Assignment 1 (ML)

Akhil Babu Manam (927000968)

September 29, 2018

1 Bias-Variance Tradeoff

1.1

The plot in fig.1 depicts the training points, testing points and the regressed curves along with the labels for each of them. The equations for the predicted polynomial fit models are as follows :

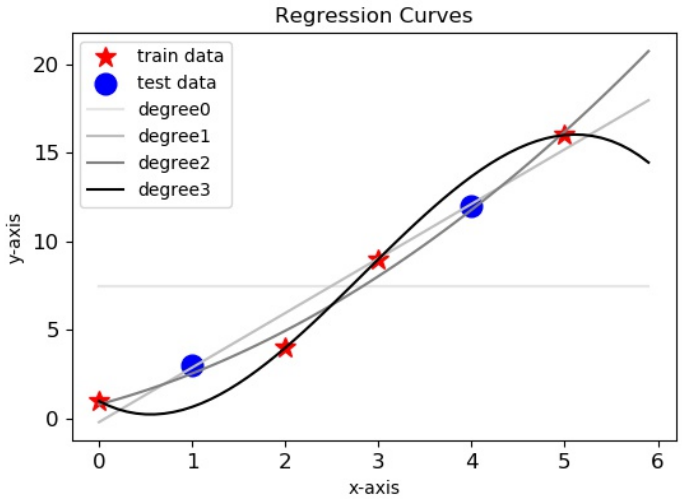

Figure 1: Polynomial regression curves

1.2

The plot in fig.2 depicts the squared bias (black), variance (gray), total error (green), train error (red) and test error (blue).

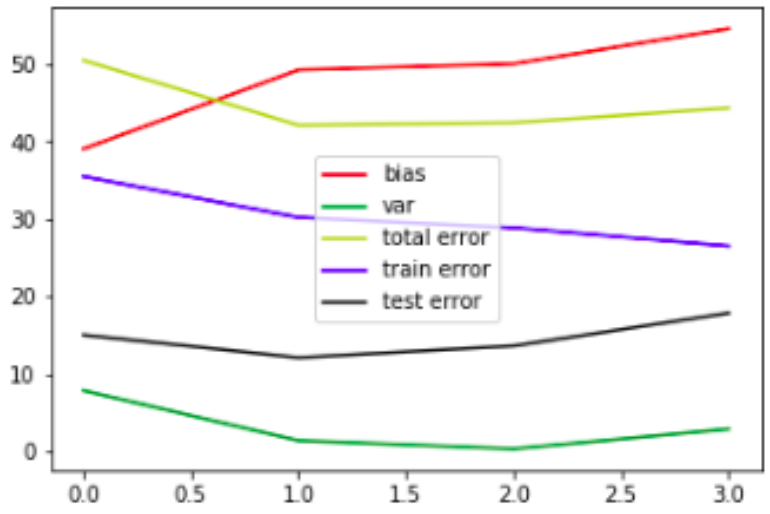

Figure 2: Bias-Variance and Error plots

1.3

According to the bias-variance tradeoff, as the squared bias of the model increases, the complexity of the model reduces and vice versa. In other words, as the model bias on the train data reduces, the model tries to learn even the noise which is present in the training data and hence the complexity of this model increases. Hence, there is a trade-off between bias and variance, it should have been clear from the plot for question 2 that as bias reduces, variance of the model increase. Also, it can be seen that while train error keeps reducing, total error and test error increase with the complexity of the model after a while.

2 Question and Proof

2.1

Let us depict tested positive as TP, tested negative as TN, having disease as D and not having disease as ND.

From the information that has been provided,
$\mathbb{P}(TP/D) = 0.92$
$\mathbb{P}(TN/ND) = 0.92$
$\mathbb{P}(D) = 0.00004$
$\mathbb{P}(ND) = 0.99996$

By Bayes rule,

$$\mathbb{P}(D/TP) = \frac{\mathbb{P}(TP, D)}{\mathbb{P}(TP)} = \frac{\mathbb{P}(\frac{TP}{D})\mathbb{P}(D)}{\mathbb{P}(TP)} \tag{1}$$

Now, let us calculate $\mathbb{P}(TP)$,
$\mathbb{P}(TP) = \mathbb{P}(\frac{TP}{D})\mathbb{P}(D) + \mathbb{P}(\frac{TP}{ND})\mathbb{P}(ND)) = 0.92 * 0.00004 + 0.08 * 0.99996 = 0.0800336$

By substituting the appropriate values,
$\mathbb{P}(\frac{D}{TP}) = 0.92 * 0.00004/0.0800336 = 0.000459 = 0.00046$

So, the chance level of the patient having the disease after testing positive is 1 in 2174 or the probability that he has the disease is 0.00046.

