MACHINE LEARNING

ASSIGNMENT - 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Answer: The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation. RSS gives the proportion of variation in target variable. So it would be good to take RSS than R-squared.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

Answer: Total variation in target variable is the sum of squares of the difference between the actual values and their mean.

$$TSS = \sum (y_i - \vec{y})^2$$

TSS or Total sum of squares gives the total variation in Y. We can see that it is very similar to the variance of Y. While the variance is the average of the squared sums of difference between actual values and data points, TSS is the total of the squared sums.

The explained sum of squares measures how much variation there is in the modelled values and this is compared to the total sum of squares (TSS), which measures how much variation there is in the observed data, and to the residual sum of squares, which measures the variation in the error between the observed data and modelled values.

Residual for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

$$Residual = actual - predicted = y\_\hat{y}$$

Using the residual values, we can determine the sum of squares of the residuals also known as **Residual sum of squares** or RSS.

$$RSS = \sum (y_i - \hat{y}_i)^2$$

3. What is the need of regularization in machine learning?

Answer: Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

4. What is Gini–impurity index?

Answer: Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

6. What is an ensemble technique in machine learning?

Answer: Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

7. What is the difference between Bagging and Boosting techniques?

Answer: Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification. It attempts to increase the weight of an observation if it was erroneously categorized. Boosting creates good predictive models in general.

8. What is out-of-bag error in random forests?

Answer: The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained

9. What is K-fold cross-validation?

Answer: K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation. Each fold is used as a testing set at one point in the process.

10. What is hyper parameter tuning in machine learning and why it is done?

Answer: Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors

11. What issues can occur if we have a large learning rate in Gradient Descent?

Answer: Gradient Descent is too sensitive to the learning rate. If it is too big, the algorithm may bypass the local minimum and overshoot. If it too small, it might increase the total computation time to a very large extent.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Answer: Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries.

13. Differentiate between Adaboost and Gradient Boosting. Adaboost Gradient Boost

Answer: Some of the differences between adaboost and Gradient Boosting are

1 In Adaboost An additive model where shortcomings of previous models are identified by high-weight data points. But in gradient boost an additive model where shortcomings of previous models are identified by the gradient.

2 In Adaboost the trees are usually grown as decision stumps. But in gradient boost the trees are grown to a greater depth usually ranging from 8 to 32 terminal nodes.

3 In Adaboost each classifier has different weights assigned to the final prediction based on its performance. But in gradient boost all classifiers are weighed equally and their predictive capacity is restricted with learning rate to increase accuracy.

4 In Adaboost it gives weights to both classifiers and observations thus capturing maximum variance within data. But in gradient boost It builds trees on previous classifier's residuals thus capturing variance in data.

14. What is bias-variance trade off in machine learning?

Answer: When a model is high on variance, it is then said to as Overfitting of Data. Basically it's for testing data. By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as Underfitting of Data. This happens when the hypothesis is too simple or linear in nature. To avoid Overfitting and Underfitting of Data bias-variance tradeoff have to be achieved.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answer: The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyperplane decision boundary between the classes. The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with view to making it linear dataset.

WORKSHEET 5 SQL

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using MySQL for the required Operation.

Table Explanations:

