CSCE 633: Machine Learning Assignment 1

Sameer Kumar Behera - (UIN: 526004296)

2018/09/28

1 Polynomial Regression : Bias-Variance Tradeoff

The d-degree polynomial regression curves were plotted with d = 0, 1, 2, 3 in the following figure.

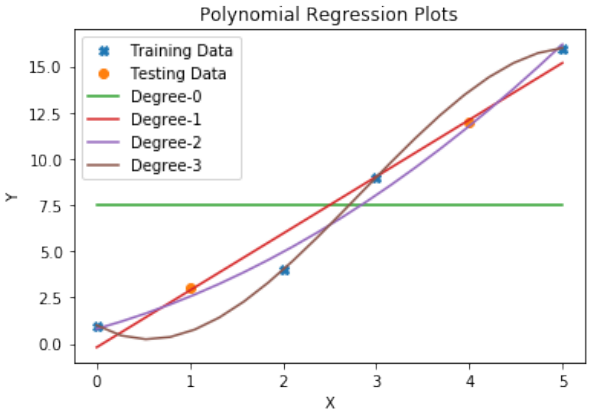

Figure 1: Polynomial Regression Curves with Varying Degree

In bias-variance trade-off, we are calculating the expected error at a given and fixed point x for the model. Here, the expectation is calculated over different choices of training set, i.e. the randomness comes from the data distribution and the noise in the ground truth model. Hence, it doesn't make sense to calculate the bias, variance on a fixed training set without considering the noises.

In statistics and machine learning, the biasvariance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa. The biasvariance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.

Training, Test and Total Errors for d = 0, 1, 2, 3 were as follows:

```
Training Error Data for Degree 0 is 129.0
Training Error Data for Degree 1 is 5.923076923076925
Training Error Data for Degree 2 is 1.9230769230769231
Training Error Data for Degree 3 is 5.546185201769476e-28


Test Error Data for Degree 0 is 40.5
Test Error Data for Degree 1 is 0.02662721893491167
Test Error Data for Degree 2 is 0.24884944115713056
Test Error Data for Degree 3 is 8.222222222222207


Total Error Data for Degree 0 is 169.5
Total Error Data for Degree 1 is 5.949704142011837
Total Error Data for Degree 2 is 2.1719263642340536
Total Error Data for Degree 3 is 8.222222222222207
```

The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data than the intended outputs (overfitting).

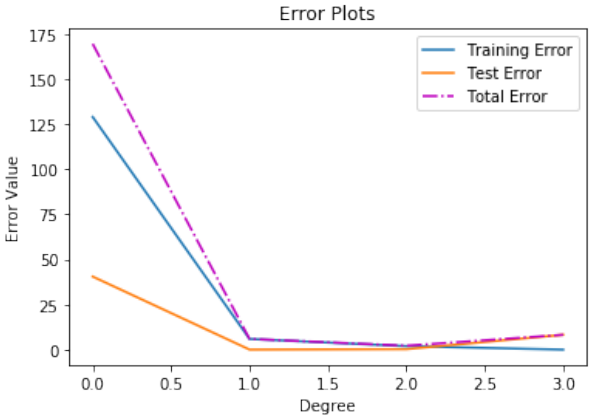

Figure 2: Error Plots with Varying Degrees

As we can see from the figure above, with increasing degrees, the errors reduce dramatically as the model prediction gets better on higher degrees. However, the from d=2 to d=3, the test error increases slightly due to the overfitting of the model on training data.

