ML : Assignment 2

Anshuman Suri (2014021)

Note: All plots are in *plots* folder of the assignment (there were too many, and weren't fitting here well. They're structured according to the part of the question)

Theta initialization : all weights set to zero (gave better performance that random normalized weights)

Question 4

(a) The mean square error decreases with increase in the percentage of data (for both the data-sets). This shows that more ground truth leads to better predictions (which is in alignment with what one would expect, as long as we don't overfit).
Hyper-parameters
- **(i)** Lin : Learning rate : 1e-3
- **(ii)** Sph : Learning rate : 1e-4

(b) The mean square error decreases with increase in the percentage of data (for both the data-sets) for both polynomial and gaussian kernels. In both the cases, polynomial kernel seems to perform better than the gaussian kernel. The degree of polynomial and sigma in the gaussian function were both chosen after trying out a lot of different values for both of them.
Hyper-parameters
- **(i)** Lin (polynomial) : Learning rate : 4e-7, Degree : 2
- **(ii)** Lin (gaussian) : Learning rate : 1e-2, Sigma : 1.0
- **(iii)** Sph (polynomial) : Learning rate : 1e-9, Degree : 2
- **(iv)** Sph (gaussian) : Learning rate : 1e-1, Sigma : 1.0

(c) For the *lin* data-set, neither of the 3 fits is good enough. However, given the distribution of data, none of them is expected to be a good fit (except maybe Gaussian, if we were to map it to higher dimensions).
For the *sph* data-set, the performance of polynomial kernel is much better than the other two, fitting the curve almost quite good.

(d) For ridge regression, polynomial kernel of degree was chose.
As expected, increasing the value of delta leads to a decrease in the values of theta, which converge to zero if delta is too large (as expected).
Choosing delta as zero for both the data-sets gives the least error. However, the purpose of using delta is to avoid overfitting, so an increase in error is expected.
The value of delta was chosen by trial such that increasing it by a factor of 10 or more

gave absurd values (inf/nan,etc).
Hyper-parameters

(i) Lin : Learning rate : 4e-7, Delta : 10

(ii) Sph : Learning rate : 1e-9, Delta : 100

(e) (MSE v/s percentage of data plots have also been plotted as a sanity check)
Linear regression (with linear kernel) and ridge regression. Values of delta were chosen in a similar way as in part (d).
Performing 10-fold cross validation yields the following values:

(i) Iris (mean: 6.06342192453, variance: 0.0889601662237)

(ii) Seeds (mean: 6.26038434857, variance: 1.18842322702)

(iii) Air (mean: 1567.45400344, variance: 3220074.38105)

Hyper-parameters

(i) Iris (learning rate: 1e-2, delta: 1000)

(ii) Seeds (learning rate: 1e-3, delta: 10000)

(iii) Air (learning rate: 1e-6, delta: 0)