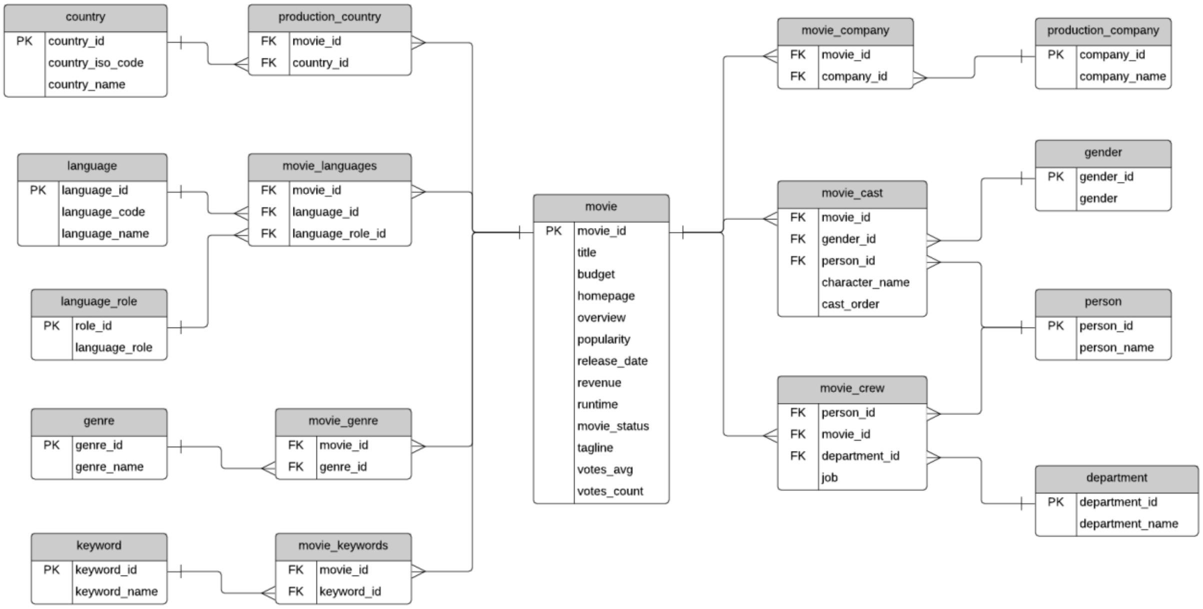

Table Explanations:

- The **movie** table contains information about each movie. There are text descriptions such as title and overview. Some fields are more obvious than others: revenue (the amount of money the movie made), budget (the amount spent on creating the movie). Other fields are calculated based on data used to create the data source: popularity, votes_avg, and votes_count. The status indicates if the movie is Released, Rumoured, or in Post-Production.
- The **country** list contains a list of different countries, and the **movie_country** table contains a record of which countries a movie was filmed in (because some movies are filmed in multiple countries). This is a standard many-to-many table, and you'll find these in a lot of databases.
- The same concept applies to the **production_company** table. There is a list of production companies and a many-to-many relationship with movies which is captured in the **movie_company** table.
- The **languages** table has a list of languages, and the **movie_languages** captures a list of languages in a movie. The difference with this structure is the addition of a **language_role** table.
- This **language_role** table contains two records: Original and Spoken. A movie can have an original language (e.g. English), but many Spoken languages. This is captured in the **movie_languages** table along with a role.
- **Genres** define which category a movie fits into, such as Comedy or Horror. A movie can have multiple genres, which is why the **movie_genres** table exists.
- The same concept applies to **keywords**, but there are a lot more keywords than genres. I'm not sure what qualifies as a keyword, but you can explore the data and take a look. Some examples as "paris", "gunslinger", or "saving the world".

 The cast and crew section of the database is a little more complicated. Actors, actresses, and crew members are all people, playing different roles in a movie. Rather than have separate lists of names for crew and cast, this database contains a table called **person**, which has each person's name.

 The **movie_cast** table contains records of each person in a movie as a cast member. It has their character name, along with the **cast_order**, which I believe indicates that lower numbers appear higher on the cast list.

 The **movie_cast** table also links to the gender table, to indicate the gender of each character. The gender is linked to the **movie_cast** table rather than the **person** table to cater for characters which may be a different gender than the person, or characters of unknown gender. This means that there is no gender table linked to the **person** table, but that's because of the sample data.

 The **movie_crew** table follows a similar concept and stores all crew members for all movies. Each crew member has a job, which is part of a **department** (e.g. Camera).

QUESTIONS:

1. Write SQL query to show all the data in the Movie table.

Answer: SELECT * FROM Movie

2. Write SQL query to show the title of the longest runtime movie.

Answer: SELECT MAX(runtime) FROM Movie

3. Write SQL query to show the highest revenue generating movie title.

Answer: SELECT MAX(revenue) FROM Movie

4. Write SQL query to show the movie title with maximum value of revenue/budget.

Answer: SELECT title FROM Movie WHERE revenue == MAX(revenue)

5. Write a SQL query to show the movie title and its cast details like name of the person, gender, character name, cast order.

Answer: SELECT name of the title, person_name, gender_id, character_name, cast_order
FROM Movie as m
LEFT JOIN movie cast AS c1.
ON m. movie _id=c1. movie_id.
LEFT JOIN person AS c2.
ON m. person_id = c2. person_id.

