INSTITUTO TECNOLÓGICO DE TIJUANA

DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

SEMESTRE ENERO- JUNIO 2020

CARRERA:

Ingeniería Tecnologías de la Información
y Comunicaciones

MATERIA:

Datos Masivos

UNIDAD POR EVALUAR:

Unidad II

TRABAJO A EVALUAR:

Tarea #2 Vector Assembler

NOMBRE DEL ESTUDIANTE

Barraza Sierra Alexis Fernando

Zamorano Garcia Osvaldo Arturo

PROFESOR:

Dr. Jose Christian Romero Hernandez

VectorAssembler Library

VectorAssembler is a transformer that combines a given list of columns into a single vector column. It is useful for combining raw features and features generated by different feature transformers into a single feature vector, in order to train ML models like logistic regression and decision trees. VectorAssembler accepts the following input column types: all numeric types, boolean type, and vector type. In each row, the values of the input columns will be concatenated into a vector in the specified order.

Examples

Assume that we have a DataFrame with the columns id, hour, mobile, userFeatures, and clicked:

```
 id | hour | mobile | userFeatures     | clicked
----|------|--------|------------------|---------
 0  | 18   | 1.0    | [0.0, 10.0, 0.5] | 1.0
```

UserFeatures is a vector column that contains three user features. We want to combine hour, mobile, and userFeatures into a single feature vector called features and use it to predict clicked or not. If we set VectorAssembler's input columns to hour, mobile, and userFeatures and output column to features, after transformation we should get the following DataFrame:

```
 id | hour | mobile | userFeatures     | clicked | features
----|------|--------|------------------|---------|-----------------------------
 0  | 18   | 1.0    | [0.0, 10.0, 0.5] | 1.0     | [18.0, 1.0, 0.0, 10.0, 0.5]
```

RootMeanSquareError

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error), and square the value. Then you add up all those values for all data points, and, in the case of a fit with two parameters such as a linear fit, divide by the number of points minus two. The squaring is done so negative values do not cancel positive values. The smaller the Mean Squared Error, the closer the fit is to the data. The MSE has the units squared

of whatever is plotted on the vertical axis.

Another quantity that we calculate is the Root Mean Squared Error (RMSE). It is just the square root of the mean square error. That is probably the most easily interpreted statistic, since it has the same units as the quantity plotted on the vertical axis.