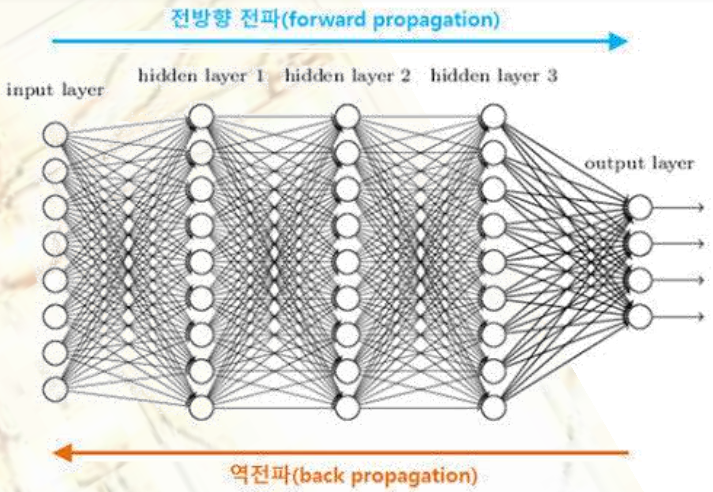
Ch. 12 혼성 모델

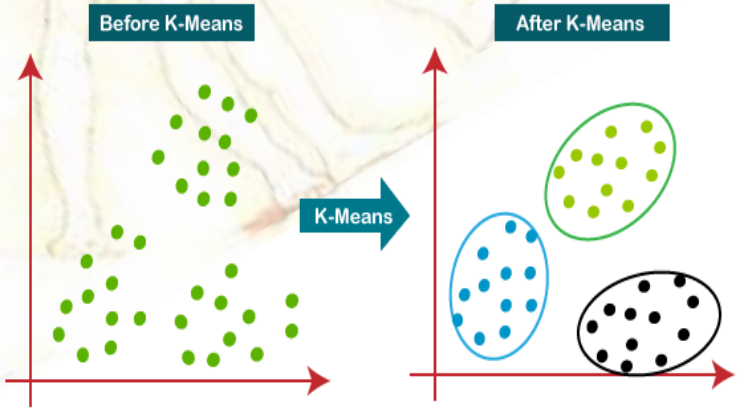
Pattern Recognition & Machine Learning Laboratory
Tae-jin Woo
Jul. 15, 2021

Introduction

- 패턴 인식 문제
  - 문제의 종류
    - 분류, 특징 추출, 군집화 등
  - 알고리즘
    - 신경망, **SVM**, 결정 트리 등
- 기존 알고리즘에 대한 의문점
  - 최선의 알고리즘 선별
    - 특정 상황에서 최선의 알고리즘
      - 충분한 데이터를 통한 실험적 성능 검증 필요
    - 보편적인 최선의 알고리즘
      - 문제 별 성능 우열 관계 상이
- 기존 알고리즘의 문제점
  - 마술 알고리즘의 부재
    - 양적 특징 벡터
      - 신경망, **SVM** 우수
    - 질적 특징 벡터
      - 트리 분류기 우수
- 현실적인 해결책
  - 혼성 모델



