**MACHINE LEARNING – WORKSHEET 2**

**In Q1 to Q5, only one option is correct, Choose the correct option:**

Qus: 1. In which of the following you can say that the model is overfitting?

Answer: D) None of the above

Qus: 2. Which among the following is a disadvantage of decision trees?

Answer: B) Decision trees are highly prone to overfitting.

Qus: 3. Which of the following is an ensemble technique?

Answer: C) Random Forest

Qus: 4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?

Answer: A) Accuracy

Qus: 5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?

Answer: B) Model B

**In Q6 to Q9, more than one options are correct, Choose all the correct options:**

Qus: 6. Which of the following are the regularization technique in Linear Regression??

Answer: A) Ridge D) Lasso

Qus: 7. Which of the following is not an example of boosting technique?

Answer: C) Random Forest B) Decision Tree

Qus: 8. Which of the techniques are used for regularization of Decision Trees?

Answer: A) Pruning C) Restricting the max depth of the tree

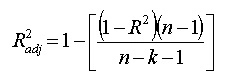
Qus: 9. Which of the following statements is true regarding the Adaboost technique?

Answer: A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

**Q10 to Q15 are subjective answer type questions, Answer them briefly.**

Qus: 10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

Answer: As we Know that, where:

* N is the number of points in your data sample.
* K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

The adjusted R2 will penalize us for adding independent variables (K in the equation) that do not fit the model. Why? In regression analysis, it can be tempting to add more variables to the data as we think of them. Some of those variables will be significant, but we can’t be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.

While values are usually positive, they can be negative as well. This could happen if our R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that our model is a poor fit for our data. Other problems with our model can also cause sub-zero values, such as not putting a constant term in our model.

Qus: 11. Differentiate between Ridge and Lasso Regression.

Answer: A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression.

|  |  |  |
| --- | --- | --- |
|  | Ridge Regression | Lasso Regression |
| Key Difference | It includes all (or none) of the features in the model. Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity. | Along with shrinking coefficients, lasso performs feature selection as well. (Remember the ‘selection‘ in the lasso full-form?) As we observed earlier, some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model. |
| Typical Use Cases | It is majorly used to prevent overfitting. Since it includes all the features, it is not very useful in case of exorbitantly high #features, say in millions, as it will pose computational challenges. | Since it provides sparse solutions, it is generally the model of choice (or some variant of this concept) for modelling cases where the #features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored. |
| Presence of Highly Correlated Features | It generally works well even in presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation. | It arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero. Also, the chosen variable changes randomly with change in model parameters. This generally doesn’t work that well as compared to ridge regression. |

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

Answer: In statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. Cuthbert Daniel claims to have invented the concept behind the variance inflation factor, but did not come up with the name.

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

Answer: It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally.

Formally, If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

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

Answer: These (R Squared, Adjusted R Squared, F Statistics, RMSE / MSE / MAE) are some metrics which you can use to evaluate your regression model.

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

Answer: Given confusion matrix: TP = 1000, FN = 50, FP = 250, TN = 1200

Recall and sensitivity are one and the same.

Sensitivity = TP / (TP+FN) = 1000 / (1000+50) = 0.9523

Specificity = TN / (TN+FP) = 1200 / (1200+250) = 0.8275

Precision = TP / (TP+FP) = 1000 / (1000+250) = 0.8

Accuracy = (TP+TN) / (TP+TN+FP+FN) = (1000+1200) / (1000+1200+250+50) =0.88