2 Probability

2.1 Bayes Theorem

$P(disease \mid test) = P(test \mid disease) * P(disease) \ / \ P(test)$
$= P(test \mid disease) * P(disease) \ / \ [P(test \mid disease) * P(disease) + P(test \mid disease!) * P(disease!)\ ]$
$= 0.92 * 0.00004 / [0.92 * 0.00004 + 0.08 * 0.99996]$
$= 0.0000368 / [0.0000368 + 0.0799968]$
$= 0.0000368 / 0.0800336$
$= 0.00045980688$

2.2 Maximum Likelihood Estimation

We are given that:

$$y_i = \omega^T x_i + \epsilon_i \tag{1}$$

where $\epsilon_i \sim$N(0, σ^2)

This implies that $y_i - \omega^T x_i$ is a gaussian function with mean 0 and standard deviation σ.

$$f(x_i|\omega, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp \frac{-(y_i - \omega x_i)^2}{2\sigma^2} \tag{2}$$

The likelihood is defined as

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} * \exp \frac{-(y_i - \omega x_i)^2}{2\sigma^2} \tag{3}$$

The constant terms can be pulled out.

$$L = (\frac{1}{\sqrt{2\pi\sigma^2}})^N \prod_{i=1}^{N} \exp \frac{-(y_i - \omega x_i)^2}{2\sigma^2} \tag{4}$$

Taking log on both the sides, we get

$$l = \log(L) = \log((\frac{1}{\sqrt{2\pi\sigma^2}})^N) - \sum_{i=1}^{N} \frac{(y_i - \omega x_i)^2}{2\sigma^2} \tag{5}$$

$$l = N * \log(\frac{1}{\sqrt{2\pi\sigma^2}})) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \omega x_i)^2 \tag{6}$$

To maximize the likelihood the derivative of l should be 0.

$$\frac{\partial l}{\partial \omega} = 0 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2 * x_i * (y_i - \omega x_i) = 0 \tag{7}$$

$$\frac{\partial l}{\partial \omega} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i (y_i - \omega x_i) = 0 \tag{8}$$

Hence, the Condition to Maximize the Likelihood is,

$$\sum_{i=1}^{N} x_i (y_i - \omega x_i) = 0 \tag{9}$$

The RSS Minimization is given as,

$$RSS = \sum_{i=1}^{N} (y_i - \omega x_i) = 0 \tag{10}$$

which on differentiation arrives at the previous equation (9).

Hence RSS Minimization is Equivalent to the MLE Condition.

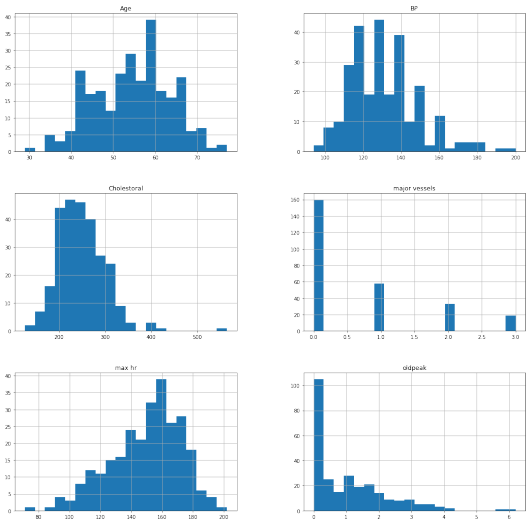

Figure 3: Plotting Histograms of All Non-Categorical Features

3 Models for Heart Disease

3.1 Data Exploration

In the given dataset, the outcome for 'heart disease' has the following features: 'Age', 'Sex', 'Chest Pain', 'BP', 'Cholestoral', 'fasting blood sugar >120', 'resting ECG', 'max hr', 'angina', 'oldpeak', 'slope', 'major vessels' and 'defect'.
The Non-Categorical Features, i.e. 'Age', 'BP', 'Cholestoral', 'major vessels', 'max hr' and 'oldpeak', etc. are Real-Valued.
A Categorical Variable can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property, i.e. 'Sex', 'Chest Pain', 'fasting blood sugar >120', 'resting ECG', 'angina', 'slope', 'major vessels' and 'defect'.

