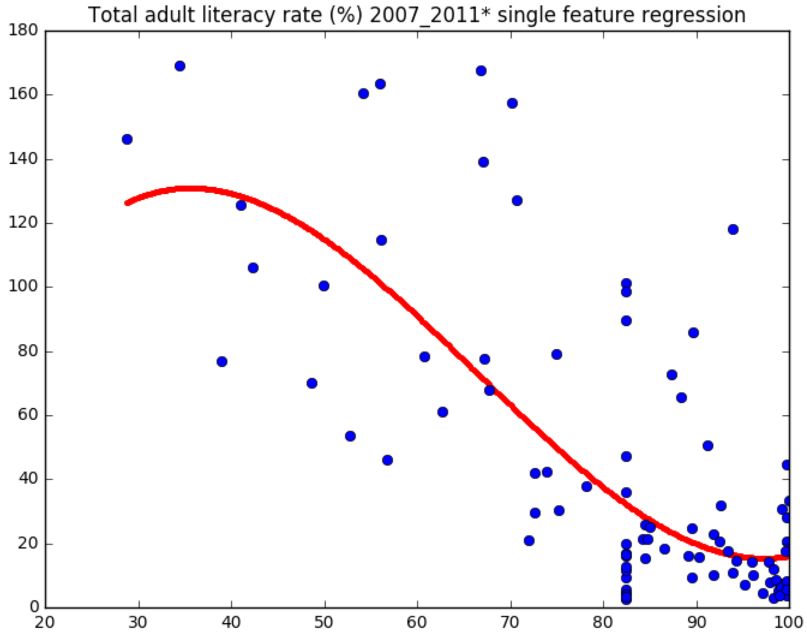
2. Single feature regression. Summary bar chart.



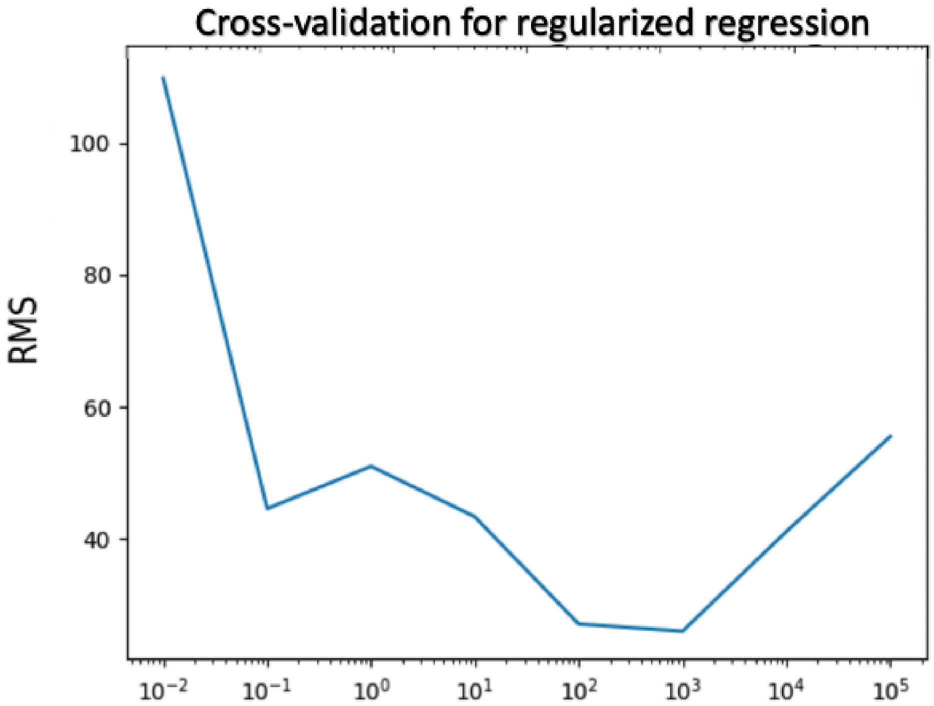
Fits for different features. Note the problems caused by outliers with large values of GNI.




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


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

2.3 Regularized Polynomial Regression



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

3 Probabilistic Modeling and Bayes' Rule (20 marks)

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

(a).

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

(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.

(a).

P (odd) = P(1) + P(3) + P(5) = 0.2 + 0.1 + 0.1 = 0.4.

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

(b).

Considering how we calculate the entropy, the entropy for the biased die is 1.6957 and for a normal die it is 1.7917