Machine Learning
Assignment-1

Kiran Kiran

1 ⇒ Probability Distribution

Ques 1: Ans ⇒ We have 3 parties (NDP, Liberals and GreenParty) Here μ represents a 3 dimensional vector, $\mu = (\mu_1, \mu_2, \mu_3)$.
μ determines the probability of each party winning. Let suppose μ_i gives the probability that party i wins, here $i = 1, 2, 3$.

We know, $\mu = (\mu_1, \mu_2, \mu_3)$

μ_1 = Probability of ~~μ_1~~ NDP wins
μ_2 = Probability of Liberal wins
μ_3 = Probability of GreenParty wins

Ques.2 Ans ⇒ For an election where the outcome is an equal chance of any party winning, μ will be:

$$\mu = (1/3, 1/3, 1/3)$$

Ques.3: Ans ⇒ The value of parameter μ for an election that is completely rigged is below:
if we assume party 1 is currently in power and surely going to win, then $\mu = (1,0,0)$
Likewise, if we assume party 2 is currently in power and definitely going to win, then $\mu = (0,1,0)$
Assuming, party 3 is currently in power and going to win, then $\mu = (0,0,1)$

Ques. 4: Ans ⇒ $P(\mu)$ encodes a belief that one party has rigged the election, but there is an equal chance that is any of the three parties:
Here prior $P(\mu)$ would remain a Dirac delta function on μ.
If party 1 rigged $= \delta(1-\mu_1)\,\delta(\mu_2)\,\delta(\mu_3)$

Ques: 4 Ans ⇒ Continuous ⇒ if party 2 rigged the election = $\delta(\mu_1).6\delta(1-\mu_2)\delta(\mu_3)$

if party 3 rigged the election = $\delta(\mu_1)\delta(\mu_2)\delta(1-\mu_3)$

$$\therefore P(\mu) = \frac{1}{3}\left[\delta(1-\mu_1)\delta(\mu_2)\delta(\mu_3) + \delta(\mu_1)\delta(1-\mu_2)\delta(\mu_3) + \delta(\mu_1)\delta(\mu_2)\delta(1-\mu_3)\right]$$

Ques: 5 Ans ⇒ Suppose my prior is that the Green Party has completely rigged the election and ~~NDP has the~~ I have a model of polls which allowed a party to win the poll even though they have no chance of winning the election, then my posterior probability on μ will be a dirac delta function, where $\mu = (0,0,1)$

We know,

posterior $\propto$ Likelihood $\times$ prior

$$\text{posterior} \propto \underbrace{\delta(\mu_1)\,\delta(\mu_2)}_{\text{NDP \& Liberals win}}\ \underbrace{\delta(1-\mu_3)}_{\text{Green Party wins}}$$

Ques: 6 Ans ⇒ Suppose if i party is selected, they will set university tuition to be t_i dollars.

Given a prior $P(\mu)$, equation for the expected amount tuition will be:

$$\text{Expected amount} = \sum_{i=1}^{3} \mu_i t_i$$

$$= \underline{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3}$$

2: Precision Per Datapoint

The likelihood function with different precision values corresponding to each data point:

$$p(t|X,w,\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T\phi(x_n),\beta_n^{-1})$$

Where β_n represents precision estimates for each data point.

By, Gaussian distribution, we know:

$$\Rightarrow \quad \mathcal{N}(t_n|w^T\phi(x_n),\beta_n^{-1}) = \left(\frac{\beta_n}{2\pi}\right)^{1/2} \exp\left\{\frac{-\beta_n}{2}(t_n-w^T\phi(x_n))^2\right\} \quad -①$$

from equation ①, placing value of $\mathcal{N}(t_n|w^T\phi(x_n),\beta_n^{-1})$ in likelihood function,

$$\Rightarrow \quad p(t|X,w,\beta) = \prod_{n=1}^{N}\left\{\left(\frac{\beta_n}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta_n}{2}(t_n-w^T\phi(x_n))^2\right\}\right\}$$

Taking log of both sides of equation,

$$\Rightarrow \quad \ln(p(t|w^T\phi(x_n),\beta_n^{-1})) = \sum_{n=1}^{N} \ln\left\{\left(\frac{\beta_n}{2\pi}\right)^{1/2} \exp\left\{\frac{-\beta_n}{2}(t_n-w^T\phi(x_n))^2\right\}\right\}$$

$$= \sum_{n=1}^{N}\left\{\frac{1}{2}\ln\beta_n - \frac{1}{2}\ln(2\pi) - \frac{\beta_n}{2}(t_n-w^T\phi(x_n))^2\right\}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\ln\beta_n - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=1}^{N}\beta_n(t_n-w^T\phi(x_n))^2$$

Above equation, represents the relation between the log likelihood function $p(t|w,\beta)$ and the sum of squares error function.