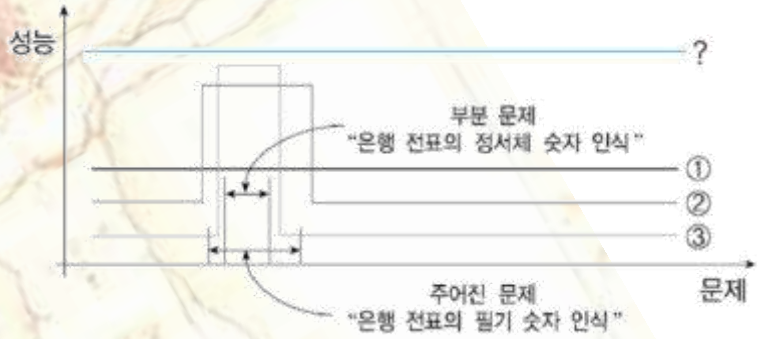
신경망 알고리즘



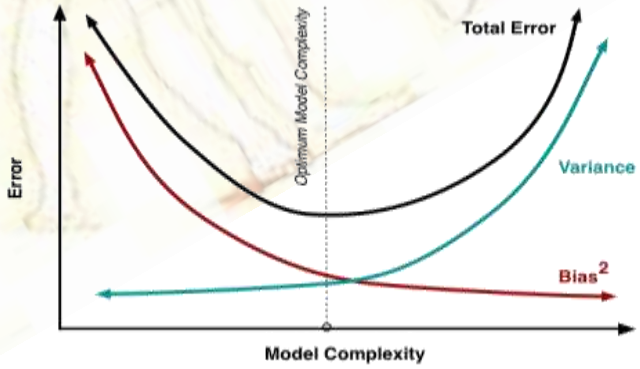
k-means 알고리즘

알고리즘의 성능 특성

- 공학적 관점
  - 목표
    - 특정 문제에 대해 가장 우수한 프로그램 설계
  - 성능 개선 방안
    - 신규 알고리즘 적용
    - 기존 알고리즘 매개 변수 튜닝
    - 충분히 크고 높은 품질의 데이터베이스 확보
  - 특징
    - 고차원 공간 작업으로 실험적 성능 우열 판단 필요
- 공짜 점심 없음
  - 보편적 최선 알고리즘
    - 이론적 불가능 증명 완료
- 바이어스-분산 딜레마
  - 개념
    - 평균 제곱 오차는 바이어스의 제곱과 분산의 합
      - 간단한 모델: 바이어스 증가, 분산 감소
      - 복잡한 모델: 바이어스 감소, 분산 증가
  - 극복 방안
    - 충분히 큰 데이터 및 적절한 모델 선택 중요



분류 알고리즘의 성능 특성

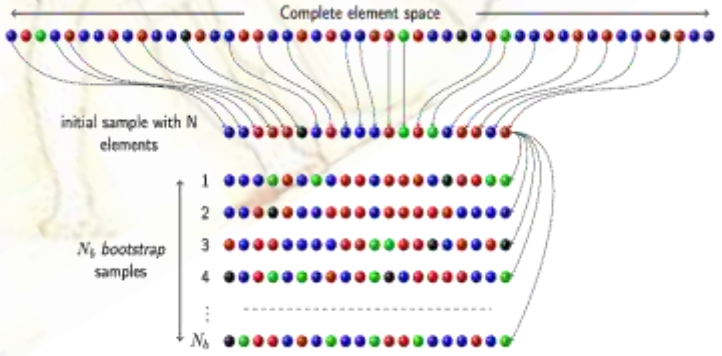


바이어스-분산 딜레마

재 샘플링에 의한 성능 평가

- 배경 및 목적
  - 데이터베이스의 중요성
    - 데이터베이스의 품질에 큰 영향을 미침
    - 검증 집합을 통한 실험적 모델 선택 필요
  - 현실적 문제
    - 대부분 데이터베이스 부족
      - 재 샘플링 기법 적용 필요
- 교차 검증
  - 방법
    - 샘플을 k개의 부분 집합으로 등분
    - 각 부분 집합으로 학습된 k개 분류기의 성능 평균
  - 장점
    - 모든 샘플이 학습에 사용될 수 있음
- 붓스트랩
  - 방법
    - 단순 랜덤 복원 추출 적용
    - 각 부분 집합으로 학습된 T개 분류기의 성능 평균
  - 장점
    - 성능 측정에 대한 통계적 신뢰도 증가

