Lab viva notes

Concepts

Data Preprocessing

Data Processing is required:

1. transforming raw data into an understandable format
2. get rid of incomplete, noisy and inconsistent data
3. better organise the data

Techniques:

1. **Data Cleaning**: Data Cleaning/Cleansing routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. We use methods like `dropna()` to drop null values.

2. **Dimensional Reduction (Feature Selection)**: Dimensional reduction techniques reduce the dimensionality of the data. It is concerned with reducing the number of input features in training data.

3. **Feature Engineering**: Feature engineering techniques are used to create new features from existing features. E.g. - decomposing categorical attributes from your dataset.

4. **Data Integration**: Data Integration is the process of combining data from different sources to form a single dataset.

5. **Data Transformation**: Data Transformation is the process of transforming data into a new form. Some of the main techniques used to deal with this issue are:

    - **Transformation for categorical variables** is the process of converting categorical variables into numeric encodings. We use methods like `get_dummies()` to convert categorical variables into dummy/indicator variables.

    - **Min-Max Scaler / Normalization**: The min-max scaler, also known as normalization, is one of the most common scalers and it refers to scaling the data between a predefined range (usually between 0 and 1).

    - **Standard Scaler / Standardization**: The standard scaler, also known as standardization, is another common scaler and it refers to scaling the data to have a *mean of 0* and a *standard deviation of 1*.

6. **Handling Data with Unequal Distribution of Classes**: There are three main techniques that we can use to address this deficiency in the dataset:

    - **Over-sampling**: Duplicating samples from the minority class.
    - **Under-sampling**: Deleting samples from the majority class.
    - **Hybrid**: It combines both over-sampling and under-sampling techniques.

Supervised learning vs Unsupervised learning

- **Supervised learning**: The data is *labelled*. The model is trained on the labelled data and then used to predict the labels of the unlabelled data. Eg: Classification, Regression

- **Unsupervised learning**: The data is *unlabelled*. The algorithms discover hidden patterns in data without the need for human intervention. Eg: Clustering

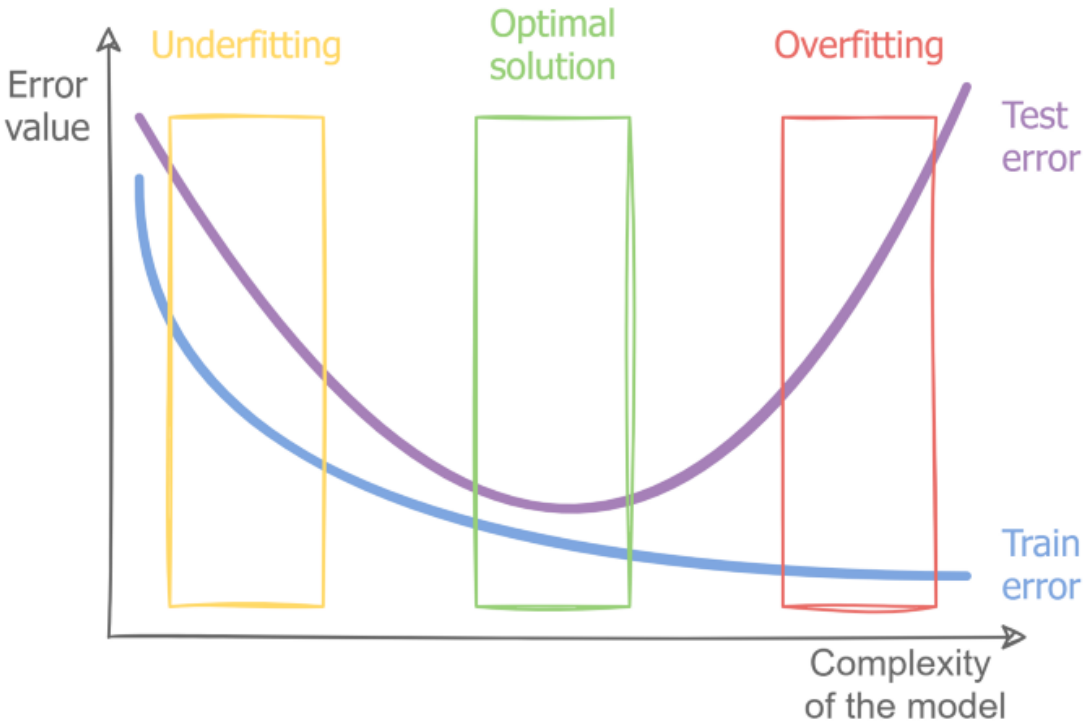
Overfitting & Underfitting

- **Underfitting**
  - When the model unable to find relations between the data.
  - This happens when the dataset has less features.
  - Model is too simple for your data
- **Overfitting**
  - Usually takes place when a model has an excessively complex structure.
  - Learns both the existing relations among data and noise.
  - Model is too complex for your data

Bias: Bias is a phenomenon that skews the result of an algorithm in favor or against an idea.