We can not take β_n outside of summation, as it has a dependency on 'n'.

$$\underbrace{\ln(p(t|w^T\phi(x_n),\beta_n^{-1}))}_{\text{log likelihood function}} = \frac{1}{2}\sum_{n=1}^{N}\ln\beta_n - \frac{N}{2}\ln(2\pi) \underbrace{- \frac{1}{2}\sum_{n=1}^{N}\beta_n(t_n-w^T\phi(x_n))^2}_{\text{Sum of squares error function}}$$

5. Training Vs Test Error

Ques. 1 Ans ⇒ If we perform unregularized regression on a dataset, in most of the cases, validation error is always higher than training error. We actually try to minimize the error on training data set while fitting the model. We can a little bit more validation error for the data we have not seen before.

This is, however, not the case always. It is possible for the validation error to be less than training error, if training set contain difficult cases to learn and validation set have easy cases to predict.

Ques. 2 Ans ⇒ Yes, for unregularized regression, training error with a degree 10 polynomial is always less than or equal to that using a degree 9 polynomial. As degree 10 polynomial has more degrees of freedom compared to degree 9 polynomial and it can fit data points even more closely on training set. Which results in lower training error.

However, there may be cases where training error for a degree 10 polynomial is equal to that for a degree 9 polynomial.

Ques. 3 Ans ⇒ Yes, testing error with a degree 20 polynomial always lower using regularized regression compared to unregularized regression.

In regularized regression, there is a penalty term or regularized term (λ) which controls the overfitting issue by encouraging weight values to decay towards zero. Value of 'λ' should be choosen carefully. If value of 'λ' is too low or too high, regularized regression can give similar results to unregularized regression and testing error may be even higher.

4: Basis function Dependent Regularization

For the case where for each weight w_n, we have a different tradeoff parameter λ_n, and a choice among one of L_1 or L_2 regularizer.

$$\nabla E(w) = ?$$

Let J_1 be the set of indices of basis functions whose weights have L_1 regularization, and J_2 be the set of indices of basis functions whose weights have L_2 regularization.

Taking J_1 as a set values which can be 0 or 1, depends on whether the weights having L_1 regularization or not.

So, $J_1 = \{0, 1, 1, 1, 0, 0, \ldots\}$ or $J_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$

Here, 1 denotes that weight is having L_1 regularization. Similarly,

$$J_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

J_2 contains the values which can be 1 if weight corresponds to L_2 regularizer, and 0 if not.

$$E(w) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2}_{\text{sum of squared error}} + \frac{1}{2}\underbrace{\sum_{m=1}^{M}\{J_{1m}\lambda_m w_m + J_{2m}\lambda_m w_m^2\}}_{\text{regularizer}} \quad -①$$

N is the number of observations and M be the number of coefficients.

J_{1m} is the mth element of J_1 which will be 1 if its weight corresponds to L1 regularizer, otherwise 0.

J_{2m} is the m^{th} element of J_2 which will be 1 if its weight corresponds to L2 regularizer, otherwise 0.

ω_m is the mth coefficient.

$$\nabla E(\omega) = \sum_{n=1}^{N} \{ t_n - \omega^T \phi(x_n) \} \phi(x_n)^T + \frac{1}{2} \sum_{m=1}^{M} \{ J_{1m} \lambda_m + 2 J_{2m} \lambda_m \omega_m \}$$

Here $\sum_{n=1}^{N} t_n = t$ & $\sum_{n=1}^{N} \phi(x_n) = \phi$

$$\therefore \nabla E(\omega) = t\phi^T - \omega^T \phi \phi^T + \frac{1}{2} \{ J_1 \lambda_{L_1} + 2 J_2 \lambda_{L_2} \omega^T \}$$

