# Machine learning

## 1. Linear Regression

### Assumptions in regression analysis:

* ~~There~~ Linear relationship: There should be a linear and additive relationship between dependent (response) and independent (predictor) variable.

To check: Residual vs fitted values.

1. If there is any pattern, then it is non-linear.
2. If a funnel shape is there, then its non-constant variance

### Multicollinearity:

It affects target variable. standard errors tend to increase. confidence become wider and less precise.

check: scatter plot. VIF < 4 means no multicollinearity. VIF > 4 Serious.

### Homoscedasticity or equal variance:

Non-constant variance arises in presence of outliers or extreme leverage value.

TO check: Residual vs fitted plot. The plot exhibits funnel shape patterns.

Normal distribution of error terms

If not, confidence intervals may become wide or narrow. Few unusual data points

check: Q-Q plot, shapiro-wilk test

The q-q quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution

The straight line shows normality

NO Autocorrelation

When the residuals are dependent on each other, there is autocorrelation.

See the x-y chart, you get to know it

NO. of observation > NO. of predictors

The standard errors are uncorrelated

Evaluation metrics of Regression model

Mean Absolute Error (MAE):

It is a simple metric which calculates the absolute difference between actual and predicted values divided by total no. of points in our dataset.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Advant when to use: you want to know how close the predictions are to the actual on average.

Advantages : The MAE is in the same unit as the output variable.

* It is most robust to outliers

Disadvantages : The graph of MAE is not differentiable So have to apply various optimizers like GD which is differentiable

Mean squared error (MSE)

It states the finding of squared difference between actual and predicted value.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Advantages:

The graph of MSE is differentiable, so we can use loss function.

Disadvantages:

* It penalizes the outlier
* It is not in same unit

Root Mean Square error (RMSE):

It is a simple square root of mean squared error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Advantages:

The output is in same unit.

It is sensitive to outliers.

When to use:

If you are concerned about large error, RMSE is good.

R2

It is how well the model performance is

$$R2 \text{ squared} = 1 - \frac{SSr}{SS_m}$$

R2 squared means the proportion of variance explained by the model

Disadvantage:

When adding new features, R2 starts increasing or remains constant or increasing, it never decreases. because it assumes that while adding more data variance of the data increases.

Adjusted $R^2$

$$R^2 = 1 - \left[ \left[ \frac{n-1}{n-k-1} \right] \times (1 - R^2) \right]$$

where

$n$ = no. of observations

$K$ = no. of independent variables

## 2. Logistic Regression

Assumptions for logistic regression:

The dependent variable must be categorical in nature.

The independent variable should not have multi-collinearity

Logistic:

Instead of fitting a regression line, we fit 'S' shaped logistic function,

Logistic regression is the ratio of correctly ide

$$e^z = \frac{P(x)}{1-P(x)}$$

After applying natural log,

$$P(x; b, w) = \frac{e^{wx+b}}{1+e^{wx+b}} = \frac{1}{1+e^{-wx+b}}$$

If we use the MSE of linear regression, then it will give a non-convex graph with local minima.

$$\text{Log loss} = \frac{1}{N}\sum_{i=1}^{N} -\left(y_i * \log(\hat{y}_i) + (1-y_i) * \log(1-\hat{y}_i)\right)$$

$$\text{Cost function} = \begin{cases} -\log(f(x)) & \text{if } y=1 \\ -\log(1-h(x)) & \text{if } y=0 \end{cases}$$

If learning rate is a big number then we miss the minimum point

# Metrics to evaluate classification model

## 1. Precision

It explains how the pre predicted positives cases actually turned out to be positive.

It is helpful where false positives is a higher concern than false negatives.

Eg: Importance of precision in recommendation leads to customer churn.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

## 2. Recall

It explains how many of the actual positive cases we were able to predict.

It is helpful metric where False negative is of higher concern than False positive.

Eg: Medical field

$$\text{Recall} = \frac{\text{True positive}}{\text{True Positive} + \text{False negative}}$$

Precision can be seen as a measure of quality and recall as a measure of quantity. 4.

Higher precision means that an algorithm returns more relevant results than irrelevant ones and high recall means that an algorithm returns most of the relevant results.

3. ~~Recall~~ (Sensitivity) F1 score

It is a harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot Precision \times Recall}{Precision + Recall}$$

It punishes extreme value.