2.2

Given are a set of data points $D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^d$,
The optimal parameter ω for fitting the model $y = \omega^T x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ can be obtained as $\text{argmax}_\omega(P(D/\omega))$

We know that

$$P(D/\omega) = P((x_1, y_1), (x_2, y_2)...(x_n, y_n)/\omega) = \Pi_i P(x_i, y_i/\omega) \tag{2}$$

by assuming all the points originate from an identically independent distribution (iid).
Taking log on both sides,

$$log(P(D/\omega)) = \Sigma_i log(P(x_i, y_i/\omega)) \tag{3}$$

Now substituting $P(x_i, y_i/\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -(\frac{(y_i - \omega^T x_i)^2}{\sigma^2})$ in eq.(3),

$$log(P(D/\omega)) = \Sigma_i (log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(y_i - \omega^T x_i)^2}{\sigma^2}) \tag{4}$$

Since maximizing $P(D/\omega)$ = maximizing $log(P(D/\omega))$ = minimizing $\Sigma_i(y_i - \omega^T x_i)^2$
Therefore, optimal parameter

$$\omega_{MLE} = argmax_\omega(P(D/\omega)) = argmin_\omega \Sigma_i (y_i - \omega^T x_i)^2) \tag{5}$$

The residual sum of squares for the linear model $y = \omega^T x + \epsilon$ given the data points $D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^d$ can be written as

$$RSS = \Sigma_i \epsilon_i^2 \tag{6}$$

where $\epsilon_i = y_i - \omega^T x_i$ Therefore, RSS error in eq(**??**) can be written as

$$RSS = \Sigma_i (_i - \omega^T x_i)^2 \tag{7}$$

Hence the optimal parameter in this case would be,

$$\omega_{RSS} = argmin_\omega \Sigma_i (_i - \omega^T x_i)^2 \tag{8}$$

Comparing eq(5) and eq(8), $\omega_{RSS} = \omega_{MLE}$

3 Models for Heart Disease

3.1

The dataset contains 270 samples with 14 columns of which the first 13 samples (corresponding to Age, Sex, Chest Pain, BP, Cholestoral, fasting blood sugar ¿ 120, resting ECG, max hr, angina, oldpeak, slope, major vessels, defect) represent the features and the final column corresponds to the class (if heart disease is present or not). Of all these 13 features, some are categorical features and some are continuous features. Categorical features are those which have discrete values that cannot be converted into simple mathematical representations like 0,1,2 etc. because they do not have any inherent logical meaning, for example, Sex is a feature which has values, male and female so we cannot assign 0,1 to both of them for performing regression because it would lead to some logical misconceptions like whether to choose 0-male 1-female vs vice versa or whether to choose the difference between the values like choosing only 0,1 or 10, 100. These values affect the linear regression and hence should be dealt properly. Whereas, continuous variables are those which are represented by numbers and hence carry logical meaning by themselves. In this dataset,

Categorical features are Sex, Chest Pain, fasting blood sugar, resting ECG, angina, slope and defect
Continuous features are Age, BP, Cholestoral, max hr, oldpeak and major vessels