6. Write a SQL query to show the country name where maximum number of movies has been produced, along with the number of movies produced.

Answer: SELECT COUNT(country_name) FROM Movie as m
LEFT JOIN production_country AS c1.
ON m. movie _id=c1. movie_id.
LEFT JOIN country AS c2.
ON m.country_id = c2.country_id.
ORDER BY COUNT(country_name) DESC;

7. Write a SQL query to show all the genre_id in one column and genre_name in second column.

Answer: SELECT genre_id, genre_name FROM Movie as m
FULL OUTER JOIN movie_gene AS c1.

ON m. movie _id=c1. movie_id.
FULL OUTER JOIN genre AS c2.
ON m. genre _id = c2. genre_id.

8. Write a SQL query to show name of all the languages in one column and number of movies in that particular column in another column.

Answer: SELECT language_ name, COUNT (titles) FROM Movie as m
RIGHT JOIN movie_ languages AS c1.
ON m. movie _id=c1. movie_id.
RIGHT JOIN language AS c2.
ON m.language_id = c2.language _id.

9. Write a SQL query to show movie name in first column, no. of crew members in second column and number of cast members in third column.

Answer: SELECT title, COUNT(person_id), COUNT(person_id) FROM Movie as m
LEFT JOIN movie_ cast AS c1.
ON m. movie _id=c1. movie_id.
LEFT JOIN person AS c2.
ON m.person_id = c2.person_id.

10. Write a SQL query to list top 10 movies title according to popularity column in decreasing order.

Answer: SELECT title FROM Movie ORDER BY popularity DESC LIMIT10

11. Write a SQL query to show the name of the 3rd most revenue generating movie and its revenue.

Answer: SELECT title FROM Movie ORDER BY revenue DESC LIMIT 3,1;

12. Write a SQL query to show the names of all the movies which have "rumoured" movie status.

Answer: SELECT title FROM Movie WHERE movie_status=="rumoured"

13. Write a SQL query to show the name of the "United States of America" produced movie which generated maximum revenue.

Answer: SELECT title FROM Movie as m
LEFT JOIN production_country AS c1.
ON m. movie _id=c1. movie_id.
LEFT JOIN country AS c2.
ON m.country _id = c2.country _id.
WHERE revenue==MAX(revenue)

14. Write a SQL query to print the movie_id in one column and name of the production company in the second column for all the movies.

Answer: SELECT movie_id,company_name FROM Movie as m
RIGHT JOIN movie_company AS c1.
ON m. movie _id=c1. movie_id.
RIGHT JOIN production_company AS c2.
ON m. company _id = c2. company _id.

15. Write a SQL query to show the title of top 20 movies arranged in decreasing order of their budget

Answer: SELECT title FROM Movie ORDER BY budget DESC LIMIT20;

STATISTICS WORKSHEET-5

Q1 to Q10 are MCQs with only one correct answer. Choose the correct option.

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.

a) Mean

b) Actual

c) Predicted

d) Expected

Answer: (c)

2. Chisquare is used to analyse

a) Score

b) Rank

c) Frequencies

d) All of these

Answer: (c)

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

a) 4

b) 12

c) 6

d) 8

Answer: (c)

4. Which of these distributions is used for a goodness of fit testing?

a) Normal distribution

b) Chisqared distribution

c) Gamma distribution

d) Poission distribution

Answer: (b)

5. Which of the following distributions is Continuous

a) Binomial Distribution

b) Hypergeometric Distribution

c) F Distribution

d) Poisson Distribution

Answer: (d)

6. A statement made about a population for testing purpose is called?

a) Statistic

b) Hypothesis

c) Level of Significance

d) TestStatistic

Answer: (b)

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

a) Null Hypothesis

b) Statistical Hypothesis

c) Simple Hypothesis

d) Composite Hypothesis

Answer: (a)

8. If the Critical region is evenly distributed then the test is referred as?

a) Two tailed

b) One tailed

c) Three tailed

d) Zero tailed

Answer: (a)

9. Alternative Hypothesis is also called as?

a) Composite hypothesis

b) Research Hypothesis

c) Simple Hypothesis

d) Null Hypothesis

Answer: (b)

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

a) np

b) n

Answer: (a)