SESSION - (2024-25)

DATA MINING AND PREDICTIVE MODELLING

LAB MANUAL

DEPARTMENT OF COMPUTER ENGINEERING AND APPLICATIONS

(CSE-AIML)

GLA. University

Submitted By-

Prince Bazad

Submitted To-

Mr.Brajesh Kumar Shukla

LAB-1

1. How do you deal with missing data? Explain all the possible situations.
2. Explain about Label and One Hot Encoding with an example.
3. What are the different types of normalizations used often? Give the formula for normalizing the values.
4. “Data cleaning, scrubbing, and normalizing became over 70% of a data scientist’s job.” Explain Why?

Ans 1) Dealing with Missing Data

- **Remove Data**: Drop rows or columns with too many missing values if they are not critical.
- **Impute Values**: Fill missing data using:
  - **Mean/Median/Mode** (for numerical or categorical data).
  - **Forward/Backward Fill** (time-series data).
  - **Prediction** using ML models.
- **Flag Missing**: Create an additional feature indicating missingness for analysis.

Ans 2) Label Encoding vs. One-Hot Encoding

- **Label Encoding**: Assigns numerical labels to categories (e.g., `Red → 0`, `Blue → 1`).
  - *Example*: `['Red', 'Blue', 'Red'] → [0, 1, 0]`
- **One-Hot Encoding**: Creates binary columns for each category.
  - *Example*: `['Red', 'Blue', 'Red'] → [[1, 0], [0, 1], [1, 0]]`
- Use **Label Encoding** for ordinal data and **One-Hot Encoding** for nominal data.

Ans 3) Types of Normalizations

1. **Min-Max Normalization**: Rescales data to [0, 1].
   - Formula: x′=x−min(x)max(x)−min(x)x' = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}x′=max(x)−min(x)x−min(x)

2. **Z-Score Normalization (Standardization)**: Scales data with mean 0 and variance 1.
    - Formula: z=x−μσz = \frac{x - \mu}{\sigma}z=σx−μ
3. **Decimal Scaling**: Scales by a power of 10.
    - Formula: x′=x10jx' = \frac{x}{10^j}x′=10jx , where jjj depends on the max value.

Ans 4). Why Does Data Cleaning Dominate?

- Real-world data is messy, containing errors, duplicates, outliers, or missing values.
- Clean data ensures accurate models and insights.
- Normalizing and structuring data reduces biases, improves performance, and prepares data for ML algorithms.

IN-LAB:

You are very much interested in Data Science, and you thought of doing a project on Data Mining. So, you want to predict the IMDB rating of a movie based on various features. But before that you realized that the

Data is not clean. So, you need to perform following operations:

a. Impute the columns containing more than 100 NaN values with suitable central tendency measure.

b. Remove the rows with NaN values in remaining columns.

c. Label Encode the columns ‘language’, ‘country’, ‘content rating’. d. OneHotEncode the column ‘country’

Colab Link-
https://colab.research.google.com/drive/1hG8Ff4n_kxmV1zRCfzE7S0mfS8-xFize?usp=sharing

LAB-2

1.How correlation analysis is useful in data reduction?

2.How does PCA impact data mining activity?

3.Differentiate b/w Pearson Correlation and Spearman’s correlation.

4.What are the limitations of PCA? best short answer

Ans 1) **Correlation Analysis in Data Reduction**

- Identifies highly correlated features to remove redundancy.

- Reduces the number of features without significant loss of information, simplifying data for analysis.

Ans 2) Impact of PCA on Data Mining

- **Dimensionality Reduction**: Simplifies datasets while retaining maximum variance.
- **Noise Reduction**: Filters out irrelevant variations.
- **Improved Performance**: Speeds up algorithms and reduces overfitting.

Ans 3) Pearson vs. Spearman Correlation

- **Pearson Correlation**: Measures linear relationships. Assumes normality. ○ Formula: r=Cov(X,Y)σXσYr = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}r=σX σY Cov(X,Y)
- **Spearman Correlation**: Measures monotonic relationships using ranks. Non-parametric.
    - Formula: rs=1−6∑di2n(n2−1)r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 1)}rs =1−n(n2−1)6∑di2 , where did_idi = rank differences.

Ans 4) Limitations of PCA

- **Interpretability**: Principal components lack clear physical meaning.
- **Loss of Information**: May discard useful variance.
- **Sensitivity**: Affected by outliers and scaling.
- **Linear Assumption**: Ineffective for non-linear relationships.