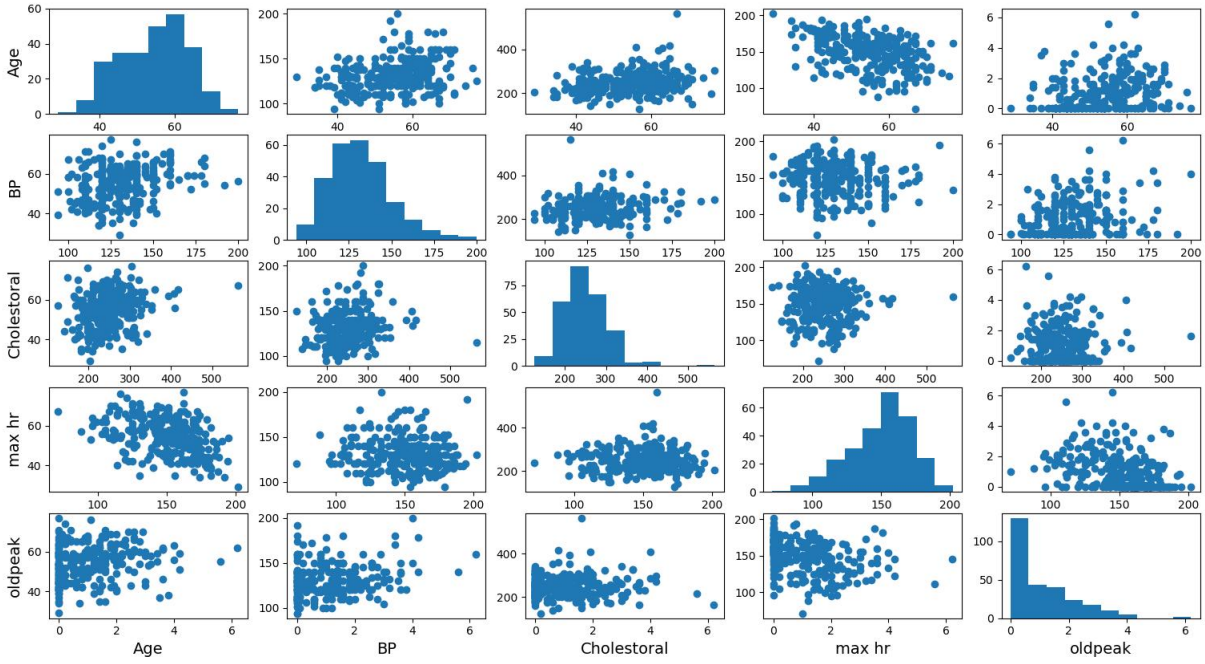

Figure 3: Histograms and scatter plots for continuous features

	heart disease
Age	0.212
Sex	-0.298
Chest Pain	-0.430
BP	0.155
Cholestoral	0.118
fasting blood sugar >120	-0.016
resting ECG	-0.182
max hr	-0.419
angina	0.419
oldpeak	0.418
slope	-0.338
major vessels	0.455
defect	0.405

It can be seen that major vessels is rather a discrete feature but since it holds values rather than textual features, we can consider it as a continuous feature. Table(3.1) represents the correlation between features and heart disease after converting all categorical features into numerical values. It can be observed from the that Chest Pain, max hr, angina, oldpeak, major vessels and slope are correlated more than other features. In general, categorical features are more correlated to heart disease than continuous features. Fig.(3) displays the histograms and scatter plots of continuous features. It can be seen from these images that no features are perfectly correlated and that age and max hr are more correlated than other pairs. Fig.(4) represents the histograms of categorical features.

3.2

Since linear regression and logistic regression cannot be performed over categorical data, all the categorical features are one-hot encoded so that they do not hold any logical meaning but are orthogonal unlike in the case of numerical encoding.

Both these strategies were implemented, where features were analyzed using label encodings as well as one-hot encodings. In the case of label encodings for categorical data, when all the features were used with out dropping any feature, the area under curve measurements were 0.9176 and 0.9076 for linear model and logistic models respectively. The ROC curves in this case are plotted in fig.(5). The optimal threshold in this situation were 0.349 and 0.231 for linear and logistic regression models respectively and the F1 score at these thresholds were 0.892 and 0.895 for linear and logistic models respectively.

Later, using the same labeled encoding for categorical features, feature selection was performed, while selecting features based on the F-statistic measure. It has been observed empirically that by choosing less than 8 features reduces the performance metrics of the linear and logistic models drastically. Hence, 8 best features with the lowest F-statistic have been chosen. The 8 features were (Age, Sex, Chest Pain, BP, Cholestoral, fasting blood sugar ¿ 120, resting ECG, slope). The area under the curve measurements were 0.8640 and

