WORKSHEET SET 4

Machine Learning

1. D) None of the above
2. B) Decision trees are highly prone to overfitting.
3. D) Decision tree
4. A) Accuracy
5. B) Model B
6. A) Ridge & D) Lasso
7. B) Decision Tree & C) Random Forest
8. A) Pruning & C) Restricting the max depth of the tree
9. 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. If we add unnecessary predictors to a model, adjusted R-squared will decrease. If we add more useful predictors, adjusted r-squared will increase. The adjusted R-squared increases when 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.
11. 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. 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. Lasso regression takes the magnitude of the coefficients, Ridge regression takes the square.
12. VIF or variance inflation factor detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect the regression results. The lower the VIFs value the better. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. VIFs between 1 and 5 suggest that the correlation is not severe enough to warrant corrective measures. VIFs value greater than 5 represents problematic levels of collinearity where the coefficient estimates may not be trusted and the statistical significance is questionable. Typically, in practice, there is a small amount of collinearity among the predictors.
13. To ensure that the model 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. Scaling of the data makes it easy for a model to learn and understand the problem.
14. Goodness of fit means how good a model performs. Different metrics are : Mean Squared Error (MSE), Mean Absolute error, R-squared, Adjusted R-squared and Root mean squared error or RMSE.
15. False positives or FP= 250
    False negatives or FN= 50
    True positives or TP=1000
    True negatives or TN=1200
    Sensitivity: Tp/(Tp+Fn) = 1000/1050 = 0.9523
    Specificity: Tn/(Tn+Fp) = 1200/1450 = 0.8275
    Precision: Tp/(Tp+Fp) = 1000/(1000+250) = 0.8
    Recall: Tp/(Tp+Fn) = 1000/(1000+50) =0.9523
    Accuracy: (Tp+Tn)/(Tp+Tn+Fp+Fn) =(1000+1200)/(1000+1200+250+50) = 0.88

SQL

1. A. Commit, C. Rollback & D. Savepoint
2. A. Create, C. Drop & D. Alter
3. C. SELECT * FROM SALES WHEN PRICE = NULL;
4. C. Authorizing Access and other control over Database
5. B. Column Alias
6. B. COMMIT
7. A. Parenthesis - (...).
8. C. TABLE
9. D. All of the mentioned
10. A. ASC
11. Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database.
12. A database cursor is an identifier associated with a group of rows. It is a pointer to the current row in a buffer.
13. Queries in SQL are Create, Read, Update, or Delete. Select query reads tables and returns rows. Insert query adds a new role to a table. Update query modifies existing rows. Delete query removes a row from a table.
14. In SQL, a constraint is any rule applied to a column or table that limits what data can be entered into it. Any time you attempt to perform an operation that changes that data held in a table such as an Insert, Update or Delete statement the RDBMS will test whether that data violates any existing constraints and, if so, return an error.
15. Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table. Often this is the primary key field that we would like to be created automatically every time a new record is inserted.

Statistics

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. a) variance
6. b) standard deviation
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.
11. 0
12. Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance
13. Examples of data that does not have a Gaussian distribution, nor log-normal: Allocation of wealth among individuals and Values of oil reserves among oil fields (many small ones, a small number of large ones)
14. Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it.
15. Probability is the measure of the likelihood that an event will occur. Likelihood is proportional to probability.