교차 검증

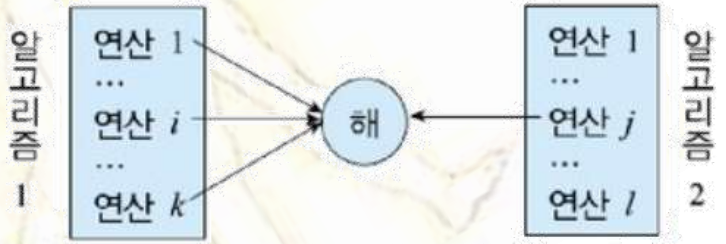


붓스트랩

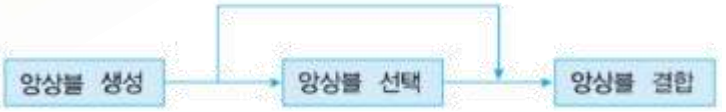
혼성 모델의 발상

- 동기
  - 의사 결정 방식 모방
    - 인간: 여러 전문가의 의견을 종합하여 결론 도출
      - 혼성 모델: 알고리즘의 결합을 통한 성능 개선
- 유형
  - 알고리즘 결합
    - 서로 다른 알고리즘이 구한 해의 결합
    - 대표 명칭: 분류기 앙상블
  - 연산 공유
    - 서로 다른 알고리즘의 협력을 통한 해 도출
- 분류기 앙상블
  - 구성
    - 앙상블 생성, 선택, 결합
  - 장점
    - 나쁜 운 회피 가능
    - 성능 향상 가능
    - 데이터 양 및 질 부족에 따른 어려움 극복 가능
    - 복잡한 결정 경계에 효과적
    - 점진 학습 가능

알고리즘 결합



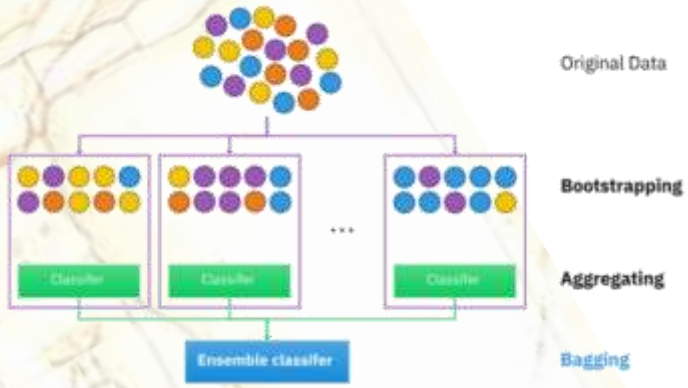
연산 공유



분류기 앙상블 시스템의 구성

앙상블 생성

- 개념
  - 샘플 집합에 대해 복수의 분류기 생성
- 생성 방법
  - 재 샘플링
    - 대표 기법: 배깅, 부스팅
  - 분류기 결합 및 특징 벡터 부분 공간 사용
    - **ex)** 랜덤 포레스트 알고리즘
- 배깅
  - 방법
    - 붓스트랩을 통한 **T**개의 샘플 집합 생성
    - 각 샘플 집합에 대해 **T**개 분류기 독립적 학습
- 부스팅
  - 개념
    - **t**번째 분류기와 **t+1**번째 분류기 사이 연관성 존재
  - 방법
    - **t**번째 분류기에서 틀린 샘플에 가중치 부여
    - 큰 가중치의 샘플에 집중하여 **t+1**번째 분류기 학습
    - 가중치를 고려하여 각 분류기 별 신뢰도 부여 **(AdaBoost)**



배깅

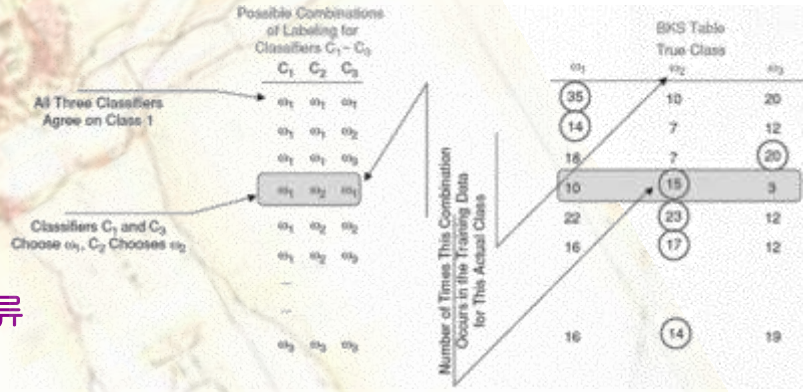
부스팅

앙상블 결합 (1/2)

