Principal Component Analysis (PCA)

Content Prepared By: Chandra Lingam, Cotton Cola Designs LLC

Principal Component Analysis

Dimension Reduction Technique – retain "information" while reducing features

Real World Datasets have several features that contain similar data – Works great!

New PCA features (*Components*) may be important predictors of target – But…how do you map the "importance" to read-world features?

PCA

Works only for numeric data

Data needs to be normalized – features with similar scale

For large dimension datasets, PCA is an option to reduce features and use that for training a model

Number of Components

Typically, various libraries allow you to specify:

Number of Components

Total Variation to Capture as a percentage (for example capture 90% of the "information") – in this case PCA will figure out required number of components

PCA on SageMaker

Two Modes

Regular - Good for Sparse Data and Moderate sized datasets

Random – Good for very large datasets – uses approximation algorithm

PCA SageMaker – Data Format

Input:

- csv
- recordio-protobuf

Inference:

- csv
- json
- recordio-protobuf

Demo 1 – Random Data Set

PCA with Random Data set

Show that random data set features cannot be reduced much

pca\ExplorePCA\random_data_pca_exploration.ipynb

Demo 2 – Correlated Data Set

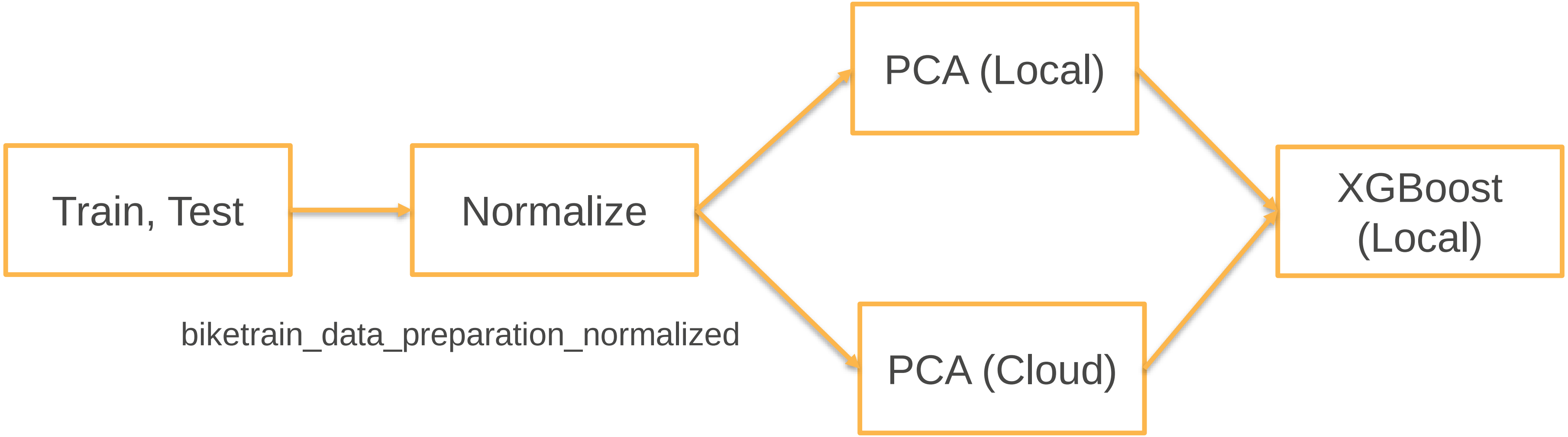
PCA with correlated data set

PCA can capture substantial amount of information with very few components

pca\ExplorePCA\correlated_data_pca_exploration.ipynb

Demo 3 – Kaggle Bike Train with PCA Components

biketrain_pca_projection_localmode



biketrain_data_preparation_normalized

pca\BikeSharingRegression

pca_cloud_training_template
biketrain_pca_projection_cloudmode

Replace: temp, atemp, humidity, windspeed
with PCA Components