UNIT-3

1)Feature engineering on categorical data

There are several types of categorical feature engineering techniques that can be applied to transform categorical variables into numerical or binary representations. Here are some commonly used techniques:

• **One-Hot Encoding:** One-hot encoding is a widely used technique where each category in a categorical variable is transformed into a binary feature. Each feature represents whether a particular category is present (1) or absent (0) in the original variable.

- **Label Encoding:** Label encoding assigns a unique numerical label to each category in a categorical variable. It replaces the original categories with their corresponding numerical labels.
- **Ordinal Encoding:** Ordinal encoding is similar to label encoding but preserves the order of the categories.
- **Frequency Encoding:** Frequency encoding replaces categories with their corresponding frequencies in the dataset. It assigns a numerical value to each category based on how frequently it appears.

Feature engineering on image data

Features that are obtainable from the image EXIF data:

- Image create date and time
- Image dimensions
- Image compression format
- Image resolution and aspect ratio
- Flash, aperture, focal length, and exposure

- ➢ An image can be represented by the value of each of its pixels as a two dimensional array.
- ➢ We can use numpy arrays.
- ➢ Color images usually have three components also known as channels.

- The R, G, and B channels stand for the red, green, and blue channels, respectively.
- This can be represented as a three dimensional array (m, n, c) where m indicates the number of rows in the image, n indicates the number of columns. These are determined by the image dimensions. The c indicates which channel it represents (R, G or B)
- Converting images to gray scale is necessary to convert color image representation as two-dimensional image.
- Each pixel value can be computed using the equation
  Y = 0.2125 x R + 0.7154 x G + 0.0721 x.

(i)Edge Detection:

- The canny edge detector algorithm is an edge detector algorithm. This algorithm typically involves using a Gaussian distribution with a specific standard deviation σ (sigma) to smoothen and denoise the image.
- Then Sobel filter has to be applied to extract image intensity gradients

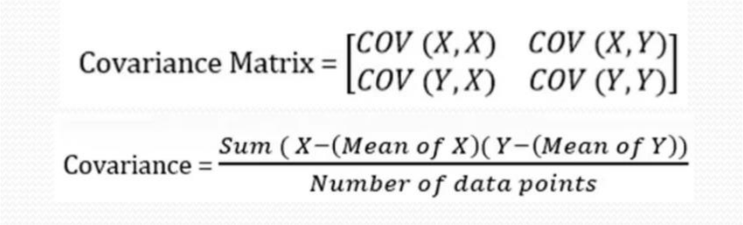
(ii) HOG algorithm:

- The image is normalized and denoised to remove excess illumination effects.
- First order image gradients are computed to capture image attributes like contour, texture, and so on.
- Gradient histograms are built on top of these gradients based on specific windows called cells.
- Finally these cells are normalized and a flattened feature descriptor is obtained, which can be used as a feature vector for models.

2)PCA (Principal Component Analysis)

- A very popular technique of linear data transformation from higher to lower dimensions is Principal Component Analysis, also known as PCA.
- Principal component analysis, is a statistical method that uses the process of linear, orthogonal transformation to transform a higher dimensional set of features that could be possibly correlated into a lower-dimensional set of linearly uncorrelated features.
- In any PCA transformation, the total number of PCs is always less than or equal to the initial number of features.
- The first principal component tries to capture the maximum variance of the original set of features.
- Each of the succeeding components tries to capture more of the variance such that they are orthogonal to the preceding components.

STEP 1: STANDARDIZATION
Calculate the Mean and Standard Deviation for each feature and then, tabulate the same
STEP 2: COVARIANCE MATRIX COMPUTATION



$$\text{Covariance Matrix} = \begin{bmatrix} COV\ (X,X) & COV\ (X,Y) \\ COV\ (Y,X) & COV\ (Y,Y) \end{bmatrix}$$

$$\text{Covariance} = \frac{Sum\ (\ X-(Mean\ of\ X)(\ Y-(Mean\ of\ Y))}{Number\ of\ data\ points}$$

STEP 3: (COV(X, Y)=COV(Y, X)).

