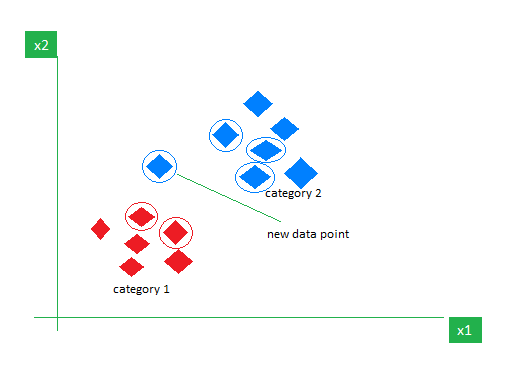
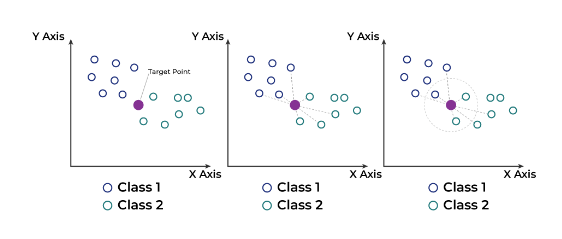
KNN

1. Khái niệm

* KNN là thuật toán thuộc miền học có giám sát và có ứng dụng mạnh mẽ trong nhận dạng mẫu
* Giả sử có 2 bộ dữ liệu X1, X2
* Các bộ phân loại lân cận gần nhất dựa trên việc học bằng cách so sánh 1 bộ dữ liệu kiểm tra nhất định với các bộ dữ liệu huấn luận tươn gtuwj với nó. Các bộ dữ liệu huấn luyện được mô tả bởi n thuộc tính. Mỗi bộ dữ liệu đại diện cho 1 điểm trong không gian n chiều.



* Khi được cung cấp một bộ dữ liệu đầu vào, KNN sẽ tìm kiếm trong không gian mẫu để tìm ra bộ dữ liệu huấn luyện gần nhất với bộ dữ liệu chưa biết. K bộ dữ liệu huấn luyện này là k " hàng xóm gần nhất" của bộ dữ liệu chưa biết



1. Khoảng cách

* "Độ gần " của bộ dữ liệu đã được huấn luyện gần nhất với bộ dữ liệu chưa biết được xác định theo thước đo khoảng cách
  + Khoảng cách Euclide giữa 2 điểm hoặc 2 bộ dữ liệu X1 = (x11,x12,…x1n) và X2 = (x21,x22,….,x2n) là:

=

* + Đối với mỗi thuộc tính số, ta lấy chênh lệch giữa các giá trị tương ứng của thuộc tính đó trong X1 và X2, căn bậc 2 tổng khoảng cách. Thông thường sẽ chuẩn hoá giá trị của từng thuộc tính trước khi tính bằng Euclide. Giúp ngăn các thuộc tín có phạm vi ban đầu lớn vượt trội hơn các thuộc tính có phạm vi nhỏ hơn ban đầu. Ví dụ chuẩn hoá tối thiểu – tối đa có thể được sử dụng để chuyển đối giá trị v của thuộc tính số A thành v' trong phạm vi [0,1] bằng cách tính

v ' =

* Đối với KNN, bộ dữ liệu chưa biết được gán cho lớp phổ biến nhất trong số k lân cận nhất của nó. Khi k=1, bộ dữ liệu chưa biết được gán lớp của bộ dữ liệu huấn luyện gần nó nhất trong không gian mẫu. Các bộ phân loại lân cận gần nhất cũng có thể được sử dụng để dự đoán số (trả về 1 dữ đoán có giá trị thực cho bộ dữ liệu chưa biết cho trước). Trong TH này, trình phân loại trả về giá trị trung bình của các nhãn có giá trị thực được liên kết với KNN của bộ dữ liệu chưa biết

? Nếu những dữ liệu không phải là số

* Đối với loại dữ liệu không phải là số mà là danh nghĩa, các dữ liệu loại text (hoặc phân loại) chẳng hạn như màu sắc thì so sánh giá trị tương ứng của thuộc tính trong bộ X1 với giá trị đó trong bộ X2. Nếu cả 2 bộ dữ liệu giống hệt nhau (VD X1 và X2 đều có màu xanh), thì hiệu giữa 2 bộ dữ liệu này được lấy là 0. Nếu 2 bộ dữ liệu này là khác nhau (vd X1 màu xanh và X2 có màu đỏ) thì sự khác biệt được coi là 1.

