Course Title

Big Data Emerging Technologies

❖ **Modules**

1. Big Data Rankings & Products
2. Big Data & Hadoop
3. Spark
4. Spark ML & Streaming
5. Storm
6. IBM SPSS Statistics Project

Big Data

Spark ML (Machine Learning)

Spark Machine Learning

❖ Apache Spark's MLlib (ML library)

- Spark's ML (Machine Learning) library
- Easy to learn and use
- Very practical functions
- Scalable performance
- High level ML functions
  - ML Algorithms, Featurization, Pipelines, persistence, utilities

Spark Machine Learning

❖ Apache Spark's ML library

- ML Algorithms
  - Tools for constructing, evaluating, and tuning ML Pipelines
- Featurization
  - Feature extraction, transformation, dimensionality reduction, and selection

Spark Machine Learning

❖ Apache Spark's ML library

- Pipelines
  - Learning algorithms for classification, regression, clustering, collaborative filtering, etc.
- Persistence
  - Saving and load algorithms, models, Pipelines, etc.
- Utilities
  - Linear algebra, statistics, data handling, etc.

Spark Machine Learning

❖ Apache Spark's ML library

- **MLlib** includes the DataFrame-based API
- DataFrame-based API in the **spark.ml** package is Spark's primary ML API
  - RDD-based APIs in the spark.mllib package are in maintenance mode since Spark 2.0

Spark Machine Learning

- DataFrame-based API (spark.ml package)

Basic statistics	**correlation, hypothesis testing**
Pipelines	**DataFrame, pipline components (transformers, estimators), pipeline, parameters, saving and loading pipelines**
Extracting, transforming and selecting features	**feature extractors, feature transformers, feature selectors, LSH (Locality Sensitive Hashing) operations, LSH algorithms**
Classification and Regression	**classification, regression, linear methods, decision trees, tree ensembles (random forests, gradient-boosted trees)**
Clustering	***k*-means, Gaussian mixture, LDA (Latent Dirichlet Allocation), bisecting *k*-means, GMM (Gaussian Mixture Model)**

Spark Machine Learning

- DataFrame-based API (spark.ml package)

Collaborative filtering	**scaling of the regularization parameter, cold-start strategy (drop any rows in the DataFrame of predictions that contain NaN values)**
Frequent pattern mining	**FP (Frequent Pattern)-growth**
Model selection and tuning	**model selction (a.k.a hyperparameter tuning), cross-validatioin, train-validation split**
Optimization of linear methods	**L-BFGS (Limited-memory BFGS), normal equation solver for weighted least squares, IRLS (Iteratively Reweighted Least Squares)**

Big Data

References

References

- Spark 2.2.0 Machine Learning Library Guide [Online]. Available: https://spark.apache.org/docs/latest/ml-guide.html