It is an effective metric in following cases:

when FP & FN are equally costly 5.

when more data doesn't effectively change the outcome.

True Negative is high.

Precision can be seen as a measure of quality and recall as a measure of quantity.

Higher precision means that an algorithm returns more relevant results than irrelevant ones and high recall means that an algorithm returns most of the relevant results.

3. ~~Recall~~ (Sensitivity) F1 Score

It is a harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot Precision \times Recall}{Precision + Recall}$$

It punishes extreme value.

It is an effective metric in following cases:

when FP & FN are equally costly

when more data doesn't effectively change the outcome.

True Negative is high.

4. AV-ROC

The Receiver operator characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against FPR(False positive Rate) at various threshold values and separates the signal from noise.

AUC (Area under curve) is the measure of ability of a classifier to distinguish between classes.

x axis -> False positive rate

y axis -> True positive rate

Higher the value of x means means high False Positive.

The choice of threshold depends on ability to balance between FP and FN.

5. Log loss

$$\text{log\_loss} = y \log(P) + (1-y) \log(1-P)$$

For Imbalanced data, ROC helps.

# 3. Naive Bayes

It is used to predict models for binary or multi-classification labels. It operates on conditional probabilities, which estimate the likelihood of a classification based on the combined factors while assuming independence between them.

## Assumption of Naive Bayes

It assumes that all features in the input data ~~and some prior~~ are independent of each other. used in Sentimental analysis, spam filtering and text classification.

## Bayes' theorem

It is used to determine the probability of a hypothesis with prior knowledge.

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

Process:

1. Convert the given dataset into frequency tables.

2. Generate likelihood table by finding the probabilities of given features.

3. Use Bayes theorem to calculate the posterior probability

Eager learner – Naive Bayes.

Types of Naive Bayes

Gaussian – It follows normal distribution. It assumes that the values are sampled from the Gaussian distribution

Multinomial – It assumes it is taken from multinomial data.

Bernoulli – Assumes the predictor variables are independent Booleans variables. used for document classification.

It (GaussNB) has segregated the datapoints with fine boundary.

* Its a eager learner

5. SVM

The shortest distance between the observations and the threshold is called the margin.

Maximal margin classifier They are super sensitive to the data.

It involves misclassification for variance-bias tradeoff.

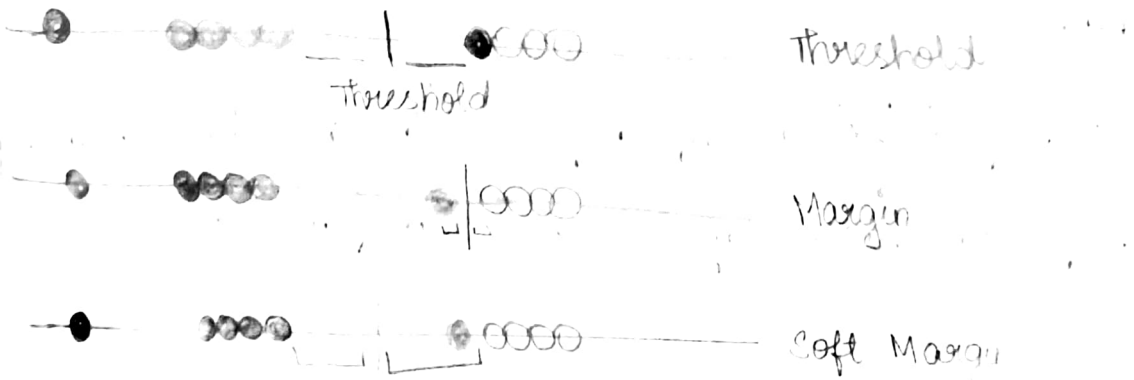
Soft Margin :

When we allow misclassifications the distance between the observations and threshold is called soft margin

Uses CV to determine how many misclassification and observations to allow inside of the soft margin.

Soft Margin classifier (support vector) ~~Machine~~

It comes from that observations on the edge and within soft margin are called support vectors



when 3-dimensional, it is plane. If 4 or more, it is subspace, it is hyperplanes.

## Support Vector Machine

Idea is start with relatively one and move to higher dimension

Find support vector classifier to get line.

How to decide the transformation?

Kernel function decides.

when degree = 1,

It computes the relationship between each pair of observations in 1 dimension and used to find the SVC.

RBF:

It uses infinite dimension. It is like weighted KNN.

