COMP 543, Tools and Models for Data Science

Chengyin Liu, cl93

Research #5

K-NN classification

This paper [1] describes the nearest neighbor classification algorithm, which assigns the class identification of the nearest of a set of classified points to an unclassified point. The important point is that not only the nearest neighbor rule is introduced in this paper, but also the proof of upper bound of the error. The author shows that the probability of error of the nearest neighbor algorithm is bounded by twice the Bayes probability of error. It is to say, at least half of the available information in the data collection of classified samples is contained in the nearest neighbor.

Spark

Speaking of Spark [2], it is a new big data framework that supports the applications which reuse data across multiple parallel operations. Hadoop may have trouble on these applications since it involves a lot of reading and writing from the disks. But Spark can handle that problem by introduce the abstraction called resilient distributed datasets (RDDs). RDDs are read-only collections of objects partitioned across a set of machines, which constitute a cluster operating system. As a result, the authors claim that Spark performs 10 times better than Hadoop in iterative machine learning jobs, which is really inspiring.

By combining Spark procedural processing with the relational model, Michael presents a bold idea called Spark SQL [3]. This platform seeks to leverage the benefits of both declarative queries and optimized storage, with the ability to do complex analytics that require a procedural language to describe the algorithm. Spark SQL achieves this through the DataFrame API, the Spark SQL equivalent of a table, which integrates with procedural code. By combining this with the optimizer Catalyst, this paper shows us a very powerful system.

References:

[1] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.

[2] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." HotCloud 10.10-10 (2010): 95.

[3] Armbrust, Michael, et al. "Spark sql: Relational data processing in spark." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015.