Summary : Text Classification

by : Dhea Fajriati Anas

iykra

Table of Contents

01

Text Preprocessing

Text Preprocessing

Case Folding

Converting a word to
lowercase

Remove White Spaces

Remove whitespaces

Tokenizing

Splitting the senteces
to words

Remove Punctuation

Remove punctuation
e.g. [!”#$%&’()*+,-
./:;<=>?@[\]^_`{|}~]

Stemmer

Remove the inflection
of a word to its basic
form

Remove Stopword

Examples of stopwords inl
ndonesian are "yang",
“dan", "di", “dari“ ,etc

02

Feature Extraction

Feature Extraction

Bag of Words (BoW)

Transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

TF-IDF

Evaluates how relevant a word is to a document in a collection of documents.

03

Modelling

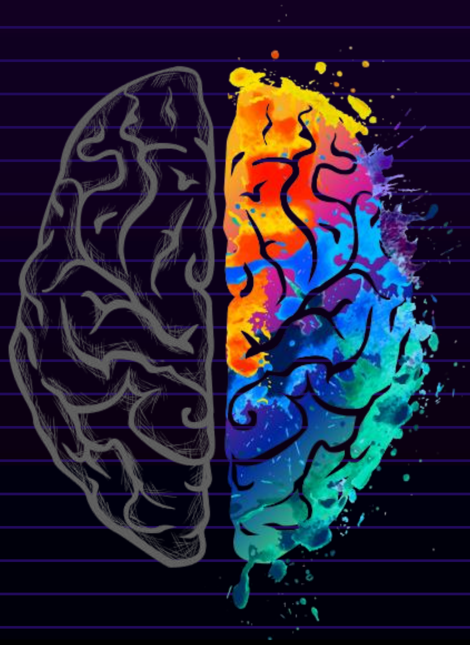
Example of Machine Learning Model

Logistic Regression

Used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.



Support Vector Machine

Supervised machine learning algorithm that can be used for both classification or regression challenges.

Random Forest

Consists of a large number of individual decision trees that operate as an ensemble.

Compare Multiple Models

Logistic Regression

Strong Area:
Linear model
Binary classification

The Core Idea:
Event occurs probability
Odds ratio

Main Hyperparameter:
{C: 0.0001, 10000}
{solver: newton-cg, lbfgs, liblinear, sag, saga}
{penalty: l1. l2}

Naive Bayes

Strong Area:
Text data
Word based classification

The Core Idea:
Bayes theorem
Conditional probability
Human-like estimation

Source: https://medium.com/shortcutnlp/05-model-application-how-to-compare-and-choose-the-best-ml-model-b7cfff804c08

Compare Multiple Models

Random Forest

Strong Area:
Compex non-linear classification
Continuous values (in case of regression trees)

The Core Idea:
Ensemble learning
Bagging (parallel)
Weak learner and strong learner

Main Hyperparameter:
n_estimators = number of trees
max_features = max number of features considered for -splitting a node
max_depth = max number of levels in each decision tree
min_samples_split = min number of data points placed in- a node before the node is split
min_samples_leaf = min number of data points allowed -in a leaf node
bootstrap = method for sampling data points (with or -without replacement

SVM

Strong Area:
Compex non-linear classification
Multiclass classification

The Core Idea:
Kernel methods
Margin Maximization
Hard margin vs Soft margin by C

Main Hyperparameter:
{kernel: rbf, linear} =
{C: 0.0001, 10000} = Regularization: sensitivity for miss classification
{gamma: 0.0001, 10000} =

Source: https://medium.com/shortcutnlp/05-model-application-how-to-compare-and-choose-the-best-ml-model-b7cfff804c08

03

Evaluation

Model Evaluation Metrics

Regression Problem

e.g. R Square, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).

Classification Problem

e.g. Classification Report, Confusion Matrix, Receiver Operating Characteristic (ROC) Curves, AUC,

Regression Problem

R Square

R Square measures how much variability in dependent variable can be explained by the model.

Formula :

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

MAE

MAE is taking the sum of the absolute value of error.

Formula :

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

MSE

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

Formula :

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

RMSE

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because sometimes MSE value can be too big to compare easily, and MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

Classification Problem

Confusion Matrix

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. The confusion matrix is useful for measuring Recall (also known as Sensitivity), Precision, Specificity, Accuracy, and, most importantly, the AUC-ROC Curve.

Accuracy:

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

Recall:

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

Precision:

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

F_1 score:

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

- True Positive (TP) : observation is positive, and is predicted to be positive
- False Negative (FN) : observation is positive, but is predicted negative
- True Negative (TN) : observation is negative, but is predicted to be negative
- False Positive (FP) : observation is negative, but is predicted positive

Thanks