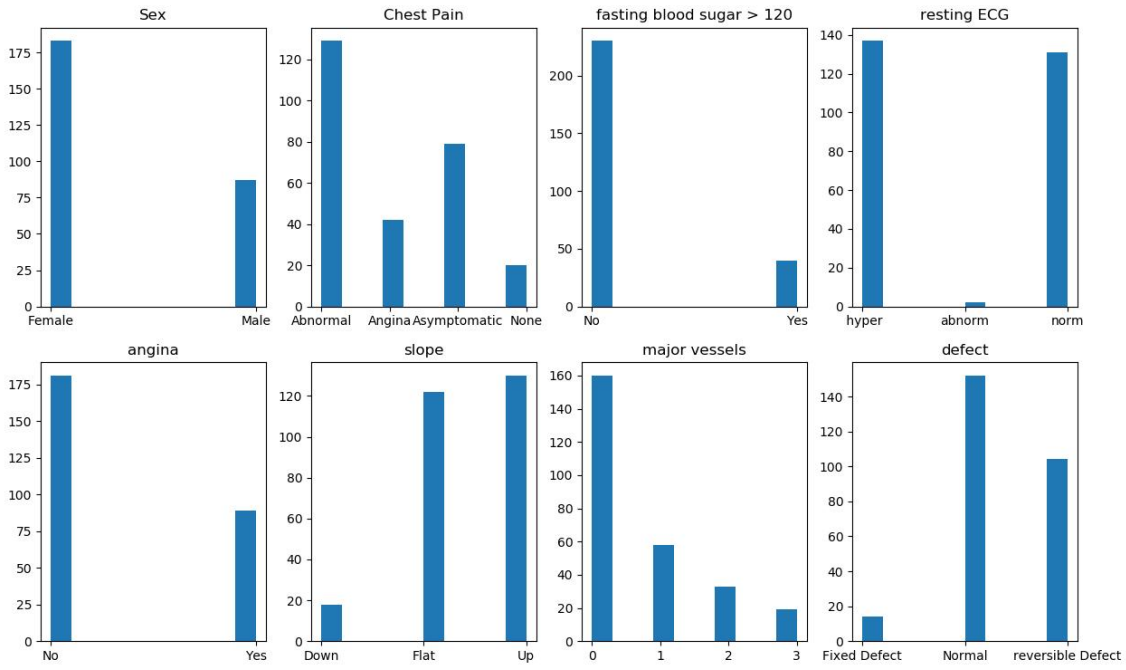

Figure 4: Histogram of categorical features

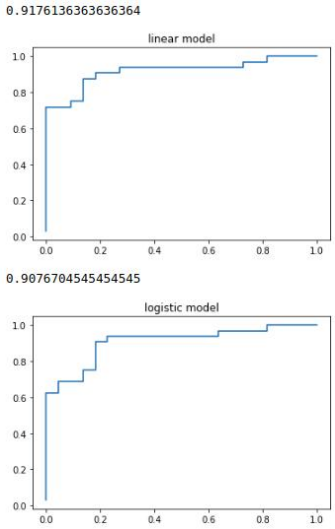

Figure 5: ROC with labeled categorical features and with all features

0.8585 for linear regression and logistic regression respectively. The area under curve of linear regression is higher probably due to the feature selection being performed with linear functions. The ROC curves in this case are plotted in fig.(6). The optimal threshold in this situation were 0.46 and 0.329 for linear and logistic regression models respectively and the F1 score at these thresholds were 0.775 and 0.792 for linear and logistic models respectively.

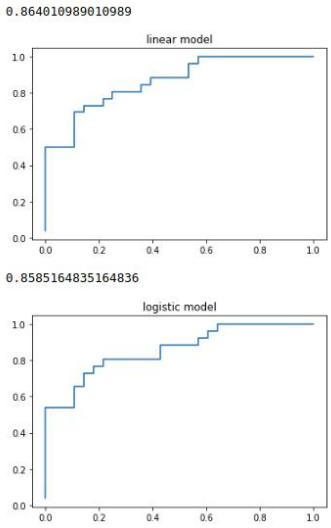

Figure 6: ROC with labeled categorical features and with 7 selected features based on F-statistic

3.3 I

n order to choose features automatically during the computation, features which had F value less than 40 were chosen for fitting the models in order to cover around 8 features from the 13. It has been observed that the features changed with each iteration but only marginally, like an addition of one extra feature or removal of one etc.
The area under the curve values for 5 folds were 0.802, 0.882, 0.920, 0.876,0.911. Therefore the mean and the 95% confidence interval will be 0.878 and 0.041.
The F1 values for the 5 folds and the 95% confidence interval will be 0.801 (mean) and 0.045