Here ϕ is a $N \times M$ matrix.

$$\phi := \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \phi_{M-1}(x_2) \\ \vdots & \vdots & \\ \phi_0(x_N) & \phi_1(x_N) & \phi_{M-1}(x_N) \end{pmatrix}$$

So the gradient for the regularized squared error will be:

$$\boxed{\nabla E(\omega) = t\phi^T - \omega^T \phi \phi^T + \frac{1}{2} \{ J_1 \lambda_{L_1} + 2 J_2 \lambda_{L_2} \omega^T \}}$$

Here, J_1 is the set of indices of basis functions whose weights have L1 regularization and J_2 is the set of indices of basis functions whose weights have L2 regularization.

By setting gradient to zero, we can get the value of ω.

5.1 Getting Started

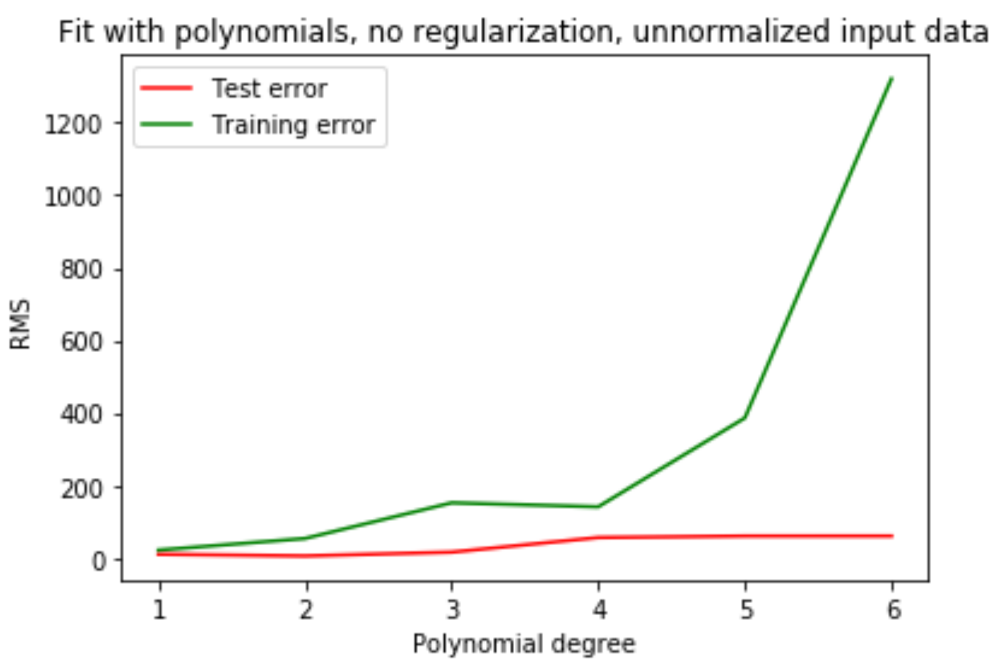
1. Which country had the highest child mortality rate in 1990? What was the rate?
Ans: The country which had the highest child morality rate is Niger. The rate was 313.7.

2. Which country had the highest child mortality rate in 2011? What was the rate?
Ans: In the year 2011, Sierra Leone had the highest child mortality rate. The rate was 185.3.

