SCALABLE AND CLOUD PROGRAMMING

EDUART UZEIR & DOMENICO CORIALE

MULTIPLE LINEAR REGRESSION

- A supervised learning algorithm that allow us to predict a value using multiple predictors in the same time.

$$Y = \beta_0 + \beta_1 X + \beta_2 X + \ldots + \beta_p X_p + \varepsilon$$

- In our case we want to predict the "price" of a house using a set of predictors, that in our model we call "features".

features = (Avg Area Income, Avg Area House Age, Avg Area Number of Rooms, Avg Area Number of Bedrooms, Area Population)

- In the implementation we have used Sparks ***ml.regression.LinearRegression*** library

LOGISTIC REGRESSION

- A Machine Learning Classification algorithm that is used to predict the probability of a categorical dependent variable. The result is binary (0, 1).
- In this implementation we want to know if a person will signup (Y = 1) or not (Y = 0) in a bank service.
- The dataset contain (41,188 records) and 21 fields for each record.

  (https://raw.githubusercontent.com/madmashup/targeted-marketing-predictive-engine/master/banking.csv)
- In the implementation we have used Sparks ***ml.classification.LogisticRegression*** library

K-MEANS CLUSTERING

- Clustering is one example of unsupervised learning where we want to group together similar unlabeled data.
- Examples of clustering may be: grouping together similar documents or customers, market segmentation etc.
- Our goal here is to group together similar seismic regions in the World.
- The dataset contain 5305 record of Earthquakes that have happened in World from 1970 to 2014. Here we are considering 4 predictors (Latitude, Longitude, Depth, Magnitude)

(https://data.humdata.org/dataset/catalog-of-earthquakes1970-2014)