FLIP ROBO

WORKSHEET- 5 MACHINE LEARNING

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 - R2 it represents the proportion of the variance in our data which is explained by our model; the closer to one, the better the fit.

   The residual sum of squares (RSS) is the sum of the squared distances between actual versus predicted values:

   $$RSS=\Sigma i=1 \text{ to } n\ [(yi - y'i)^2]$$

   Where yi is a given data point and y'i is the fitted value for yi.

   The actual number we get depends largely on the scale of our response variable. Taken alone, the RSS isn't so informative.

   Therefore, R2 is a better measure.

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 - The residual sum of squares (RSS) is the sum of the squared distances between actual versus predicted values:

   $$RSS=\Sigma i=1 \text{ to } n\ [(yi - y'i)^2]$$

ESS: The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

$$ESS=\Sigma i=1 \text{ to } n [(y'i - ymean)^2]$$

TSS: Total sum of squares (TSS) = explained sum of squares (ESS)+ residual sum of squares (RSS).

$$TSS = \Sigma i=1 \text{ to } n [(y'i - ymean)^2] + \Sigma i=1 \text{ to } n [(yi - y'i)^2]$$

The relation between the above 3 could be linearly expressed as:

$$TSS = RSS + ESS$$

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

Answer - Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. Regularization is a penalty faced by in case of regressions. Regularization constraints or shrinks the coefficient towards zero. This means that this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. Regularization significantly reduces the variance of the model, without substantial increase in its bias.

4. What is Gini–impurity index?

Answer – Gini index or Gini impurity measures the probability of a particular variable to be wrongly classified when chosen randomly. This measure is calculated where the modelling contains Tree Algorithms like Decision Tress or random forest.

If we have C total classes and p(i) is the probability of picking a data point with class i, then the Gini Impurity is calculated as

$$G=i=1\Sigma C\ p(i)*(1-p(i))$$

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class, then it can be called pure.

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

Answer - Yes, unregularized decision trees are prone to overfitting. Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.

But unlike other algorithms decision tree does not use regularization to fight against overfitting. Instead it uses pruning. There are mainly to types of pruning performed: -

- **Pre-pruning** that stop growing the tree earlier, before it perfectly classifies the training set.
- **Post-pruning** that allows the tree to perfectly classify the training set, and then post prune the tree.

6. **What is an ensemble technique in machine learning?**

Answer - Ensemble techniques combine the decisions from multiple models to improve the overall performance. Bagging and Boosting are two of the most used techniques in machine learning.

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

Answer -

- Bagging is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.
- Boosting is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

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

Answer - Out-of-bag error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating.

9. What is K-fold cross-validation?

Answer - In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 sub samples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation.

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

Answer - While defining the parameters, often the default values are not the ones that give the best result. In machine learning, hyper parameter optimization or tuning is the problem of choosing a set of optimal hyper parameters for a learning algorithm. A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

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

Answer - A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution.

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

Answer - No, logistic regression only forms 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.**

Answer - Gradient Boosting is a generic algorithm to find approximate solutions to the additive modelling problem, while AdaBoost be a special case with a particular loss function. Hence, Gradient Boosting is much more flexible.

On the other hand, AdaBoost can be interpreted from a much more intuitive perspective and can be implemented without the reference to gradients by reweighting the training samples based on classifications from previous learners.

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

Answer - If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and under fitting the data.

This trade off in complexity is why there is a trade off between bias and variance. An algorithm can't be more complex and less complex at the same time.

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

Answer -

- **Linear Kernel** is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are many Features in a particular Data Set.
- **Gaussian RBF(Radial Basis Function)** is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point. Gaussian Kernel is of the following format
- In the **polynomial kernel**, we simply calculate the dot product by increasing the power of the kernel.

FLIP ROBO

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.