- If the value of the Covariance Matrix is positive, then it indicates that the variables are correlated.
- If the value of the Covariance Matrix is negative, then it indicates that the variables are inversely correlated.

STEP 4: FEATURE VECTOR

- To determine the principal components of variables, you have to define eigen value and eigen vectors for the same.
- Let A be any square matrix. A non-zero vector v is an eigenvector of A if Av = λv
- Then, substitute each eigen value in (A-λI)v=0 equation and solve the same for different eigen vectors.

- Now, calculate the sum of each Eigen column, arrange them in descending order and pick up the topmost Eigen values. These are the Principal components

STEP 5: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

Final Data Set = (Standardized Original Data Set) * (FeatureVector)

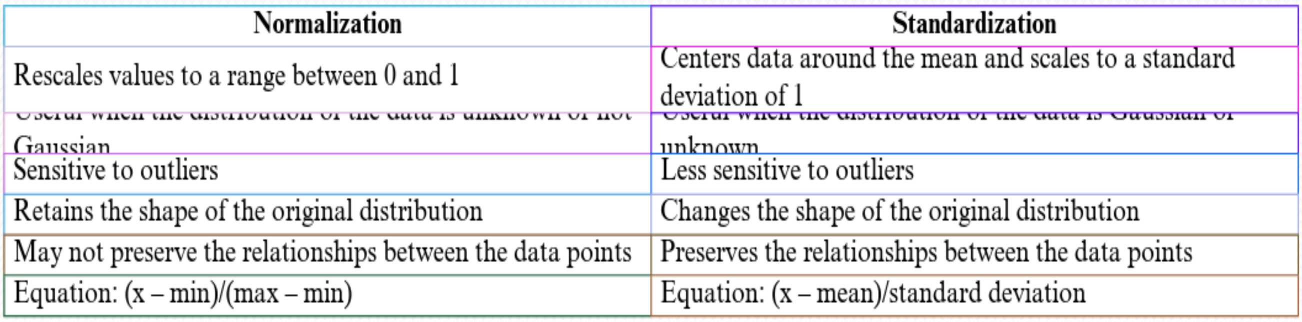
3)Feature Scaling

- When dealing with numeric features, certain attributes may be completely unbounded in nature, like view counts of a video or web page hits.
- Models are sensitive to the magnitude or scale of features like linear or logistic regression
- Standardized Scaling(Standardization technique)- This is also popularly known as Z-score scaling.

$$SS(X_i)=\frac{X_i-\mu_X}{\sigma_X}$$

- Min-Max Scaling(Normalization technique)- We can transform and scale our feature values such that each value is within the range of [0, 1].

$$MMS(X_i)=\frac{X_i-\min(X)}{\max(X)-\min(X)}$$



Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

4)Feature Selection

- We have to select an optimal number of features to train and build models that generalize very well on the data and prevent overfitting.
- They are classified as: Filter methods, Wrapper methods, Embedded methods

1.Filter methods

- These techniques select features purely based on metrics like correlation, mutual information etc.
- These methods do not depend on results obtained from any model and usually check the relationship of each feature with the response variable to be predicted.
- Popular methods include threshold based methods and statistical tests
- **Threshold based methods** This is a filter based feature selection strategy, where you can use some form of cut-off or thresholding for limiting the total number of features during feature selection
- **Statistical Methods**
  - To select features based on univariate statistical tests.
  - Techniques available are: Mutual information, ANOVA (analysis of variance) and chi-square tests.
  - Based on scores obtained from these statistical tests, you can select the best features on the basis of their score.

2. Wrapper methods

- These techniques use a recursive approach to build multiple models using feature subsets and select the best subset of features giving us the best performing model.
- Methods like backward selecting and forward elimination are popular wrapper based methods.
- Recursive Feature Elimination (RFE): This strategy is also popularly known as backward elimination. Recursive Feature Elimination, also known as RFE is a popular Wrapper Method

3. Embedded methods

- These techniques try to combine the benefits of the other two methods by leveraging Machine Learning models themselves to rank and score feature variables based on their importance.
- Tree based methods like decision trees and ensemble methods like random forests are popular examples of embedded methods