Thuật toán học có giám sát k-nearest neighbor

Thuật toán học có giám sát k-nearest neighbor (kNN) Có rất nhiều thuật toán học có giám sát, ở đây em sẽ giới thiệu một thuật toán học có giám sát điển hình, đó là k-nearest neighbor (kNN hay k-láng giềng gần nhất) kNN là phương pháp truyền thống khá nổi tiếng theo hướng tiếp cận thống kê đã được nghiên cứu trong nhiều năm qua. kNN được đánh giá là một trong những phương pháp tốt nhất được sử dụng từ những thời kỳ đầu trong nghiên cứu về phân loại văn bản Ý tưởng của phương pháp này đó là khi cần phân loại một văn bản mới, thuật toán sẽ xác định khoảng cách (có thể áp dụng các công thức về khoảng cách như Euclide, Cosine, Manhattan, …) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất, gọi là k nearest neighbor – k láng giềng gần nhất, sau đó dùng các khoảng cách này đánh trọng số cho tất cả các chủ đề. Khi đó, trọng số của một chủ đề chính là tổng tất cả các khoảng cách ở trên của các văn bản trong k láng giềng có cùng chủ đề, chủ đề nào không xuất hiện trong k láng giềng sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo giá trị trọng số giảm dần và các chủ đề có trọng số cao sẽ được chọn làm chủ đề của văn bản cần phân loại. Trọng số của chủ đề cj đối với văn bản x được tính như sau :



Trong đó : y (di, c) thuộc {0,1}, với: y = 0: văn bản di không thuộc về chủ đề cj y = 1: văn bản di thuộc về chủ đề cj sim (x, d): độ giống nhau giữa văn bản cần phân loại x và văn bản d. Chúng ta có thể sử dụng độ đo cosine để tính khoảng cách:



bj là ngưỡng phân loại của chủ đề cj được tự động học sử dụng một tập văn bản hợp lệ được chọn ra từ tập huấn luyện. Để chọn được tham số k tốt nhất cho thao tác phân loại, thuật toán cần được chạy thử nghiệm trên nhiều giá trị k khác nhau, giá trị k càng lớn thì thuật toán càng ổn định và sai sót càng thấp.