IN-LAB:

You are thinking to sell your house and you are also aware of Data Science. So, you want to predict the price of your house based on various features. You are using the House Prices dataset. After doing hours of modelling. You realized that the model performance is low due to high number of features. Now, you want to eliminate the features using Pearson Correlation as follows:

1. Load & describe the Data.

2. Build correlation matrix.

3. Plot a Heatmap of correlation matrix.

Keep only the features which are having correlation with price > abs (0.5)

Dataset- train.csv

https://colab.research.google.com/drive/1Qf0rN-JpUkctbjLq6cacRe5CJjcfTJEu?usp=sharing

LAB-3

1. What is regression? What are the different types of regression?

2. Explain the 3 different evaluation approaches i.e., train test on same data, splitting train and test & K-Fold Cross Validation.

3. What are evaluation metrics used in Regression models?

4. What is Under Fitting and Over Fitting?

Ans 1) Regression predicts a continuous target variable using one or more predictors.

- **Types**:
  - **Linear Regression**: Predicts a linear relationship.
  - **Multiple Linear Regression**: Involves multiple predictors.
  - **Polynomial Regression**: Handles non-linear relationships.
  - **Logistic Regression**: For binary classification tasks.

Ans 2) Evaluation Approaches

- **Train-Test on Same Data**: Tests on training data, often overestimates performance.
- **Train-Test Split**: Divides data into separate train and test sets for unbiased evaluation.
- **K-Fold Cross Validation**: Splits data into kkk subsets, ensuring every subset is tested once, reducing bias and variance.

Ans 3) Evaluation Metrics

- **MAE**: Average absolute error.
- **MSE**: Average squared error.
- **RMSE**: Square root of MSE for interpretability.
- **R-Squared**: Proportion of variance explained by the model.

Ans 4) Underfitting vs. Overfitting

- **Underfitting**: Model too simple; high bias, low accuracy on train/test.
- **Overfitting**: Model too complex; excellent on train data, poor on test data.

IN-LAB:

You are a weather data analyst. While doing your daily job you have encountered a linear relationship between salinity and temperature. So, you want to predict temperature based on salinity. So, follow the below steps:

Data: bottle.csv

a. Perform essential Data cleaning.

b. Plot the correlation between salinity and temperature.

c. Split the data set into Test and Train data set.

d. Build a Linear Regression Model

e. Display the coefficient and intercept of the Regression model.

f. Evaluate the Regression model using MSE and R2(R-Squared) value.

ColabLink

https://colab.research.google.com/drive/1g_91yD0uxnuU4kvs5U_8AeVEdlL1pN0W?usp=sharing

LAB-4

1.How does Decision Tree Algorithm work?
2.What are various classification Metrics? Explain them Briefly.
3.Write down the process for implementing any classification algorithm

Ans1)

Splits data into subsets based on feature values using a criterion (e.g., Gini Index, Entropy).

Recursively creates branches until nodes are pure (one class) or meet a stopping condition.

Predictions are made by traversing the tree based on feature values.

Ans2) **Accuracy**: Proportion of correctly predicted instances.

- Accuracy=TP+TNTotal\text{Accuracy} = \frac{\text{TP} +

\text{TN}}{\text{Total}}Accuracy=TotalTP+TN

Precision: Proportion of true positives among predicted positives.

- Precision=TPTP+FP\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}Precision=TP+FPTP

Recall (Sensitivity): Proportion of true positives among actual positives.

- Recall=TPTP+FN\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}Recall=TP+FNTP

F1-Score: Harmonic mean of precision and recall.

- F1=2×Precision×RecallPrecision+Recall\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F1=2×Precision+RecallPrecision×Recall

ROC-AUC: Measures model's ability to distinguish classes.

Confusion Matrix: Summarizes prediction outcomes (TP, TN, FP, FN).

Ans 3) Understand the Problem: Define objectives and gather data.

Preprocess Data:

- Clean data (handle missing values, outliers).
- Encode categorical features.
- Scale/normalize numerical features if required.

Split Data: Into training and testing sets.

Select Algorithm: Choose an appropriate classifier (e.g., Logistic Regression, SVM, Decision Tree).

Train Model: Fit the algorithm to the training data.

Evaluate Model:

- Use metrics like accuracy, precision, recall, etc. ● Validate using cross-validation if needed.

Optimize: Tune hyperparameters to improve performance.

Deploy: Use the model for predictions.

IN-LAB:

A multinational company wants to promote their products in a metropolitan city. They do not want to call everyone and promote. So, they took the data of the people containing theattributes Gender, Age, Estimated Salary, and they allocated User-Id to each one. They want to classify which customers want to purchase their product. Help them to make predictions with decision tree and check the accuracy of the tree.