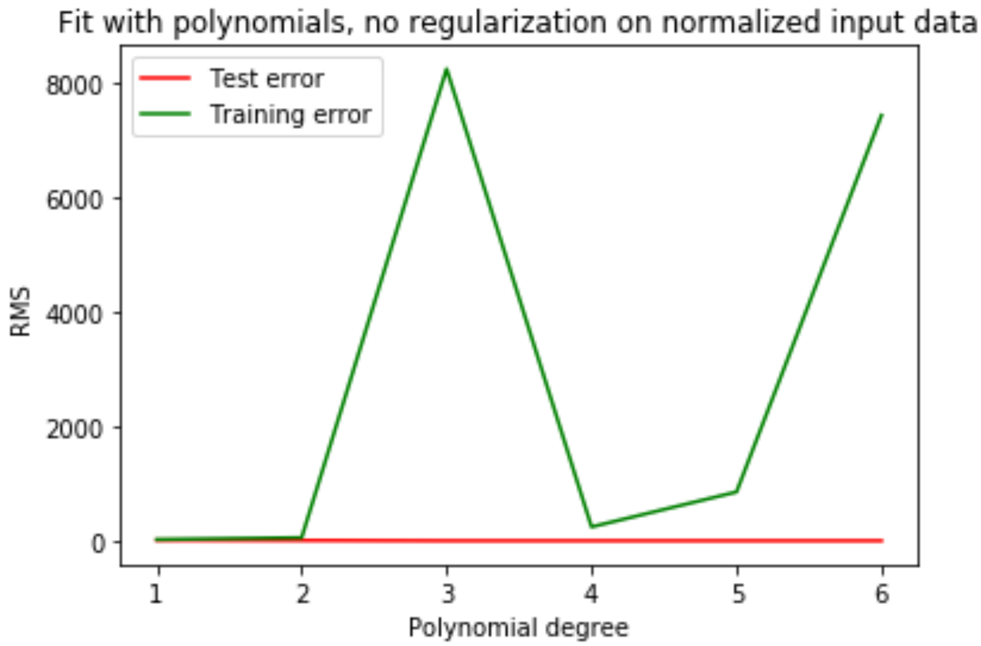
3. Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function assignment1.load unicef data()?
Ans: The function assignment1.load unicef data() handled the missing features by replacing them with the mean of values of feature for other countries.

5.2 Polynomial Regression

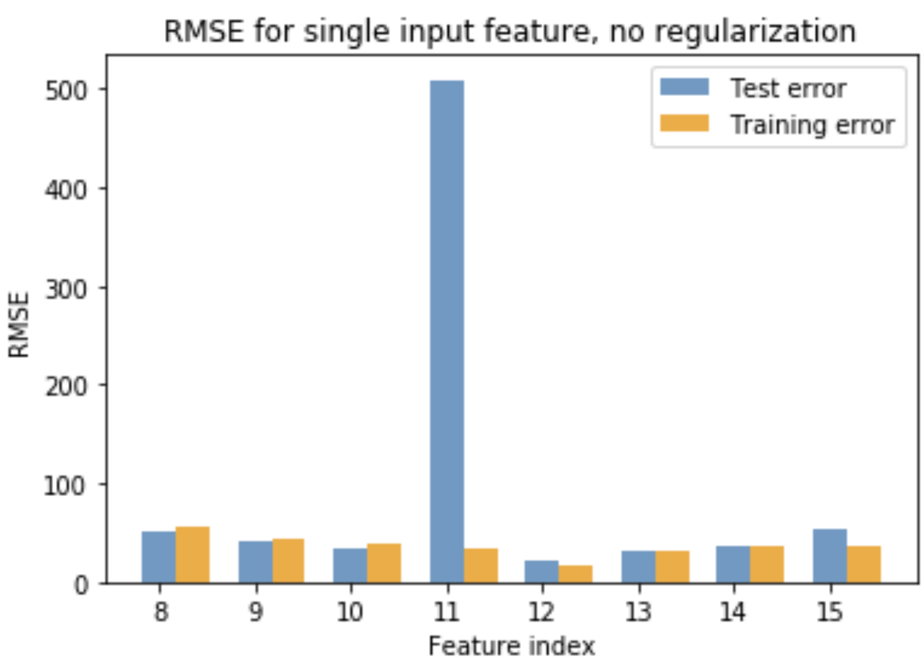
1.1) Plot training error and test error (in RMS error) versus polynomial degree.



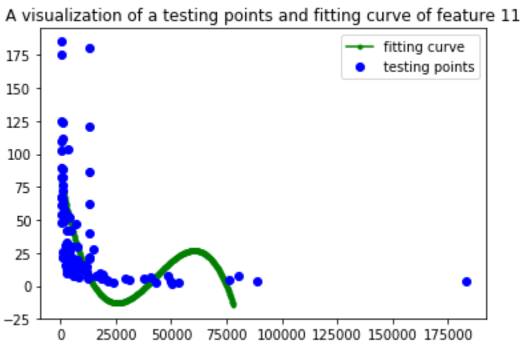
1.2) Normalize the input features before using them (not the targets, just the inputs x).

