MACHINE LEARNING FORMULA CHEAT SHEET

– Omkar Jagtap

Index

- Normalization
- Robust Scaling
- Maximum Absolute Scaling
- Min Max Scaling

Evaluation Metrics Formulas

- Mean Absolute Error
- Mean Square Error
- Root Mean Square Error
- Sum Square Error
- Mean Square Error
- Root Mean Square Error
- R Square
- Accuracy
- Precision
- Recall
- F1 Score

1. STATISTICS

MEAN (AVERAGE)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MEDIAN

if *n* is odd,

$$median = \left(\frac{n+1}{2}\right)^{th}$$

if *n* is even,

$$median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

n = number of terms
th = n(*th*) number

MODE

(Most recurring value)

$$M_o = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

Where

l = lower limit of the modal class,

h = size of the class interval (assuming all class sizes to be equal),

f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class.

VARIANCE

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

where:

- x_i represents each individual value in the data set,
- $\bar{x}$ is the mean (average) of the data set, and
- n is the total number of values in the data set.

STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

In this formula:

- x_i represents each individual value in the data set,
- $\bar{x}$ is the mean (average) of the data set,
- $\sum$ denotes the sum over all values, and
- n is the total number of values in the data set.

CO-VARIANCE

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n}$$

where:

- x_i and y_i are the individual values of variables X and Y in the data set,
- $\bar{X}$ and $\bar{Y}$ are the means (averages) of variables X and Y, and
- n is the total number of paired observations.

CO-RELATION

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- $\mathrm{cov}(X, Y)$ is the covariance between variables X and Y,
- σ_X is the standard deviation of variable X,
- σ_Y is the standard deviation of variable Y.

CONFIDENCE INTERVAL

$$\text{Confidence Interval} = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

where:

- $\bar{x}$ is the sample mean,
- s is the sample standard deviation,
- n is the sample size, and
- Z is the Z-score associated with the desired level of confidence.

P-value

If p-value is low (compared to alpha) let the null Hypothesis GO

RANGE

$$R = \text{Maximum Value} - \text{Minimum Value}$$

QUARTILE

The Quartile Formula for Q1 = $\frac{1}{4}$ $(n + 1)^{th}$term

The Quartile Formula for Q3 = $\frac{3}{4}$ $(n + 1)^{th}$term

The Quartile Formula for Q2 = Q3 – Q1 (Equivalent to Median)

Inter Quartile Range

$$\text{IQR} = Q3 - Q1$$

MACHINE LEARNING

Linear Regression

$$Y = b_0 + b_1 \cdot X + \varepsilon$$

where:

- Y is the dependent variable,
- X is the independent variable,
- b_0 is the y-intercept (the value of Y when X is 0),
- b_1 is the slope of the line (the change in Y for a one-unit change in X),
- ε is the error term (representing the difference between the observed and predicted values).

Logistic Regression (Sigmoid Function)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

KNN/ K Means Euclidean Distance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

In three-dimensional space, the formula extends to:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

And more generally, for n-dimensional space:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

where a_i and b_i are the coordinates of the two points in each dimension.

KNN/ K Means Manhattan Distance

$$d = |x_2 - x_1| + |y_2 - y_1|$$

In three-dimensional space, it extends to:

$$d = |x_2 - x_1| + |y_2 - y_1| + |z_2 - z_1|$$

And more generally, for n-dimensional space:

$$d = \sum_{i=1}^{n} |a_i - b_i|$$

Decision Tree (Entropy)

$H(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$

where:

- S is the set of labels (e.g., the target variable in a dataset),
- p_1 is the proportion of instances in S belonging to class 1,
- p_2 is the proportion of instances in S belonging to class 2 (for binary classification, $p_2 = 1 - p_1$),
- $\log_2$ is the logarithm base 2.

Decision Tree (Information Gain-Entropy)

$IG(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot H(S_v)$

where:

- $H(S)$ is the entropy of the set S (a measure of impurity),
- $\text{values}(A)$ is the set of possible values for feature A,
- S_v is the subset of instances in S for which feature A has value v,
- $|S|$ and $|S_v|$ are the sizes of sets S and S_v, respectively.