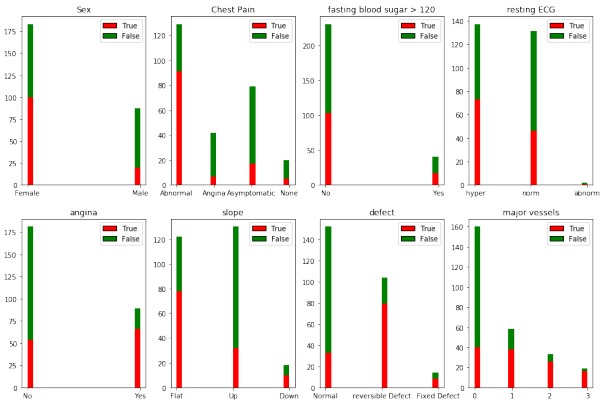

Figure 4: Plotting Stacked Histograms of Discrete Variables

A Continuous Variable has an infinite number of possible values, i.e. 'Age', 'BP', 'Cholestoral', 'max hr' and 'oldpeak'.

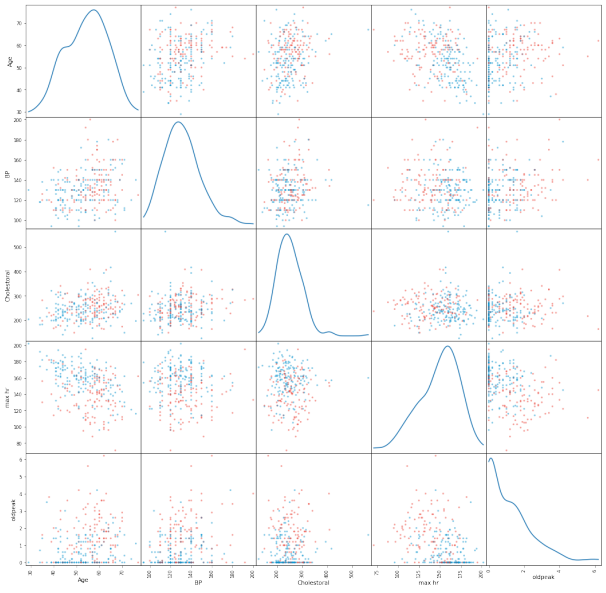

Figure 5: Plotting Scatter Plots of Continuous Variables in Scatter Matrix

3.2 Prediction

Before prediction, some features are hot-encoded by which categorical variables which do not have any certain order are converted into a form that could be provided to ML algorithms to do a better job in prediction, i.e. 'Chest Pain' and 'defect'. Rest all the other categorical variables had some order and weren't hot-encoded.
Definitely, all inputs are not necessary as the dependency of the outcome to some of the variables is non-existent. The t-test and its probability can be used to remove some unnecessary variables. The higher the P-value of the test, the lesser is its necessity.

OLS Regression Results

Dep. Variable:	heart disease	R-squared:	0.560
Model:	OLS	Adj. R-squared:	0.525
Method:	Least Squares	F-statistic:	15.85
Date:	Fri, 28 Sep 2018	Prob (F-statistic):	1.42e-27
Time:	19:03:35	Log-Likelihood:	-66.669
No. Observations:	216	AIC:	167.3
Df Residuals:	199	BIC:	224.7
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1917	0.214	0.894	0.373	-0.231	0.615
Age	-0.0032	0.003	-1.036	0.301	-0.009	0.003
Sex	0.1503	0.059	2.547	0.012	0.034	0.267
BP	0.0029	0.001	2.007	0.046	5.08e-05	0.006
Cholestoral	0.0003	0.000	0.517	0.606	-0.001	0.001
fasting blood sugar > 120	-0.0634	0.069	-0.916	0.361	-0.200	0.073
resting ECG	0.0298	0.025	1.207	0.229	-0.019	0.078
max hr	-0.0028	0.001	-2.158	0.032	-0.005	-0.000
angina	0.0942	0.058	1.612	0.108	-0.021	0.209
oldpeak	0.0368	0.028	1.294	0.197	-0.019	0.093
slope	-0.0352	0.051	-0.689	0.492	-0.136	0.066
major vessels	0.1539	0.028	5.497	0.000	0.099	0.209
Chest Pain_Abnormal	0.2152	0.064	3.375	0.001	0.089	0.341
Chest Pain_Angina	0.0321	0.075	0.430	0.668	-0.115	0.179
Chest Pain_Asymptomatic	0.0153	0.065	0.235	0.814	-0.113	0.144
Chest Pain_None	-0.0710	0.096	-0.742	0.459	-0.260	0.118
defect_Fixed Defect	0.0717	0.111	0.648	0.518	-0.147	0.290
defect_Normal	-0.0569	0.079	-0.724	0.470	-0.212	0.098
defect_reversible Defect	0.1768	0.088	2.010	0.046	0.003	0.350