Variance: Variance refers to the changes in the model when using different portions of the training data set.

- low bias, low variance => Good Model
- low bias, high variance => Overfitting
- high bias, low variance => Underfitting
- high bias, high variance => Bad Model (*ML chere dao*)



To fix

- Underfitting
  - Increase the number of features
  - Increase the complexity of the model (use SVM with different kernels instead of logistic regression)
- Overfitting

- Reduce the number of features
- Use regularization
- Use cross validation
- Sometimes having more data helps

Regularization is an indirect and forced simplification of the model. The regularization term requires the model to keep parameters values as small as possible, so requires the model to be as simple as possible.

Feature Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

Methods of feature selection:

1. **Filter Method**: Features are dropped based on their correlation with dependent variable. Eg: Pearson correlation, Chi-Square test
2. **Wrapper Method**: Split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again. Eg: Forward selection, Backward elimination
3. **Embedded Method**: This method combines the filter and wrapper method. It uses a machine learning algorithm to select the features. Eg: Lasso regression, Ridge regression

Regression

We have two different variable sets:

- **Independent variables:** These are the variables that are used to predict the dependent variable.
- **Dependent variables:** the variable we want to predict

Linear Regression

- The dependent variables are conitnuous in nature.
- The relationship b/w indep and dep variables is **linear**.

Types of linear regression:

1. Simple linear regression

- Only one independent variable

$$ y = \alpha_0 + \alpha_1 x \ \begin{aligned} \alpha = \text{Regression coefficient} \ x = \text{Independent variable} \ y = \text{Dependent variable} \
\end{aligned} $$

2. Multiple linear regression

- More than one independent variable

$$ y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n \ \begin{aligned} \alpha_i = \text{Regression coefficient} \ x_i = \text{Independent variables} \ y = \text{Dependent variable} \
\end{aligned} $$

Polynomial regression

Regression analysis in which the relationship between the independent variables and dependent variables are modeled in the nth degree polynomial.

$$ y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots + \alpha_n x^n \ \begin{aligned} \alpha_i = \text{Regression coefficient} \ x = \text{Independent variable} \ y = \text{Dependent variable} \ \end{aligned} $$

Logistic Regression

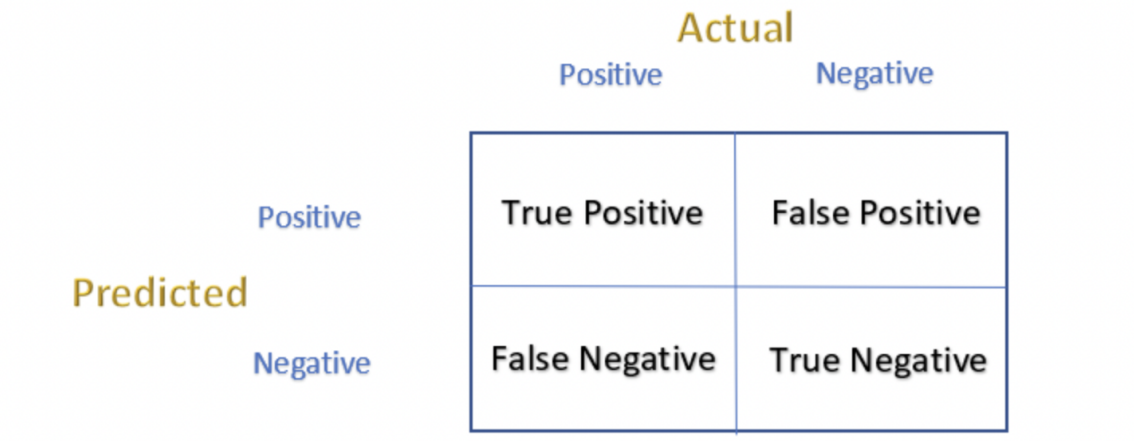
The dependent variable is binary in nature, its value is either 0 or 1.

Examples:

- Whether an email is spam or not
- Whether cancer is malignant or benign

We can also use `multiclass` logistic regression to predict more than two classes.

Confusion Matrix



Accuracy: It is how often did the model predict the event correctly. The ratio of correctly predicted events to the total events. It is given by:

$$ Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} $$

Precision: It is how often the model predicted the event to be positive and it turned out to be true. The ratio of correctly predicted positive events to the total predicted positive events. It is given by:

$$ Precision = \frac{TP}{(TP+FP)} $$

Recall: Out of the total positive, what percentage are predicted positive. It is given by:

$$ Recall = \frac{TP}{(TP+FN)} $$

> *NOTE: Both Precision and Recall has TP (True Positive) in the numerator.*

F1 Score: It is the harmonic mean of precision and recall. It is given by:

$$ \text{F1 Score} = \frac{2*Recall*Precision}{(Recall+Precision)} $$

Handling Catagorical Data

One Hot Encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Integer Encoding

Here, each unique category value is assigned an integer value.

