**ML Notes**

**AI vs ML vs DL vs DS**

AI: the smart application that can perform its own task without human intervention.

e.g Self driving car, robots

ML: It provides stats tool to analyze, visualize, predictive models, forecasting

e.g. Recommendation system

Deep learning DL: Mimic the human brain.

e.g. Object detection, image recognition.

Data Science DS: all above skill

Type of LEarning

1. Supervised ML: We know input and output, We know the expected result.
   1. Classification: O/P feature of the dataset, o/p categorise data
   2. Regressions: O/P continuous value.
2. Unsupervised ML: No O/P, The Target is to create cluster.
3. Semi-supervised; combo of above.
4. Reinforcement Learning: e.g. As child, we have to train.

Type of dataset

1. Training dataset: We will train our model on this dataset.
2. Validation Dataset: Hyper Tuning of the model, to improvised accuracy.
3. Test Dataset: Model will be test.
4. Model Performance
5. Overfitting, Underfitting:
6. Bias vs Variance

**Overfitting**: with train data, the model gives good accuracy but test data gives low

**Underfitting**: with train data, the model gives bad accuracy but test data gives low

**Generalized model**: with train and test, accuracy is good.

**Bias is related to train data, and Variance is related to test data.**

**Bias:** When Train data accuracy is high then High bias, When Train data low is high then Low bias

**Variance:** When test data accuracy is high then High Variance, When Test data accuracy is low then low bias

In the generalized model: high bias and high variance.

07-10-2023

**Missing value:** No specific data

**Missing data at random MAR:** the missing values are systematically related to the observed data, but not to the missing data.

**Missing data not at random MAR:**

MAR (Missing at Random): Definition: Data is missing at random (MAR) if the probability of missing a data point depends only on observed values and not on unobserved values. Characteristics: Missingness may be related to observed variables in the dataset, but not to the unobserved (missing) values. Once the observed variables are taken into account, the missingness is random.

MNAR (Not Missing at Random): Definition: Data is not missing at random (MNAR) if the probability of missing a data point is related to the unobserved (missing) values. Characteristics: The missingness is related to the specific values that are missing. The missing values may be systematically different from the observed values, and the reason for missingness is tied to the unobserved characteristics of the data.

**Imbalnced data**

**Upsampling**

**downsampling**

13-10-2023

**Feature extraction**

It is process of selecting and extracting the most important feature from raw data.

**Feature selection**: We just pick the most imp feature.

Filter method, embedded method

**Principal component analysis (PCA):**

Feature Scaling:

standardization in ml and normalization in deep learning

* Standardization: Z-Score
* Min – Max Scaling: normalization
* Unit Vector:

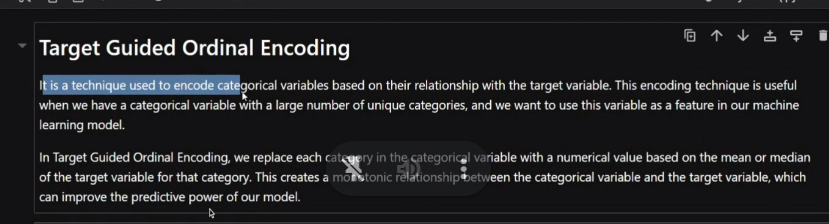
**17-10-2023**

**Nominal / one hot encoding technique**

**It is technique used to transform categorical variable that have no intrinsic ordering into numerical values that can be used in machine learning model. One common**

**18-10-2023**

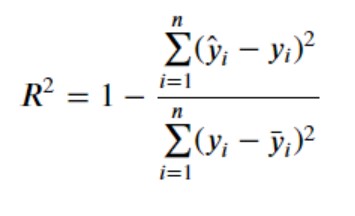
**Target guilded ordinal encoding**



**---**

**Linear regression**

1. **R square :** 1- (Sum(sqare Residual) / sum(square total))



1. **Adjusted R square :**

A black and white math equation

Description automatically generated with medium confidence

**MSE (mean square error):**

**MAE ( mean Absolute error): roburst with outliener**

**RMSE (Root mean square error):**

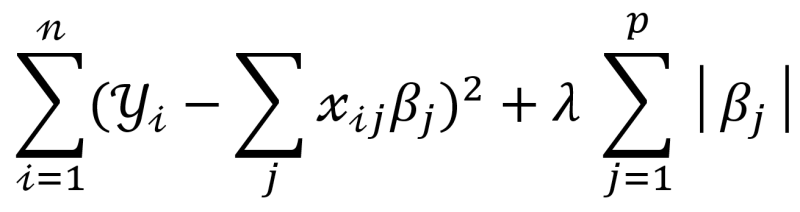
**Regression**

1. **Ridge: reducing overfitting**

A black text with black letters

Description automatically generated with medium confidence

1. **Lasso: Feature selection**



1. **Elastic regression :**

**Use for reducing overfitting**

**Feature selection**

**28-11-2023 – PW –**

**Support Vector classifier in-depth intuition:**

1. **SVC (support Vector classifier) – Classification**

Marginal plans – With the best-fit line, we create a marginal line through the nearest element. That element is called the Support Vector.

If we tried to create a marginal line with different elements then we will consider the highest marginal plan distance i.e. d > d’ (d = marginal plan distance)

**Soft margin and hard margin in SVC:**

**Soft margin:** Some data pints are misclassified but it’s fine. At end of the day we are able predict some point.



**Hard margin:** No points are misclassified.

**SVC Mathematical intuition:**

**Equestion of straight line = y = b[w1x1 + w2x2 + w3x3] =**

**Where w = [w1, x1**

**W2, x = x2 wT = w1, w2, w3**



**W3 x3**

**Y = wTx+ b**

**If the angle is greater than with best-fit line origin then the -ve value.**

So, to find d i.e. marginal plan distance we will subtract first marginal value to another.

**wTx+ b = 0 🡪** best fit line equation

**wTx1+ b = +1 🡪** positive marginal line equation

**wTx2+ b = -1 🡪** negative marginal line equation



**wTx1+ b = +1**

**wTx2+ b = -1**

wT(x1-x2) = ***2/||w||* 🡪** Distance between marginal plan line

1. **SVR (Support Vector Regressor) – Regression**

**Epsilon :** marginal error



**wTx+ b = 0 🡪** best fit line equation

**wTx1+ 🡪** positive marginal line equation



**wTx2- 🡪** negative marginal line equation



**etaa : error above the the margin**



**How it is different from SVC?**

It is used for regression, which is a continuing process. So ideally all elements are supposed to fall in the marginal plane. But practically it’s not happening. So add/submission of element which we represent with etta



**SVM Kernel:**

When data is scattered and presented in a linear format drawing the best-fit line is difficult. So, to differentiate the data we must **transform** data in 2D/3D. After transformation data will appear as separate data and be able to draw the best-fit line.

After drawing a best-fit line, we can process as before method.

**Below are methods to transform data:**

1. **Polynomial Kernel:**
2. **RBF Kernel:**
3. **Sigmoid Kernel:**