Decision Tree (Gini Impurity)

$Gini(S) = 1 - \sum_{i=1}^{K} p_i^2$

In this formula:

- p_i is the proportion of instances in class i in the set S.
- The summation goes over all K classes present in the set.

Decision Tree (Gini Impurity – Information Gain)

$\text{Information Gain} = Gini(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot Gini(S_v)$

In this formula:

- $Gini(S)$ is the Gini impurity of the entire set S.
- $\text{values}(A)$ represents the distinct values that feature A can take.
- $|S_v|$ is the size of the subset of instances where feature A has the value v.
- $|S|$ is the size of the entire set S.
- $Gini(S_v)$ is the Gini impurity of the subset S_v.

Bayes’ Theorem

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

Here's the breakdown of the terms:

- $P(A|B)$: This is the probability of event A occurring given that event B has occurred. This is often referred to as the posterior probability.
- $P(B|A)$: This is the probability of event B occurring given that event A has occurred. This is known as the likelihood.
- $P(A)$: This is the prior probability of event A occurring. It represents our initial belief in the probability of A before observing any evidence.
- $P(B)$: This is the probability of event B occurring. It acts as a normalization factor.

SVM (Linear Kernel)

$$K(x_1, x_2) = x_1^T x_2$$

SVM (Polynomial Kernel)

$$K(x_1, x_2) = (x_1^T x_2 + r)^d$$

SVM (Radial Basis Function [rbf] Kernel)

$$K(x_1, x_2) = \exp(-\gamma . || x_1 - x_2 ||^2)$$

SVM (Sigmoid Kernel)

$$K(x_1, x_2) = \tanh(\gamma . x_1^T x_2 + r)$$

PRE-PROCESSING

Standard Scaler

$$Z = \frac{(X-\mu)}{\sigma}$$

where:

- Z is the standardized value,
- X is the original value of the feature,
- μ is the mean of the feature values,
- σ is the standard deviation of the feature values.

Normalization (L1 & L2)

1. **L1 Normalization:**

$$X_{\text{normalized}} = \frac{X}{\sum_{i=1}^{n} |x_i|}$$

2. **L2 Normalization:**

$$X_{\text{normalized}} = \frac{X}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Robust Scaler

$$X_{\text{scaled}} = \frac{X - Q_1(X)}{Q_3(X) - Q_1(X)}$$

where:

- X_{scaled} is the scaled value of X,
- X is the original value of the feature,
- $Q_1(X)$ is the first quartile (25th percentile) of the feature X,
- $Q_3(X)$ is the third quartile (75th percentile) of the feature X.

Max Absolute Scaler

$$X_{\text{scaled}} = \frac{X}{\max(|X_{\max}|, |X_{\min}|)}$$

where:

- X_{scaled} is the scaled value of X,
- X is the original value of the feature,
- $X_{\max}$ is the maximum absolute value of the feature in the dataset,
- $X_{\min}$ is the minimum absolute value of the feature in the dataset.

Min Max Scaler

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

- X_{scaled} is the scaled value of X,
- X is the original value of the feature,
- $X_{\min}$ is the minimum value of the feature in the dataset,
- $X_{\max}$ is the maximum value of the feature in the dataset.

EVALUATION METRICS

REGRESSION

MAE (Mean Absolute Error)

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

where:

- n is the number of observations or data points,
- y_i is the actual or observed value for the i-th data point,
- $\hat{y}_i$ is the predicted value for the i-th data point.

SSE (Sum of Squared Error) ,

MSE (Mean of Squared Error),

RMSE (Root Mean of Squared Error)

1. **Sum of Squared Errors (SSE):**

   $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

2. **Mean Squared Error (MSE):**

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

3. **Root Mean Squared Error (RMSE):**

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

R- SQUARE

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

CLASSIFICATION

Confusion Matrix

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Type I, Type II Error

		Reality	
		True	False
Measured or Perceived	True	Correct ☺	**Type 1 error** False Positive
	False	**Type 2 error** False Negative	Correct ☺

Accuracy, Precision, Recall, F1 Score

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

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

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

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$