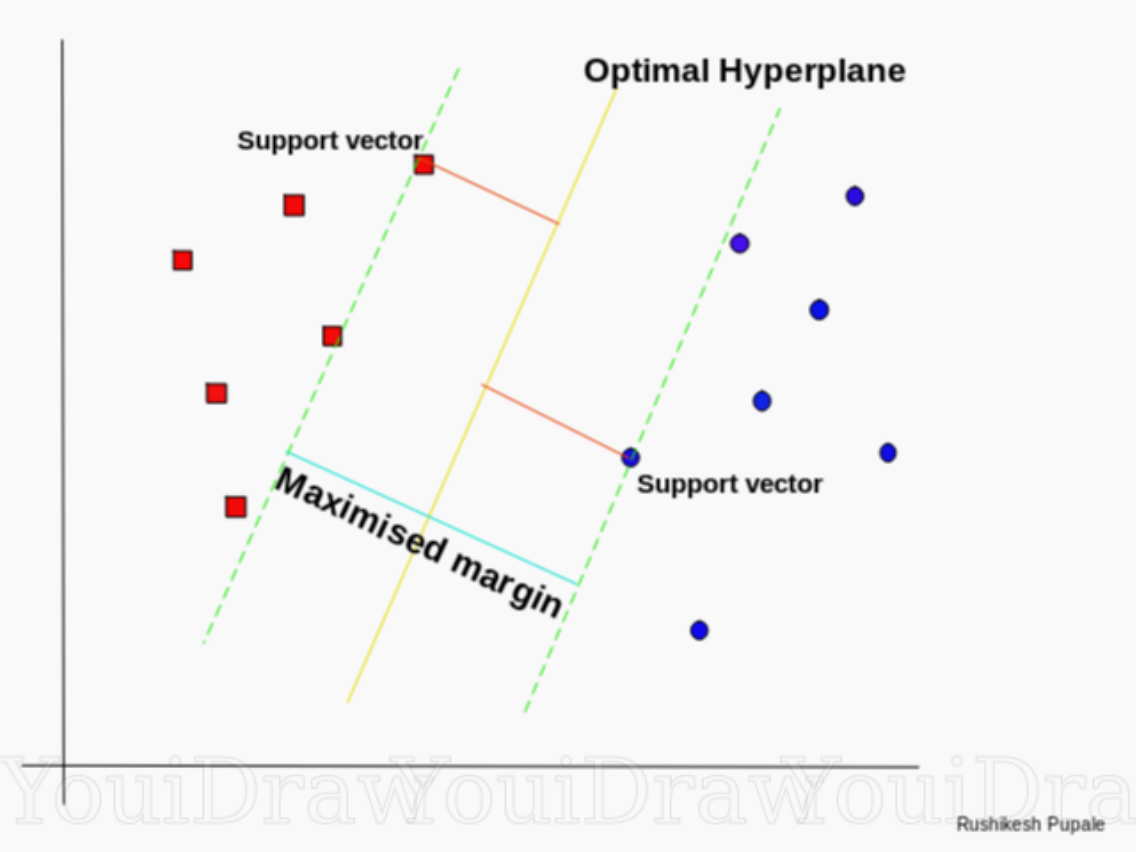
Problem with Integer Encoding is that the model might make the assumption that two nearby values are more similar than two distant values.

SVM (Support Vector Machine)

- linear model for classification and regression problems.
- algorithm creates a line or a hyperplane which separates the data into classes.

Support Vectors: The data points that are closer to the hyperplane.

Hyperplane: A hyperplane in an n-dimensional Euclidean space is a flat, $n-1$ dimensional subset of that space that divides the space into two disconnected parts.



Decision Tree

- A decision tree is a graphical representation of all possible solutions to a decision based on certain conditions.
- used for both classification and regression problems.
- linear but mostly used for non-linear data.
- use CART algorithm (Classification and Regression Tree).

Entropy: It gives the measure of impurity or randomness in the data.

Information Gain: The information gain is the decrease in the entropy after the dataset is split on the basis of an attribute.

Naive Bayes classifier

- Naive Bayes classifier is a probabilistic machine learning model that's used for classification task.

Bayes Theorem: It is used to find the probability of an event based on the prior knowledge of conditions that might be related to the event.

$$ P(A|B) = \frac{P(B|A)P(A)}{P(B)} $$

Types of Naive Bayes Classifier:

Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

Bernoulli Naive Bayes:

This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

Reference

1. Overfitting and Underfitting
2. Logistic Regression
3. Oversampling and Undersampling
4. Error Metrics
5. One Hot Encoding
6. SVM
7. Decision Tree
8. Naive Bayes Classifier

TODO

- K-fold