

Data Mining, CSCI 347, Fall 2019
Top Data Mining Algorithms, Nov. 20

Top data mining algorithms according to a 2017 survey by TechLeer

<https://www.techleer.com/articles/438-a-list-of-top-data-mining-algorithms/>

1. C4.5 – Decision tree “divide and conquer”. Uses information gain, rather than gain ratio. Can use continuous and discrete data, creating ranges for the continuous data. Uses a single-pass pruning process to mitigate over-fitting.
2. K-means clustering
3. Support vector machines – learns a hyperplane to classify data into 2 classes; uses kernel tricks
4. Apriori – association rule learner
5. Expectation-maximization, EM, generally used as a clustering algorithm. K can be specified, or the algorithm can decide, iterating over and over (user can specify the maximum iterations); based on statistics algorithm that optimizes the likelihood of seeing observed data while estimating the parameters of a statistical model with unobserved variables
6. Page rank – a number between 0 and 1 that measures the page’s prestige
7. Adaboost, ensemble learning – using multiple algorithms together (bagging – voting, boosting – voting but with weights and stacking – using one to get initial results and then using others to refine those results)
8. kNN – k nearest neighbor, instance based learning
9. Naïve Bayes
10. CART – another version of first C4.5 (J48) but uses a different pruning strategy