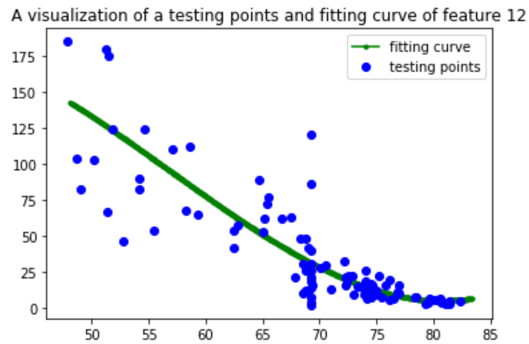


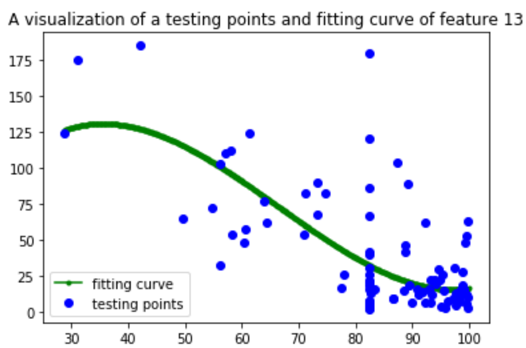
2.1) Plotted the training error and test error (in RMS error) for each of the 8 features using a bar chart.



2.2) Plotted the visualization of testing points and fitting curve of feature 11-13. The testing error for feature 11 (GNI per capita) is very high.

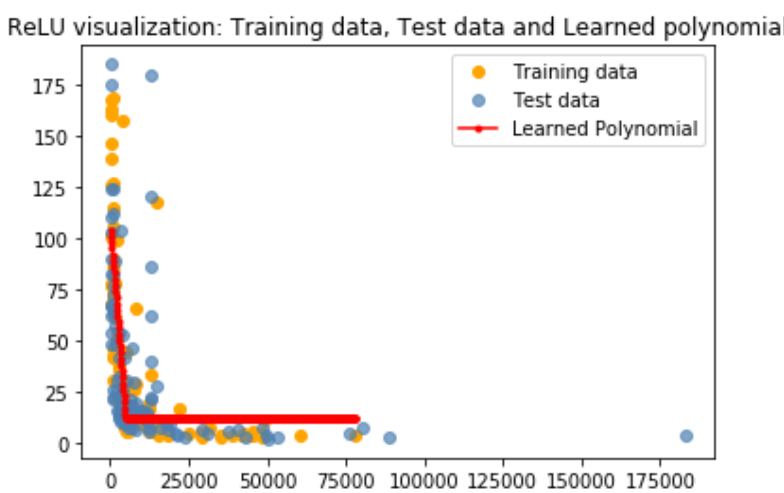




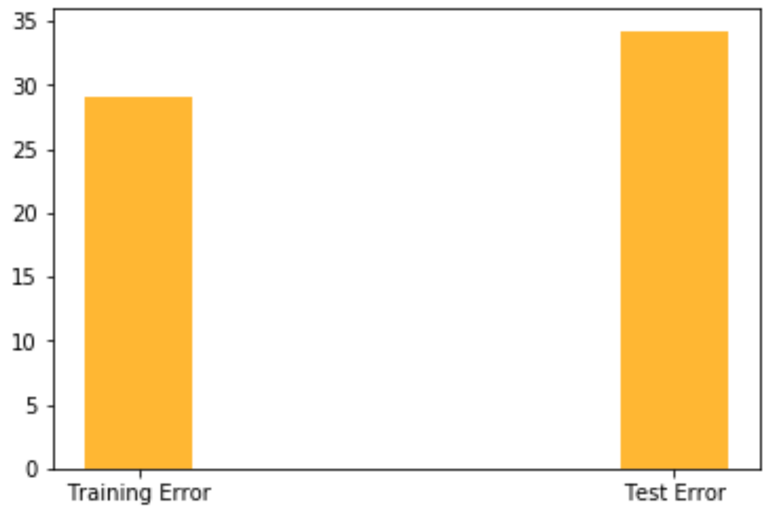


5.3 ReLU Basis Function

3.1) Plotted of the fit for feature 11 (GNI).



3.2) Plotted the training and testing error for this regression model.



<u>5.4 Regularized Polynomial Regression</u>

4.1) Plotted the plot of average validation set error versus λ(regularizer). The value of cross-validation error for λ = 1000 is lowest, at 28.64.