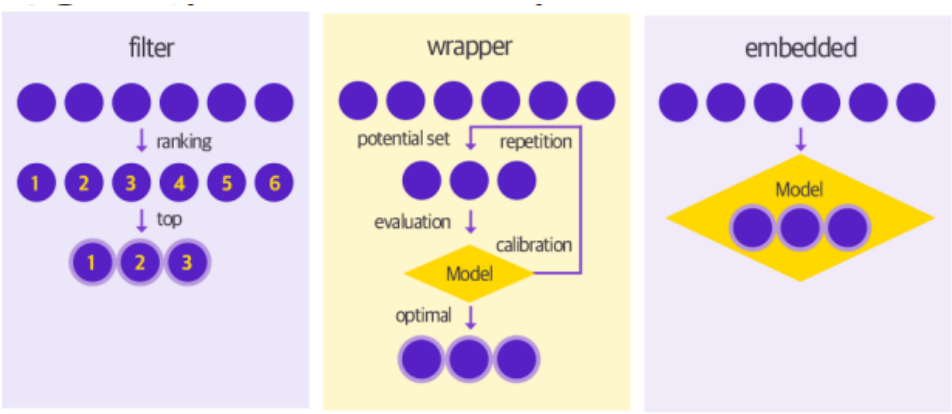
06.개요, 특성 선택 방법론

- 특성 공학의 개요
  - 특성 공학 : 머신러닝 알고리즘에 사용되는 입력데이터에 해당하는 특성 변수들에 대한 처리
  - 특성 선택 : 전체 특성변수 중 최적의 조합을 선택하는 문제
  - 특성 추출 : 특성변수들을 적절하게 조합하여 새로운 특성변수를 만드는 문제
- 특성공간 차원축소의 필요성
  - 모델의 해석력 향상
  - 모델 훈련시간의 단축
  - 차원의 저주방지
  - 과적합(overfitting)에 의한 일반화 오차를 줄여 성능 향상
- 특성 선택(Feature Selection) 방법론
  - 주어진 특성 변수들 가운데 가장 좋은 특성 변수의 조합만 선택
  - 불필요한 특성 변수를 제거
  - Filtering, Wrapper, Embedded 방식으로 분류



  - Filter 방식 : 각 특성변수를 독립적인 평가함수로 평가
    - **각 특성변수 x_i와 목표변수 (Y)** 와의 연관성을 측정한 뒤, 목표변수를 잘 설명할 수 있는 특성변수만을 선택하는 방식.
    - x_i와 Y의 1:1 관계로만 연관성을 판단

        - 연관성 파악을 위해 t-test, chi-square test, information gain 등의 지표가 활용됨

- Wrapper 방식 : 학습 알고리즘을 이용
    - 다양한 특성변수의 조합에 대해 목표변수를 예측하기 위한 알고리즘을 훈련하고,
      cross-validation 등의 방법으로 훈련된 모델의 예측력을 평가.
      그 결과를 비교하여 최적화된 특성변수의 조합을 찾는 방법
    - 특성 변수의 조합이 바뀔때 마다 모델을 학습함
    - 특성변수에 **중복된 정보가 많은 경우** 이를 효과적으로 제거함
    - 대표적인 알고리즘
        - forward selection : 중요한 변수부터 순차적으로 포함하면서 더이상 포함될 중요변수가 없으면 멈추는 방식
        - backward selection : 모두 포함하고 순차적으로 중요하지 않은 변수를 제거해 나가는 방식
        - stepwise selection : 매 스탭을 반복하면서 중요하면 선택하고 중요하지 않으면 제거하는 방법
- Filter 방식과 Wrapper 방식의 차이점
    - Filter는 1:1 로 판단하는 데에 비해 Wrapper방식은 변수들의 조합과 y를 평가 (다대일)
    - Filter는 모델링을 하지않고 기초통계에서 나오는 추론방법들에 기초해서 독립적으로 판단
    - Wrapper는 조합이 주어질때마다 모델에 fitting하여 판단
- Filter 방식과 Wrapper 방식 장단점

	장점	단점
Filter	- 계산비용이 적고 속도가 빠름.	- 특성 변수간의 상호작용을 고려하지 않음.
Wrapper	- 특성변수 간의 상호작용을 고려함. - 주어진 학습 알고리즘에 대해 항상 최적의 특성변수 조합을 찾음.	- 모델을 학습해야 하므로, 계산비용이 크고 속도가 느림. - 과적합(overfitting)의 가능성 있음.

- Embedded 방식 : 학습 알고리즘 자체에 feature selection을 포함하는 경우

- Wrapper 방식은 모든 특성변수 조합에 대한 학습을 마친 결과를 비교하는데 비해, Embedded 방식은 **학습과정에서 최적화된 변수를 선택**한다는 점에서 차이가 있음
- 대표적인 방법으로는 특성변수에 규제를 가하는 방식인 Ridge, Lasso, Elastic net