Finally, you will be given a dataset with 5 attributes: - User-Id, Gender, Age, Estimated Salary, Purchased You can follow the following steps: -

1. **Load the dataset.**
2. **Find out dependent and independent variables.**
3. **Split dataset into train and test**
4. **Fit the model in the Decision Tree classifier.**
5. **Make predictions and check accuracy.**
6. **Visualization of the tree using python-graphviz and pydotplus**

https://colab.research.google.com/drive/1-NhKcTPAGB2XYJTFFXl3E6Fi6-4BXfnD?usp=sharing

LAB-5

1.What mathematical concept Naive Bayes is based on?
2.What are basic assumption in the case of the Naive Bayes classifier?
3.How Naïve Bayes Classifier works?
4.What are advantages and disadvantages of Naive Bayes classifier?

Ans 1) Naive Bayes is based on Bayes' Theorem:

P(A∣B)=P(B∣A)⋅P(A)P(B)P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}P(A∣B)=P(B)P(B∣A)⋅P(A)

Ans 2) It assumes conditional independence of features.

- **Features are conditionally independent given the class label. ● Each feature contributes equally to the outcome.**

Ans 3) Calculates prior probabilities for each class.

Computes likelihood
(P(Feature∣Class)P(\text{Feature}|\text{Class})P(Feature∣Class)) for each feature given the class.

Uses Bayes' Theorem to compute posterior probability for each class:
P(Class∣Features)∝P(Class)∏P(Feature∣Class)P(\text{Class}|\text{Features}) \propto P(\text{Class}) \prod
P(\text{Feature}|\text{Class})P(Class∣Features)∝P(Class)∏P(Feature∣Class)

Predicts the class with the highest posterior probability.

Ans 4) Advantages:

- Fast and Efficient: Simple to implement with low computational cost.
- Works Well with Small Data: Effective even with limited data.
- Handles Categorical Data: Works well with discrete data.

Disadvantages:

- Strong Independence Assumption: Rarely holds true in real-world data.
- Not Suitable for Continuous Features Without Transformation: Requires techniques like Gaussian assumption for numerical data.
- Poor Performance with Highly Correlated Features: Assumption of independence breaks down.

LAB WORK-

You are working as a data scientist for a company. They have asked you to predict and classify with naive bayes classifier using the salary range of the adults in the country using the census income dataset given below:

You can follow the following steps:

1. Load and Explore the Dataset.
2. Perform required pre-processing.
3. Split the dataset into training and testing sets.
4. Train the Naïve Bayes Classifier
5. Check the accuracy score and construct confusion matrix.

6. Calculate various classification metrics. **Dataset**:

https://www.kaggle.com/qizarafzaal/adult-dataset

https://colab.research.google.com/drive/1cLOwUtQGlMFqe68Srwe8JOetYayfEZcT?usp=sharing

LAB-6

1.Explain the K-Means algorithm briefly?
2. What are the hyper parameters used in K-means algorithm?
3 Explain the K-medoids Algorithm briefly.
4. What are all the differences between K-Means and K-Medoids algorithm?

Ans 1) K-Means clusters data by minimizing the sum of squared distances between data points and their cluster centroids.

Steps:

1. Initialize kkk cluster centroids randomly.
2. Assign each point to the nearest centroid.
3. Recalculate centroids as the mean of assigned points.
4. Repeat until centroids stabilize or a stopping criterion is met.

Ans 2) Number of Clusters (kkk): Determines the number of groups.

Max Iterations: Limits the number of iterations.

Initialization Method: Strategy to initialize centroids (e.g., random, K-Means++).

Distance Metric: Usually Euclidean distance.

Ans 3) A clustering algorithm similar to K-Means but uses medoids (actual data points) as cluster centers.

Steps:

1. Initialize kkk medoids randomly.
2. Assign points to the nearest medoid.

3. Update medoids by minimizing total intra-cluster distance.
4. Repeat until medoids stabilize or a stopping criterion is met**.**

Ans 4)

Aspect	K-Means	K-Medoids
Centroids	**Mean of cluster points.**	**Actual data points (medoids).**
Distance Metric	**Uses squared Euclidean distance.**	**Supports general distance metrics.**
Robustness	**Sensitive to outliers.**	**Robust to outliers.**
Computation Cost	**Faster, lower cost.**	**Slower due to medoid computation.**

IN-LAB:

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categorize the countries using some socio-economic and health factors.

that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

You are given a country-data.csv dataset, it has columns like country, child mortality, exports, health, imports, income, inflation, life expectation, total fertility, GDP etc. Your task is to separate the set of countries which need the aid by separating data into different clusters using K Means Clustering.

1. Import the dataset and understand it

2. Now explore the data and implement data pre-processing.

1. Find optimal number of clusters(k-value) by plotting a graph.

2. Reduce the data set using Principle Component analysis (n_components=2)

3. **Now apply K means Clustering and plot the scatter plot to differentiate data points via clusters.**

4. **Identify the set of countries which need the aid.**

Dataset[**https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data**](https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data)

https://colab.research.google.com/drive/1xk9EEeRoCkZABUMmr0xJD7ERqhe3fGjY?usp=sharing