SVM: They behave as if they are in the higher dimensions. They don't actually do the transformation.

Polynomial kernel:

$$(a \times b + r)^d$$

r is coefficient, d is the degree, a & b are observation

We get a dot product.

We need to calculate the high dimensional relationship is to calculate the Dot products between each pair of points.

Radial Basis Kernel:

$$e^{-\gamma(a-b)^2}$$

It uses infinite dimensions. It behaves like KNN

$\gamma$ is determined by cross validation.

The less the value, further they are.

Taylor series

Turn f(x) into infinite sum.

# 6. DECISION TREE

Step 1: Take a colum and target and build a tree. Repeat those for all column

Chest pain

Yes ↙ ↘ No

| Heart disease | | Heart Disease | |
|---|---|---|---|
| Yes | No | Yes | NO |
| 105 | 39 | 34 | 125 |

Because none of the leaf nodes are 100% Yes or No. They all are Impure.

To determine which separation is best, we need a way to measure and compare impurity.

For a leaf,

The Gini impurity = 1 − (Probability of Yes)$^2$ − (Probability of No)$^2$

Gini Impurity for a column node = weighted average of Gini impurities for the leaf nodes.

$= \left(\frac{144}{144+159}\right) 0.395 + \left(\frac{159}{144+159}\right) 0.336$

The one =

Step 2 : The one with the lowest Impurity is made the root node.

Step 3 : Take another column for left and select lowest branche entropy.

Step 4 : DO the same thing for right side.

Re

REGRESSION TREES :

Find the mean of the sort adjacent values.

Find Gin's Impurity for each of them.

# STATISTICS

## CENTRAL LIMIT THEOREM:

In order to apply center limit theorem, there are four conditions to be met.

1. Randomization
2. Variables should be independent and identical
3. When sample is drawn without replacement, the sample size should be no longer than 10% of the population
4. Sample size needs to be sufficiently large.

Statement:

Suppose that a large sample of observation is obtained, each observation being randomly produced in a way does not depend on other, the average of observed values is computed. If the procedure is performed many times, resulting in an collection of

observed averages, the CLT says if the sample size was large enough, the probability distribution of these averages will closely approximate a normal distribution.

Law of large numbers:

It is a theorem that describes, the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to expected value as more trials are performed.

The LLN applies to the average. Therefore, while

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{x_i}{n} = \overline{x}$$

Standard deviation vs standard error

Standard error: The standard deviation of the means of the mean is called standard error.

S.D: It quantifies the variation within a set of measurements.

$$S.E = \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals

It is interval that covers the percentage of means of ~~means~~ sample means. It is calculated after Bootstrapping.

Population parameters

Why dividing by N underestimates the variance?

When we divide by n, the value around the sample mean is always less than the value around population mean.

We need squares to find the derivative and finding the minimum value would be harder.

Likelihood.

L (mean = 32 & S.D = 2.5 | mouse weighs 34 grams)

= 0.12

while probability is P( mouse >34 grams | m = 32 & SD = 2.5)

probabilities are the areas under a fixed distribution. Likelihood are the y-axis values for fixed data points with distributions that can be moved.

Maximum likelihood Estimation.

MLE is a statistical method used to estimate the parameters of a probability distribution that best explains a given set of data.

It aims to find the values of these parameters that make the observation most probable.

Conditional probability

$$P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{i=1}^{n} P(E_k)P(A|E_k)}$$

is the bayes theorem

According to conditional probability,

$$P(E_i|A) = \frac{P(E_i \cap A)}{P(A)}$$

1. A bag contains 4 white and 6 black balls while bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bag and it is found to be black. Find it is from bag I

$$P(I|B) = \frac{\frac{1}{2} \times \frac{6}{10}}{\frac{1}{2} \times \frac{6}{10} + \frac{1}{2} \times \frac{3}{7}} = \frac{\frac{6}{20}}{\frac{6}{20} + \frac{3}{14}}$$

2. A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the no. obtained is a four. Find the probability that no. obtained is actually a four.

$$P(E_i|A) = \frac{P(E_i)\,P(A|E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A|E_i)}$$

$$(\text{four}|\text{four}) = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{2}{3} \times \frac{1}{6} + \frac{1}{3} \times \frac{5}{6}}$$

$$= \frac{\frac{2}{18}}{\frac{2}{18} + \frac{5}{18}}$$

$$= \frac{2}{7}$$