Omnibus:	4.348	Durbin-Watson:	2.150
Prob(Omnibus):	0.114	Jarque-Bera (JB):	3.954
Skew:	0.304	Prob(JB):	0.139
Kurtosis:	3.264	Cond. No.	4.73e+18

Figure 6: Linear Regression Summary

Logit Regression Results

Dep. Variable:	heart disease	No. Observations:	216
Model:	Logit	Df Residuals:	199
Method:	MLE	Df Model:	16
Date:	Fri, 28 Sep 2018	Pseudo R-squ.:	0.5388
Time:	19:03:35	Log-Likelihood:	-68.429
converged:	True	LL-Null:	-148.38
		LLR p-value:	8.568e-26

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4454	2.36e+07	1.89e-08	1.000	-4.62e+07	4.62e+07
Age	-0.0315	0.030	-1.044	0.296	-0.091	0.028
Sex	1.4460	0.642	2.253	0.024	0.188	2.704
BP	0.0302	0.013	2.292	0.022	0.004	0.056
Cholestoral	0.0062	0.005	1.340	0.180	-0.003	0.015
fasting blood sugar > 120	-0.4329	0.715	-0.606	0.545	-1.834	0.968
resting ECG	0.2190	0.228	0.962	0.336	-0.227	0.665
max hr	-0.0227	0.013	-1.800	0.072	-0.047	0.002
angina	0.6824	0.504	1.354	0.176	-0.305	1.670
oldpeak	0.3160	0.272	1.164	0.245	-0.216	0.848
slope	-0.3986	0.460	-0.867	0.386	-1.300	0.503
major vessels	1.3651	0.318	4.298	0.000	0.743	1.988
Chest Pain_Abnormal	-1.7534	2.08e+07	-8.45e-08	1.000	-4.07e+07	4.07e+07
Chest Pain_Angina	-2.9984	2.11e+07	-1.42e-07	1.000	-4.13e+07	4.13e+07
Chest Pain_Asymptomatic	-3.4172	2.08e+07	-1.64e-07	1.000	-4.08e+07	4.08e+07
Chest Pain_None	-3.8644	2.07e+07	-1.86e-07	1.000	-4.06e+07	4.06e+07
defect_Fixed Defect	-0.9208	nan	nan	nan	nan	nan
defect_Normal	-1.3345	nan	nan	nan	nan	nan
defect_reversible Defect	0.2050	nan	nan	nan	nan	nan

Figure 7: Logistic Regression Summary

3.2.1 Mixed Pass Method

I used Forward Backward (Mixed) Pass Method to select the optimal model with least RSE. Starting with an empty model, I kept on adding features iteratively and simultaneously the unnecessary variables got removed whose P-value were greater than 0.01.
The List of Selected Variables in the Final Optimum Linear Model were: 'oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect'.
The List of Selected Variables in the Final Optimum Logistic Model were: 'angina', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect'.

