Worksheet Set 6

Statistics Solutions:

1. d) All of the above
2. a) Discrete
3. a) pdf (Probability density function)
4. c) mean
5. c) empirical mean
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal)..A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value. When graphing this five-number summary, only the horizontal axis displays values. Within the quadrant, a vertical line is placed above each of the summary numbers. A box is drawn around the middle three lines (first quartile, median, and third quartile) and two lines are drawn from the box's edges to the two endpoints (minimum and maximum).Although histograms and box plots are collectively part of the chart aid category, they do represent very different types of charts. Both charts effectively represent different data sets; however, in certain situations, one chart may be superior to the other in achieving the goal of identifying variances among data. The type of chart aid chosen depends on the type of data collected, rough analysis of data trends, and project goals.
11. For classification we can use : Precision-Recall, ROC-AUC, Accuracy, Log-loss.
    For regression we can use - MSE, MAE , R square, Adjusted R square.
    For Unsupervised models - Rand Index, Mutual Information
    Others - CV error, Heuristic methods to find k, BLEU (NLP).
12. 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 unpooled 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 instance: are the two regression models equals? F-test)
– Regression F-test (i.e: is at least one of the predictor useful in predicting the response

13. Distribution of numbers that show up on the top of a fair die after a large number of throws.
14. Customer satisfaction survey
15. 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.

Machine Learning Solutions :

1. a) High R-squared value for train-set and High R-squared value for test-set.
2. b) Decision trees are highly prone to overfitting.
3. c) Random Forest
4. a) Accuracy
5. b) Model B
6. a) Ridge and d) Lasso
7. b) Decision Tree and c) Random Forest
8. d) all of the above
9. b) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
10. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase. Adjusted R2 will always be less than or equal to R2.
11. There are a number of reasons to regularize regressions. Typically, the goal is to prevent overfitting, and in that case, L2 (Ridge) has some nice theoretical guarantees built into it. Another purpose for regularization is often interpretability, and in that case, L1(Lasso)-regularization can be quite powerful.
12. 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 the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
13. 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.
14. For linear regression we can use - MAE,MSE,RMSE , R square, Adjusted R square.

15. Accuracy = (TP + TN)/ (TP + FP + TN + FN)
    = (1000 + 1200) / (1000 + 50 + 250 + 1200)
    = 2200 / 2500 = 0.88
    Precision = TP / (TP + FP)
    = 1000 / (1000+50)
    = 0.95
    Recall = TP / (TP + FN)
    = 1000 / (1000 + 250)
    = 0.8
    Specificity = FP / (TN + FP)
    = 50 / (1200 + 250)
    = 0.034

SQL Worksheet Solutions :

1. a) Commit, c) Rollback and d) Savepoint
2. a) create, c) drop and d) alter
3. b) Select name from sales;
4. c) Authorizing Access and other control over Database
5. d) all of the mentioned
6. b) commit
7. a) Parentheses
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. Note that denormalization does not mean not doing normalization. It is an optimization technique that is applied after doing normalization.
12. In computer science, a database cursor is a mechanism that enables traversal over the records in a database. Cursors facilitate subsequent processing in conjunction with the traversal, such as retrieval, addition and removal of database records. The database cursor characteristic of traversal makes cursors akin to the programming language concept of iterator.
13. It is commonly accepted that there are three different types of search queries:
    - Navigational search queries.
    - Informational search queries.
    - Transactional search queries.
14. SQL constraints are used to specify rules for the data in a table.
    Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation between the constraint and the data action, the action is aborted.
    Constraints can be column level or table level. Column level constraints apply to a column, and table level constraints apply to the whole table.
    The following constraints are commonly used in SQL:
    - NOT NULL - Ensures that a column cannot have a NULL value

- UNIQUE - Ensures that all values in a column are different
- PRIMARY KEY - A combination of a NOT NULL and UNIQUE. Uniquely identifies each row in a table
- FOREIGN KEY - Prevents actions that would destroy links between tables
- CHECK - Ensures that the values in a column satisfies a specific condition
- DEFAULT - Sets a default value for a column if no value is specified
- CREATE INDEX - Used to create and retrieve data from the database very quickly

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.