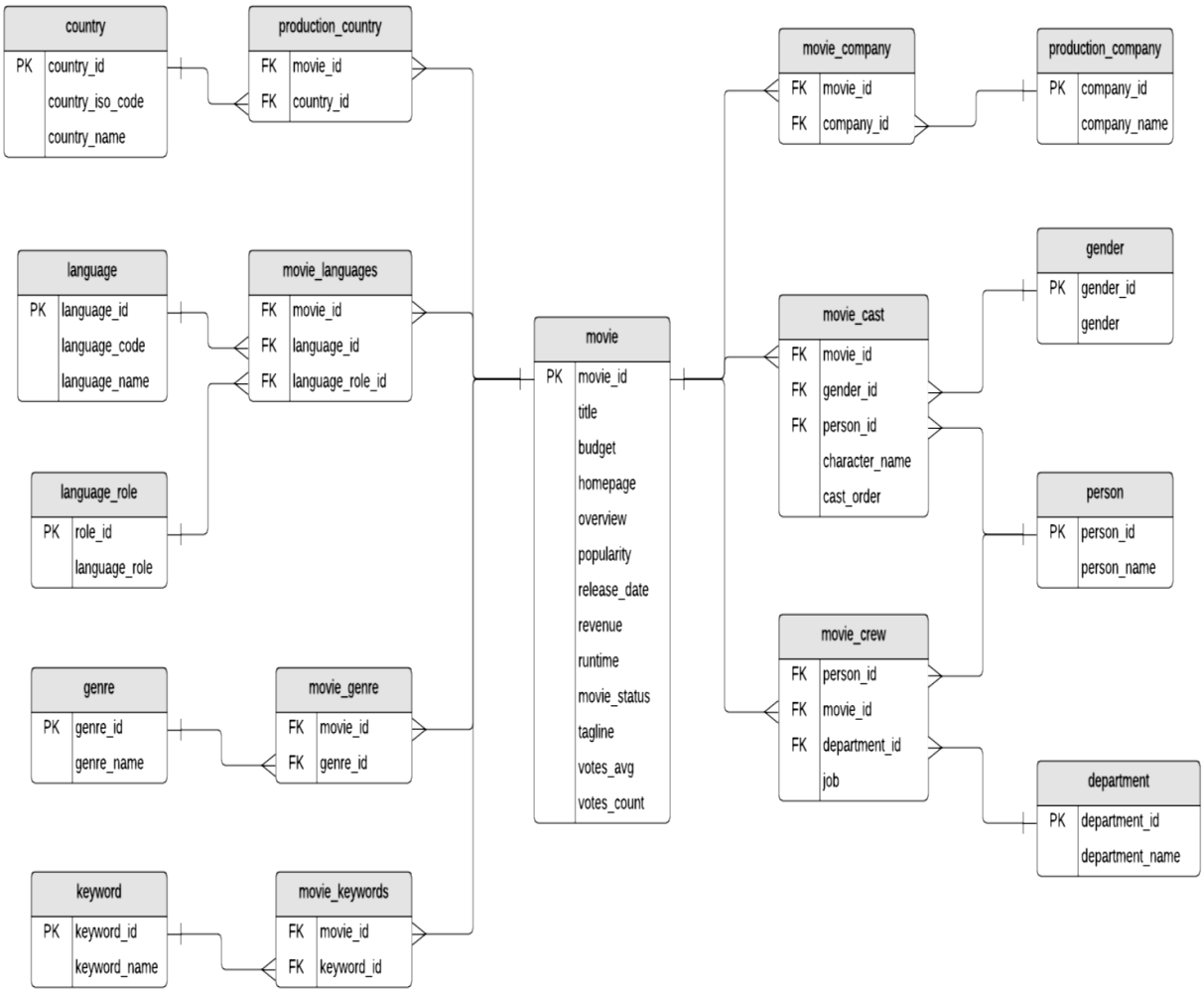

Table Explanations:

- 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.**

Select * from movie;

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

Select title from movie order by runtime desc limit 1;

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

Select title from movie order by revenue desc limit 1;

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

Select title from movie order by budget desc limit 1;

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

Select title, gender, character_name, cast_order, person_name from movie a inner join movie_cast b on a.movie_id=b.movie_id inner join gender c on c.gender_id=b.gender_id inner join person d on d.person_id= b.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.**

Select country_name, count(country_name) as count from country as a inner join production_country as b on b.country_id=a.country_id group by country_name order by count desc limit 1;

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

Select * from genre;

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.**

Select language_name,movie_id,count(language_name) from movie_languages as a join language as b on a.language_id=b.language_id group by language_name order by count(language_name) desc;

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.**

Select m.title as movie_name, count(cr.person_id) as no_of_crews, count(ca.person_id) as no_of_cast from movie as m inner join movie_crew as cr on cr.movie_id=m.movie_id inner join movie_cast ca on ca.person_id=cr_person_id;

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

Select title from movie order by popularity desc limit 10;

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

Select title from movie order by revenue desc offset 3 limit 1;

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

Select title from movie where movie_status like 'rumored';

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

Select title, revenue from movie a inner join production_country b on b.movie_id = a.movie_id inner join country c on c.country_id = b. country_id where country_name= 'United State of America';

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.**

Select m.movie_id, pc.company_name from movie m inner join movie_company mc on mc.movie_id = m.movie_id inner join production_company pc on pc.company_id =mc.company_id;

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

Select title from movie order by budget desc limit 20;

FLIP ROBO

WORKSHEET - 5 STATISTICS

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 – d) Expected

2. Chisquare is used to analyse

a) Score
b) Rank
c) Frequencies
d) All of these

Answer – c) Frequencies

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) 6

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) Chisqared distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution
b) Hypergeometric Distribution
c) F Distribution
d) Poisson Distribution

Answer – c) F Distribution

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

a) Statistic
b) Hypothesis
c) Level of Significance
d) TestStatistic

Answer – b) Hypothesis

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) Null Hypothesis

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) Two Tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis
b) Research Hypothesis
c) Simple Hypothesis
d) Null Hypothesis

Answer – b) Research Hypothesis

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) np