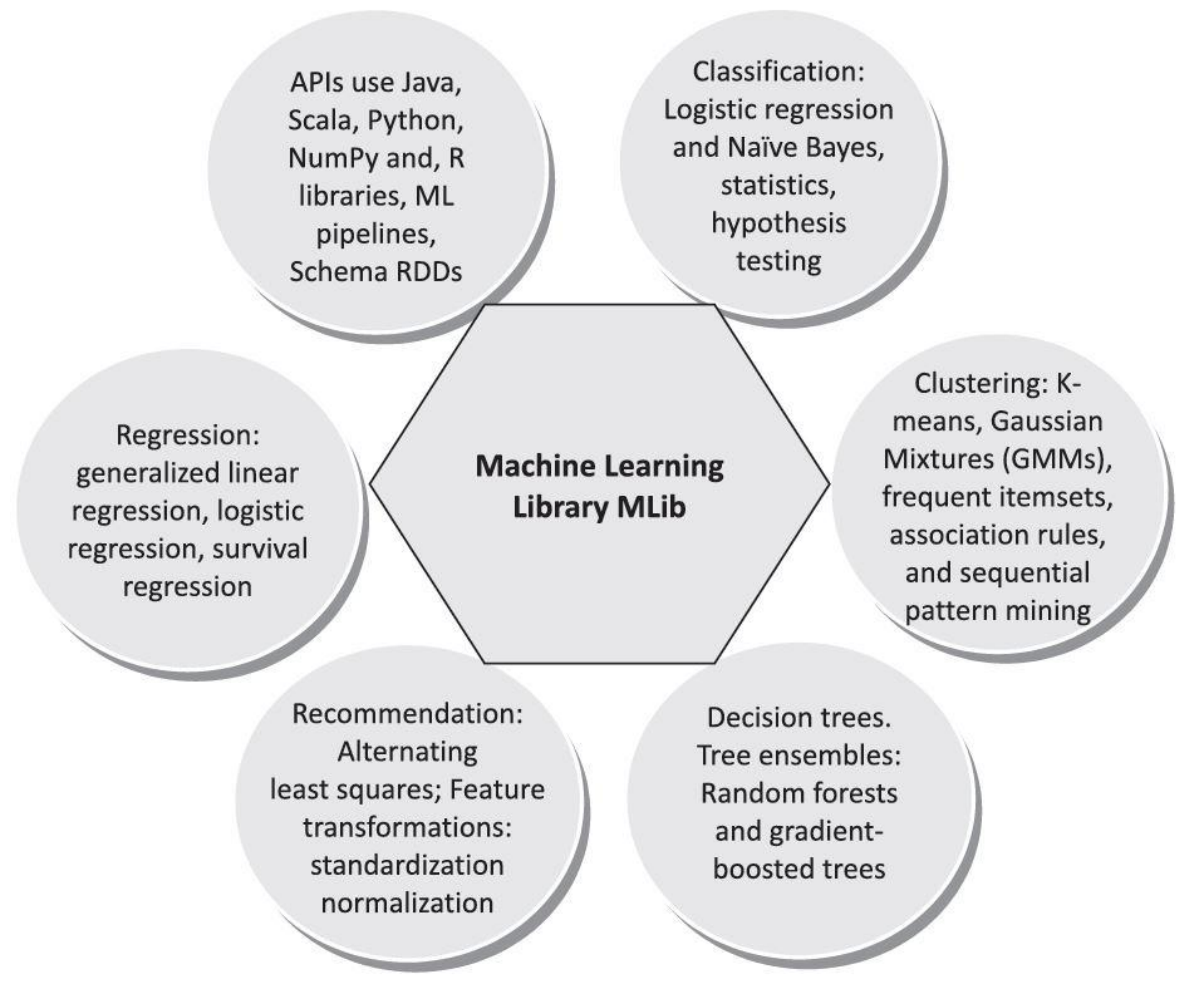
Lesson 6

Machine Learning with Apache® Spark™ MLib

ML Datasets

- ML datasets as RDDs, thus use HDFS, HBase or local files
- MLib APIs are interoperable with Spark SQL MLib Python implementation adds Python APIs
- MLib interoperates with NumPy in Python

Figure 5.8 Main usages of machine learning library MLib



Machine Learning Library MLib

APIs use Java, Scala, Python, NumPy and, R libraries, ML pipelines, Schema RDDs

Classification: Logistic regression and Naïve Bayes, statistics, hypothesis testing

Clustering: K-means, Gaussian Mixtures (GMMs), frequent itemsets, association rules, and sequential pattern mining

Decision trees. Tree ensembles: Random forests and gradient-boosted trees

Recommendation: Alternating least squares; Feature transformations: standardization normalization

Regression: generalized linear regression, logistic regression, survival regression

Spark Support Machine Learning

- ML pipelines
- Data taken from data sources, passes through the machine learning programs in between and the output becomes input to the application

MLIB Pipeline Functions

- Decision tree
- knowledge discovery
- clustering and mining

ML applications

- Use the Python UDFs, VUDFs, block-level UDFs with block-level arguments and return types, complex object types (array map and structure), and conversions or transformations of object types

Summary

- Main usages of machine learning library MLib
- Python and Spark for ML pipelines
- MLib applications

End of Lesson 6 on **Machine Learning with Apache® Spark™ MLib**