? Nếu có những giá trị bị thiếu

* Còn những giá trị của thuọc tính A nhất định bị thiếu trong bộ X1 và/hoặc X2 giả định mức chênh lệch tối đa có thể có. Giả sử rằng mỗi thuộc tính đã được ánh xạ với phạm vi [0,1]. Đối với các thuộc tính danh nghĩa, chúng ta lấy giá trị chênh lệch là 1 nếu thiếu 1 hoặc cả 2 giá trị tương ứng của A. Nếu A là số và bị thiếu trong cả 2 bộ X1 và X2 thì hiệu cũng được lấy là 1. Nếu chỉ thiếu 1 giá trị và giá trị kia (gọi là v ' ) không thiếu và đã được chuẩn hoá, thì chúng ta có thể lấy chênh lệch là [1-v '] hoặc [0-v '] tuỳ theo giá trị nào lớn hơn

? Làm thế nào để có thể xác định giá trị phù hợp cho k, số lượng hàng xóm

* Xác định bằng thực nghiệm. Bắt đầu với k =1, sử dụng bộ kiểm tra để ước tính tỷ lệ lỗi của bộ phân loại. Quá trình có thể được lặp lại mỗi lần bằng cách tăng k để cho phép thêm 1 hàng xóm. Giá trị k cho tỷ lệ lỗi tối thiểu có thể được chọn. Nói chung, số lượng bộ dữ liệu huấn luyện càng lớn thì giá trị của k sẽ càng lớn (do đó các quyết định phân loại và dự đoán số có thể dựa trên phần lớn hơn của các bộ dữ liệu được lưu trữ). Khi số lượng bộ dữ liệu huấn luyện tiến tới vô cùng và k=1 thì tỷ lệ lỗi có thể không tệ hơn 2 lần tỷ lệ lỗi Bayes . Nếu k cũng tiến tới vô cùng thì tỷ lệ lỗi sẽ tiến tới tỷ lệ lỗi Bayes. Nếu dữ liệu đầu vào có nhiều ngoại lệ hoặc nhiễu hơn thì giá trị k cao hơn sẽ tốt hơn. Nên chọn giá trị lẻ cho k để tránh bị ràng buộc trong phân loại.

1. Các bước để thực hiện thuật toán KNN

* Bước 1: chọn giá trị tối ưu của k (k đại diện cho số lượng hàng xóm gần nhất cần được xem xét khi đưa ra dự đoán)
* Bước 2: Tính khoảng cách
  + Để đo lường sự tương tự giữa điểm dữ liệu mục tiêu và điểm dữ liệu huấn luyện, khoảng cách Euclide được sử dụng (Khoảng cách được tính giữa từng điểm dữ liệu trong tập dữ liệu và điểm mục tiêu)
* Bước 3: Tìm hàng xóm gần nhất (k điểm dữ liệu có khoảng cách nhỏ nhất tới điểm đích là những điểm lân cận gần nhất)
* Bước 4: Xem xét chọn phân loại hoặc lấy trung bình để hồi quy
  + Trong bài toán phân loại, nhãn lớp được xác định bằng cách biểu quết theo đa số. Lipws có nhiều lần xuất hiện nhất trong số các lớp lân cận sẽ trở thành lớp được sự đoán điểm dữ liệu mục tiêu
  + Trong bài toán hồi quy, nhãn lớp được tính bằng cách lấy trung bình các giá trị đích của k lân cận gần nhất. Giá trị trung bình được tính toán sẽ trở thành đầu ra dự đoán cho điểm dữ liệu mục tiêu

1. Ưu điểm của thuật toán KNN

* Dễ thực hiện vì độ phức tạp của thuật toán không cao
* Dễ thích ứng: Hoạt động của KNN, nó lưu trữ tất cả dữ liệi trong bộ nhớ lưu trữ và do đó, bất cứ khi nào 1 điểm dữ liệu mới được thêm vào thì thuật toán sẽ tự điều chỉnh theo điểm dữ liệu đó và cũng đóng góp cho các dự đoán trong tương lai
* Ít siêu tham số: Tham số duy nhất được yêu cầu trong quá trình đào tạo thuật toán KNN là giá trị của k và lựa chọn khoảng cách

1. Nhược điểm của thuật toán KNN

* Không mở rộng quy mô: Thuật toán đòi hỏi rất nhiều sức mạnh tính toán cũng như khả năng lưu trữ dữ liệu. Điều này làm cho thuật toán tốn thời gian và tốn tài nguyên