- 개념
  - 다중 분류기의 출력 결합을 통해 하나의 분류 결과 도출
    - 요소 분류기의 출력 특성에 따라 결합 방식 상이
- 출력 특성
  - 부류 표지
    - 표지 벡터 $\mathbf{L} = (l_1, l_2, \ldots, l_M)^T$ 로 표현
    - 부류에 속하면 1, 속하지 않으면 0
  - 부류 순위
    - 순위 벡터 $\mathbf{R} = (r_1, r_2, \ldots, r_M)^T$ 로 표현
    - 부류에 속할 가능성의 순위를 $[1, M]$ 사이 정수로 표현
    - 부류 표지로 변환 가능
  - 부류 확률
    - 표지 벡터 $\mathbf{P} = (p_1, p_2, \ldots, p_M)^T$ 로 표현
    - 부류에 속할 확률을 $[0, 1]$ 사이 실수로 표현
    - 부류 표지 및 부류 순위로 변환 가능
- 그 외 출력 특성
  - 신경망, **SVM** 등은 부류 별 실수 값 출력
    - 특정 부류에 속할 신뢰도로 해석 가능
    - **Softmax** 함수를 통해 확률로 간주 가능

앙상블 결합 (2/2)

- 부류 표지
  - 다수 투표
    - 최다 득표자 선출 방식
    - $q = \boldsymbol{arg} \max_{j=1,M} \sum_{t=1}^{T} l_{tj}$를 만족하는 ω_q로 분류
  - 가중 다수 투표
    - 신뢰도 고려 최다 득표자 선출
    - $q = \boldsymbol{arg} \max_{j=1,M} \sum_{t=1}^{T} \alpha_t l_{tj}$를 만족하는 ω_q로 분류
  - 행위 지식 공간
    - 요소 분류기의 사전 정보를 통한 참조 표 생성 및 적용
    - 해당 출력 조합에 해당하는 **BKS**표에 가장 큰 값의 부류로 분류
- 부류 순위
  - **Borda** 계수
    - 순위 벡터의 점수 벡터 변환을 통한 순위 분류: $q = \boldsymbol{arg} \max_{j=1,M} \sum_{t=1}^{T} s_{tj}$
- 부류 확률
  - 합, 가중 합
    - 가장 높은 확률 부류 선택: $q = \boldsymbol{arg} \max_{j=1,M} \sum_{t=1}^{T} p_{tj}$ 또는 $q = \boldsymbol{arg} \max_{j=1,M} \sum_{t=1}^{T} \alpha_t p_{tj}$



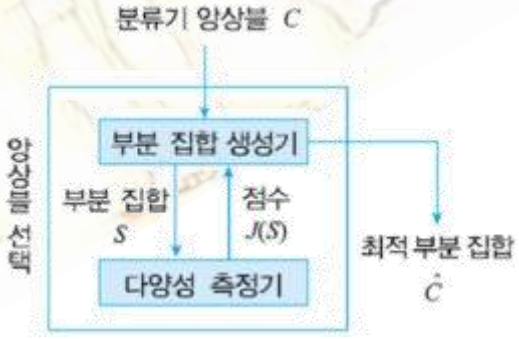
행위 지식 공간

앙상블 선택

- 개념
  - 생성된 앙상블 중 기준에 따라 특정 분류기 선별
    - 선택 기준: 다양성이 클수록 성능 개선에 유리
- 다양성 척도
  - 분류기 쌍 다양성 정의
    - **Q**-통계: 두 분류기의 경향 일치 정도
    - 상관 계수: 두 분류기의 경향 일치 정도
    - 불일치: 두 분류기의 의견이 다른 정도
    - 이중 과실: 두 분류기가 모두 틀리는 정도
  - 분류기 전체 다양성 정의
    - 엔트로피: 얻을 수 있는 정보량의 기댓값 정도
    - **Kohavi-Wolpert** 분산 및 평가자 동의
- 선택 알고리즘
  - 구성
    - 다양성 측정을 통해 최적의 앙상블 선택 가능
  - 혼성 유전 알고리즘
    - 자식 해를 해 집단에 넣기 전에 개선 알고리즘 도입
    - 기존 알고리즘 대비 미세 조정력 향상

	c_k 맞춤	c_k 틀림
c_i 맞춤	n^{11}	n^{10}
c_i 틀림	n^{01}	n^{00}

분류기 쌍 간 관계



앙상블 선택 알고리즘의 구성