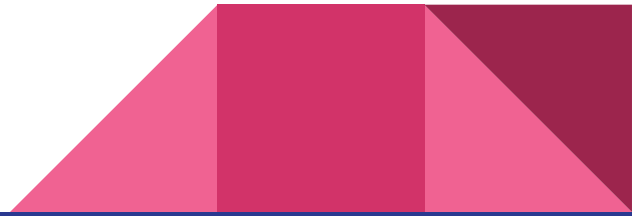
게으른 학습: 최근접 이웃 분류

Machine Learning with R

Contents

- 최근접 이웃(Nearest Neighbor) 분류기
- 게으른 학습자(Lazy learner)
- 거리를 이용한 두 관찰값(example)의 유사도 측정 방법
- k-NN 적용 방법



최근접 이웃 분류의 이해

- 정지 영상 및 동영상에서 광학 글자 인식과 얼굴 인식을 포함하는 컴퓨터 비전 응용
- 영화나 음악 추천에 대한 개인별 선호 예측
- 특정 단백지 및 질병 발견에 사용 가능한 유전자 데이터의 패턴 인식

k-NN(k-Nearest Neighbor) 알고리즘

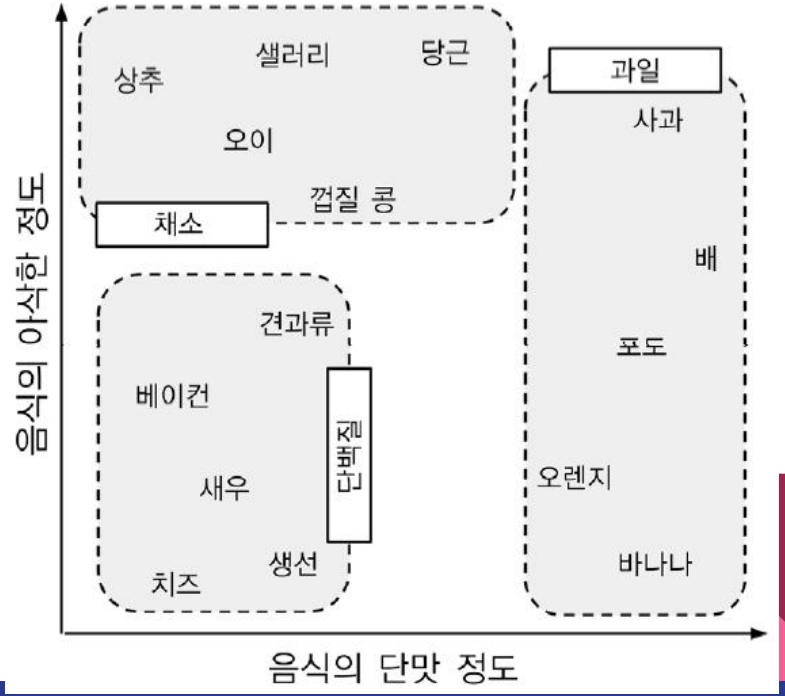
장점	단점
• 단순하고 효율적 • 기저 데이터 분포에 대한 가정을 하지 않음 • 훈련 단계가 빠름	• 모델을 생성하지 않아 특징과 클래스 간의 관계를 이해하는 능력이 제약됨 • 적절한 k의 선택이 필요 • 분류 단계가 느림 • 명목 특징 및 누락 데이터를 위한 추가 처리가 필요

k-NN(k-Nearest Neighbor) 알고리즘

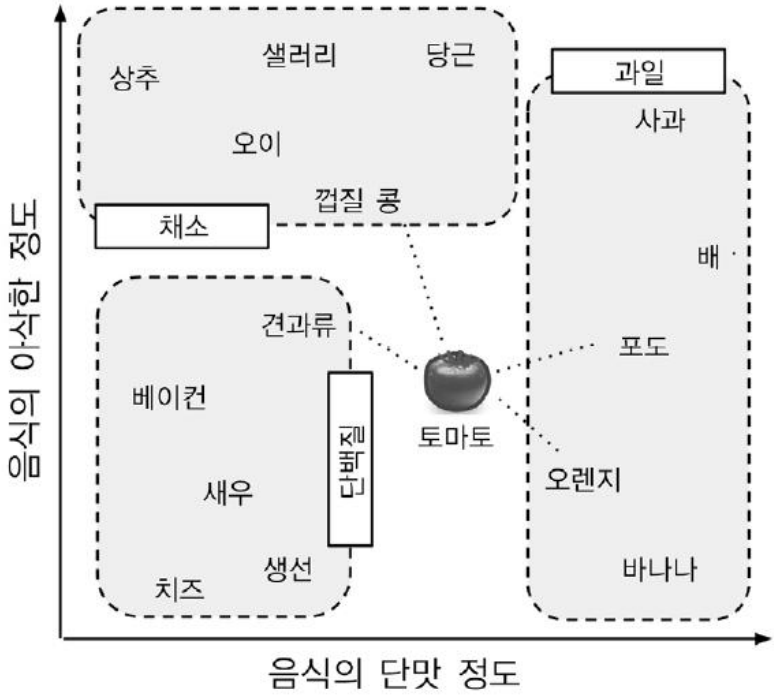
재료	단맛	아삭한 맛	음식 종류
사과	10	9	과일
베이컨	1	4	단백질
바나나	10	1	과일
당근	7	10	채소
샐러리	3	10	채소
치즈	1	1	단백질

k-NN(k-Nearest Neighbor) 알고리즘

- k-NN 알고리즘: 유유상종



k-NN(k-Nearest Neighbor) 알고리즘



- 유클리드 거리(Euclidean distance)

$$\text{dist}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

$$\text{dist}(\text{토마토, 껍질 콩}) = \sqrt{(6-3)^2 + (4-7)^2} = 4.2$$

k-NN(k-Nearest Neighbor) 알고리즘

재료	단맛	아삭한 맛	음식 종류	토마토와의 거리
포도	8	5	과일	sqrt((6 − 8)^2 + (4 − 5)^2) = 2.2
껍질 콩	3	7	채소	sqrt((6 − 3)^2 + (4 − 7)^2) = 4.2
견과	3	6	단백질	sqrt((6 − 3)^2 + (4 − 6)^2) = 3.6
오렌지	7	3	과일	sqrt((6 − 7)^2 + (4 − 3)^2) = 1.4

- 적절한 k 값 선택
  - k = 1: 오렌지 → 토마토는 과일이다.
  - k = 3: 오렌지, 포도, 견과 → 토마토는 과일이다.
- 편향-분산 트레이드오프(bias-variance tradeoff): overfitting과 underfitting 사이의 균형 문제

k-NN 알고리즘을 위한 데이터 준비

- 특징(변수)들마다 단위가 다르기 때문에, 각 특징이 거리 공식에 상대적으로 동일하게 기여할 수 있도록 범위를 줄이거나 늘려줘야 할 필요가 있음.
  - 정규화(Normalization)
  - 표준화(Standardization)
- 최소-최대 정규화(min-max normalization)

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- z-점수 표준화(z-score standardization)

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - \text{Mean}(X)}{\text{StdDev}(X)}$$

k-NN 알고리즘 예제

- 위스콘신 유방암 진단 데이터셋(Wisconsin Breast Cancer Diagnosis Dataset)
  - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
  - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
- 569개의 암 조직검사 관찰값(observation, example), 32개 변수(variable, featrue)
  - Malignant - 악성 종양
  - Benign - 양성 종양