MACHINE LEARNING

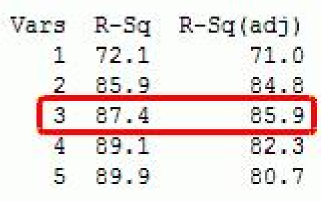
1. C
2. D
3. B
4. B
5. B
6. A & D
7. C
8. D
9. A & B
10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

Answer: Use adjusted R-squared to compare the goodness-of-fit for regression models that contain differing numbers of independent variables.

Let's say you are comparing a model with five independent variables to a model with one variable and the five variable model has a higher R-squared. Is the model with five variables actually a better model, or does it just have more variables? To determine this, just compare the adjusted R-squared values!

The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone. The adjusted R-squared value actually decreases when the term doesn't improve the model fit by a sufficient amount.

The example below shows how the adjusted R-squared increases up to a point and then decreases. On the other hand, R-squared blithely increases with each and every additional independent variable.



```
Vars  R-Sq  R-Sq(adj)
   1  72.1       71.0
   2  85.9       84.8
   3  87.4       85.9
   4  89.1       82.3
   5  89.9       80.7
```

In this example, the researchers might want to include only three independent variables in their regression model. My R-squared blog post shows how an under-specified model (too few terms) can produce biased estimates. However, an overspecified model (too many terms) can reduce the model's precision. In other words, both the coefficient estimates and predicted values can have larger margins of error around them. That's why you don't want to include too many terms in the regression model!

11.Differentiate between Ridge and Lasso Regression.

Answer: Lasso, Ridge and ElasticNet are all part of the Linear Regression family where the x (input) and y (output) are assumed to have a linear relationship. In sklearn, LinearRegression refers to the most ordinary least square linear regression method without regularization (penalty on weights) . The main difference among them is whether the model is penalized for its weights. For the rest of the post, I am going to talk about them in the context of scikit-learn library.

Linear regression (in scikit-learn) is the most basic form, where the model is not penalized for its choice of weights, at all. That means, during the training stage, if the model feels like one particular feature is particularly important, the model may place a large weight to the feature. This sometimes leads to overfitting in small datasets. Hence, following methods are invented.

Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros. During training, the objective function become:

$$\frac{1}{2m}\sum_{i=1}^{m}(y-Xw)^2+alpha\sum_{j=1}^{p}\left|w_j\right|$$

As you see, Lasso introduced a new hyperparameter, *alpha*, the coefficient to penalize weights.

Ridge takes a step further and penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. The objective function becomes:

$$\sum_{i=1}^{n}(y - Xw)^2 + alpha\sum_{j=1}^{p}{w_j}^2$$

12.What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Answer: The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

As a rule of thumb, a VIF of three or below is not a cause for concern. As VIF increases, the less reliable your regression results are going to be.

13.Why do we need to scale the data before feeding it to the train the model?

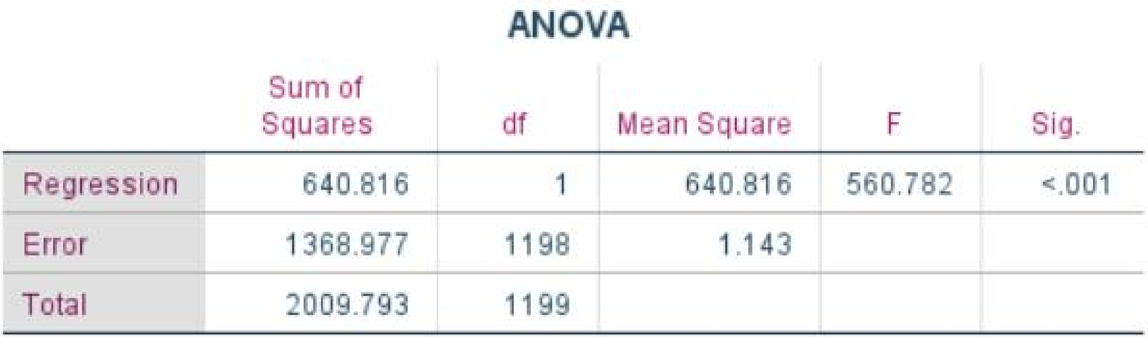
Answer: To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.

Specifically, in the case of Neural Networks Algorithms, feature scaling benefits optimization by:

- It makes the training faster
- It prevents the optimization from getting stuck in local optima
- It gives a better error surface shape
- Weight decay and Bayes optimization can be done more conveniently

14.What are the different metrics which are used to check the goodness of fit in linear regression?

Answer: Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE).



ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	640.816	1	640.816	560.782	<.001
Error	1368.977	1198	1.143		
Total	2009.793	1199			

SST measures how far the data are from the mean, and SSE measures how far the data are from the model’s predicted values. Different combinations of these two values provide different information about how the regression model compares to the mean model.

R-squared

The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model.

R-squared has the useful property that its scale is intuitive. It ranges from zero to one. Zero indicates that the proposed model does not improve prediction over the mean model. One indicates perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

One pitfall of R-squared is that it can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the model’s fit. To remedy this, a related statistic, Adjusted R-squared, incorporates the model’s degrees of freedom.

Adjusted R-squared

Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Adjusted R-squared should always be used with models with more than one predictor variable. It is interpreted as the proportion of total variance that is explained by the model.

There are situations in which a high R-squared is not necessary or relevant. When the interest is in the relationship between variables, not in prediction, the R-squared is less important.

An example is a study on how religiosity affects health outcomes. A good result is a reliable relationship between religiosity and health. No one would expect that religion explains a high percentage of the variation in health, as health is affected by many other factors. Even if the model accounts for other variables known to affect health, such as income and age, an R-squared in the range of 0.10 to 0.15 is reasonable.

The F-test

The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not. An equivalent null hypothesis is that R-squared equals zero.

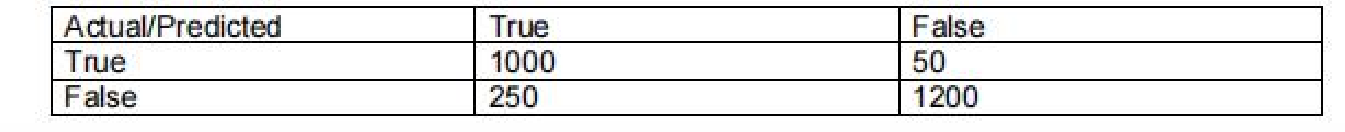
A significant F-test indicates that the observed R-squared is reliable and is not a spurious result of oddities in the data set. Thus the F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable. It can be useful when the research objective is either prediction or explanation.

RMSE

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance. It has the useful property of being in the same units as the response variable.

Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response. It’s the most important criterion for fit if the main purpose of the model is prediction.

15.From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.



Actual/Predicted	True	False
True	1000	50
False	250	1200

Answer: Sensitivity- TP/TP+FN

= 1000/1050 = 0.9

Specificity- TN/TN+FP

= 1200/1450 = 0.8

Precision- TP/Predicted True

= 1000/1050 = 0.9

Recall- TP/ Actual True

= 1000/1050 = 0.9

Accuracy- TP+TN/ TP+TN+FP+FN

= 2200/2500

= 0.88