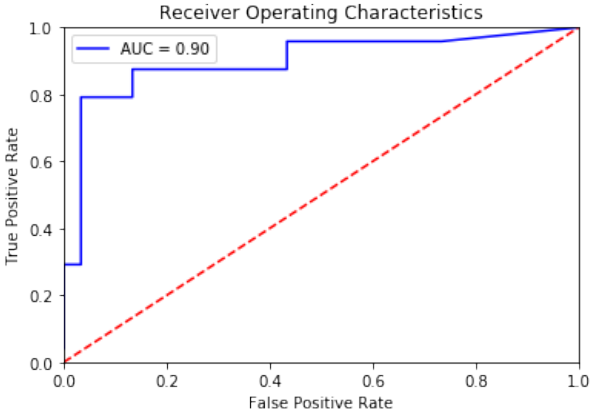

Figure 8: Linear Regression ROC: The AUROC is: 0.9

Also, Optimum Threshold for Linear Model is 0.342908598788742 for Max F1 Score as 0.8571428571428572

The Optimum Threshold for Logistic Model is 0.36491209452987045 for Max F1 Score as 0.8695652173913043

3.2.2 5-Fold Cross Validation

Yes, the features change in every fold.
K = 1
The Feature Selection of the Optimised Linear Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
The Feature Selection of the Optimised Logistic Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal',

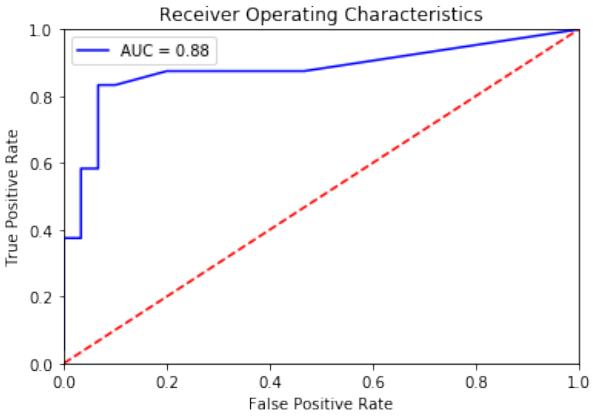

Figure 9: Logistic Regression ROC: The AUROC is: 0.8784722222222222

'defect_reversible Defect']
K = 2
The Feature Selection of the Optimised Linear Model is ['Sex', 'angina', 'oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
The Feature Selection of the Optimised Logistic Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
K = 3
The Feature Selection of the Optimised Linear Model is ['Sex', 'max hr', 'oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
The Feature Selection of the Optimised Logistic Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
K = 4
The Feature Selection of the Optimised Linear Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
The Feature Selection of the Optimised Logistic Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
K = 5
The Feature Selection of the Optimised Linear Model is ['oldpeak', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']
The Feature Selection of the Optimised Logistic Model is ['angina', 'major vessels', 'Chest Pain_Abnormal', 'defect_reversible Defect']

Mean of the AUROCs of Linear Model: 0.8864606920856921
Confidence Interval Lower Bound of the AUROCs of Linear Model: 0.8215454467570331
Confidence Interval Lower Bound of the AUROCs of Linear Model: 0.9513759374143511
Mean of the AUROCs of Logistic Model: 0.8862037962037961
Confidence Interval Lower Bound of the AUROCs of Logistic Model: 0.8249646265977543
Confidence Interval Lower Bound of the AUROCs of Logistic Model: 0.9474429658098379
Mean of the F1 Scores of Linear Model: 0.4060831761321323
Confidence Interval Lower Bound of the AUROCs of Linear Model: 0.36684575279481224
Confidence Interval Lower Bound of the AUROCs of Linear Model: 0.44532059946945235
Mean of the F1 Scores of Logistic Model: 0.40898041743358543
Confidence Interval Lower Bound of the AUROCs of Logistic Model: 0.3654473745851148
Confidence Interval Lower Bound of the AUROCs of Logistic Model: 0.4525134602820561