WorkSheet-6 Machine Learning

1. In which of the following you can say that the model is overfitting?
   Ans. C) High R-squared value for train-set and Low R-squared value for test-set.
2. Which among the following is a disadvantage of decision trees?
   Ans. B) Decision trees are highly prone to overfitting.
3. Which of the following is an ensemble technique?
   Ans. C) Random Forest
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?
   Ans. B) Sensitivity
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?
   Ans. B) Model B
6. Which of the following are the regularization technique in Linear Regression?
   Ans. A) Ridge & D) LASSO
7. Which of the following is not an example of boosting technique?
   Ans. B) Decision Tree & C) Random Forest
8. Which of the techniques are used for regularization of Decision Trees?
   Ans. A) Pruning & C) Restricting the max depth of the tree
9. Which of the following statements is true regarding the Adaboost technique?
   Ans. 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.
10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?
    Ans. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.
11. Differentiate between Ridge and Lasso Regression.
    Ans. Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 whereas lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients. It's a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares (l2 penalty).

    As we know that ridge regression can't zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficient of collinear variables. Here it helps to select the variables out of given n variables while performing lasso regression.
12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
    Ans. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to

WorkSheet-6 Machine Learning

the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like x and x2.

13. Why do we need to scale the data before feeding it to the train the model?
    Ans. 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 can help the gradient descent converge more quickly towards the minima.
14. What are the different metrics which are used to check the goodness of fit in linear regression?
    Ans. These (R Squared, Adjusted R Squared, F Statistics, RMSE / MSE / MAE) are some metrics which are used to check the goodness of fit in linear regression
15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy. Actual/Predicted True False True 1000 50 False 250 1200.
    Ans. Sensitivity = TPR = TP / (TP + FN) = 0.8000
    Specificity = SPC = TN / (FP + TN) = 0.9600
    Precision = PPV = TP / (TP + FP) = 0.9524
    Recall = TPR = TP / (TP + FN) = 0.8000
    Accuracy = ACC = (TP + TN) / (P + N) = 0.8800

WorkSheet-6 Statistics

1. Which of the following can be considered as random variable?
   Ans. d) All of the mentioned
2. Which of the following random variable that take on only a countable number of possibilities?
   Ans. a) Discrete
3. Which of the following function is associated with a continuous random variable?
   Ans. a) pdf
4. The expected value or _____ of a random variable is the center of its distribution.
   Ans. c) mean
5. Which of the following of a random variable is not a measure of spread?
   Ans. a) Variance
6. The _____ of the Chi-squared distribution is twice the degrees of freedom.
   Ans. a) variance
7. The beta distribution is the default prior for parameters between _____
   Ans. c) 0 and 1
8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
   Ans. b) bootstrap
9. Data that summarize all observations in a category are called _____ data.
   Ans. b) summarized
10. What is the difference between a boxplot and histogram?
    Ans. Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

    Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.
11. How to select metrics?
    Ans. Step 1 Why is the measurement required?
    Step 2 What needs to be measured?
    Step 3 What is the precision of measurement required?
    Step 4 How will it be measured?
    Step 5 What use will the measurement be put to? By whom?
12. How do you assess the statistical significance of an insight?
    Ans. Statistical significance can be accessed using hypothesis testing: – Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B) – Then, we choose a suitable statistical test and statistics used to reject the null hypothesis – Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value) – We calculate the observed test statistics from the data and check whether it lies in the critical region Common tests:
    – One sample Z test – Two-sample Z test – One sample t-test – paired t-test – Two sample pooled equal variances t-test – Two sample unpoled unequal variances t-test and unequal sample sizes (Welch's t-test) – Chi-squared test for variances – Chi-squared test for goodness of fit – ANOVA (for

WorkSheet-6 Statistics

instance: are the two regression models equals? F-test) – Regression F-test (i.e: is at least one of the predictors useful in predicting the response?)

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

    Ans.

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

14. Give an example where the median is a better measure than the mean.

    Ans. When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean. For example, consider the following **distribution of salaries** for residents in a certain city: The median does a better job of capturing the “typical” salary of a resident than the mean.

15. What is the Likelihood?

    Ans. In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values.