CPSC 6430: Machine Learning Implementation and Evaluation

Project 2: Linear Regression – Model Prediction and Evaluation

Student: Adithya Suresh, C18590622

Choosing a Model for Predicting on Unseen Data:

Goal:

Using the regression program from Project 1, model has to be chosen between linear, quadratic and cubic model and to predict future times for the women's Olympic 100-meter race dataset.

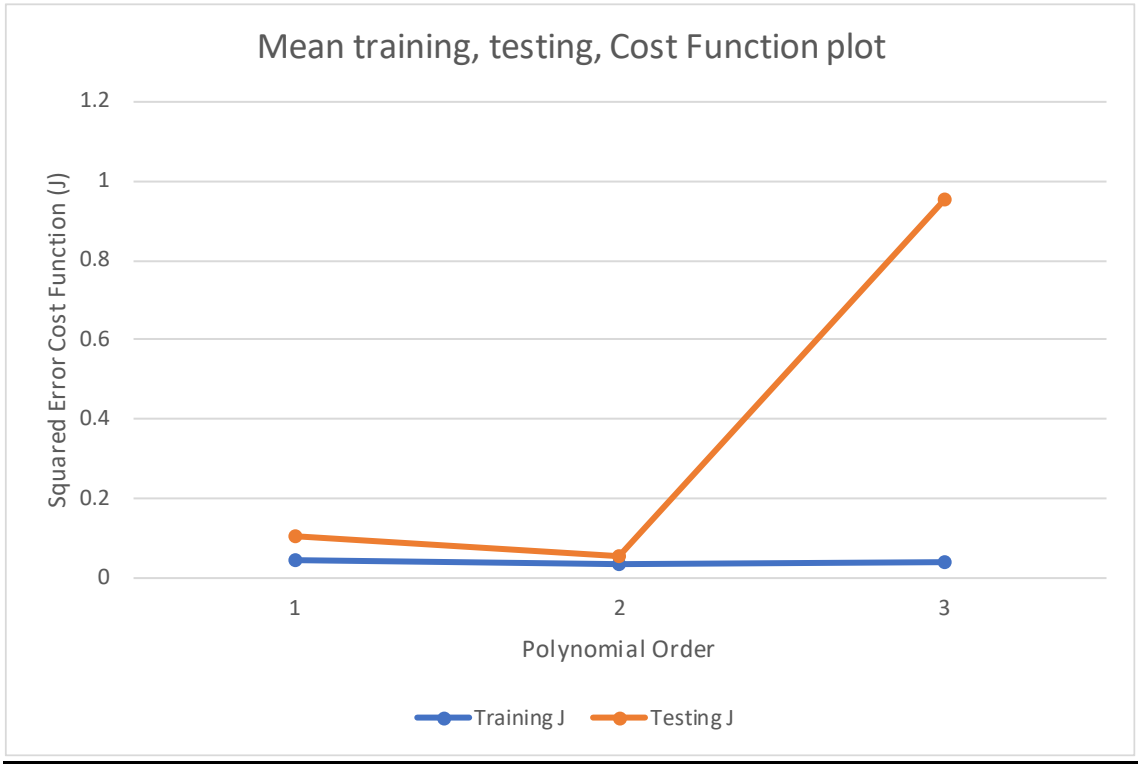
Model selection process:

The Women's Olympic 100-meter race data is taken, split into respective folds and trained with weight calculation and J (cost function) value was found for different X matrices which contains x_1, x_1^2, x_1^3 from order one to three – polynomial degree. The J values for training, and testing are finally tabulated and the mean J value for both training and testing are plotted with respect to the polynomial degree.

Chart with J value for multiple folds:

		J value		
		Linear	Quadratic	Cubic
Folds	1234	0.038363881	0.035470849	0.034935869
	5	0.176175883	0.067758297	0.035102454
	1235	0.046015392	0.032789809	0.032693591
	4	0.064397254	0.051659399	0.05729373
	1245	0.053637365	0.04265187	0.041095964
	3	0.031457255	0.007397201	0.009454863
	1345	0.049941619	0.03514046	0.034413505
	2	0.042401024	0.036221937	0.034103128
	2345	0.032835722	0.025533753	0.041614841
	1	0.199418976	0.115478451	4.63690816
Mean for training		0.044158796	0.034317348	0.036950754
Mean for testing		0.102770078	0.055703057	0.954572467

Plot of Mean training and testing J value with polynomial degree up to 3:



Model selection reason:

From the above, chart we can infer that the mean error value for the quadratic model's testing set is comparatively lower than both linear and cubic models. When the plot is inferred, while reaching the third polynomial degree, the mean testing error is high and raises upwards, which means high variance. The smallest variance was achieved, and the error produced with the quadratic model is less. Hence the quadratic model has a comparatively low bias, with the variance between training and testing being less. This makes the quadratic model of regression be the best choice for predicting the values for new data.

Code summary:

The code will take the entire W100MTimes.txt dataset and train with quadratic regression model and the weights are calculated. Then the code prompts for the user to feed the input year for winning women's prediction time. The data will be taken as 90 for 1990 and 08 for 2008 (examples) and the prediction equation, $hw(x) = w0 + w_1x + w_2x^2$ will be predict the race time for that year.