MathWorks®

지도 학습 적용

지도학습을 고려해야 하는 경우

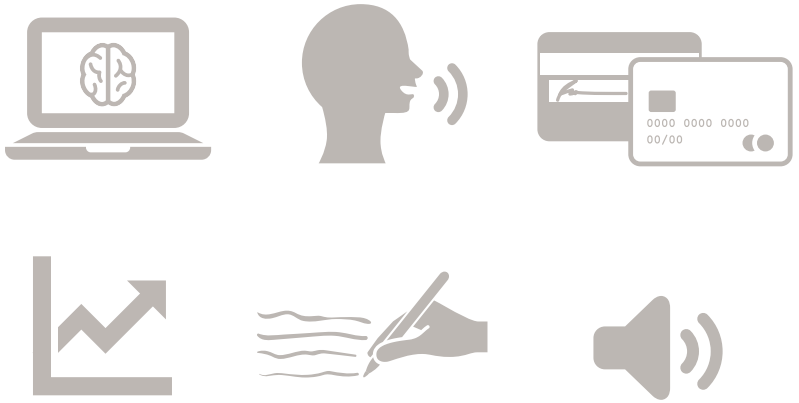
지도학습 알고리즘은 알려진 입력 데이터 세트(훈련 세트) 및 알려진 데이터에 대한 응답(출력)을 사용하고 새 입력 데이터에 대한 응답을 위해 합리적인 예측을 생성하도록 모델을 훈련합니다. 예측하려고 하는 출력에 대한 기존 데이터가 있는 경우 지도학습을 사용합니다.

지도학습 기법

모든 지도학습 기법은 분류 또는 회귀의 형태입니다.

분류 기법은 이메일이 진짜 또는 스팸인지 여부, 종양이 작은 크기인지, 중간 크기인지 또는 큰 크기인지 등의 개별 응답을 예측합니다. 분류 모델은 데이터를 범주로 분류하도록 훈련됩니다. 응용 분야에는 의료 이미징, 음성 인식, 신용 평가 등이 있습니다.

회귀 기법은 온도 변화 또는 전기 수요 변동 등의 연속 응답을 예측합니다. 응용 분야에는 주가 예측, 자필 인식, 음향 신호 처리 등이 있습니다.



- 데이터에 태그를 지정하거나 데이터를 분류할 수 있습니까? 데이터를 특정 그룹이나 클래스로 구분할 수 있다면 분류 알고리즘을 사용합니다.
- 데이터 범위를 사용하고 있습니까? 응답의 특성이 온도나 장비 오류 발생까지의 시간 같은 실수이면 회귀 기법을 사용합니다.

적합한 알고리즘 선택

섹션 1에서 살펴본 대로 머신 러닝 알고리즘을 선택하는 것은 시행착오 과정입니다. 또한 다음과 같은 알고리즘의 특정 특성 간 균형을 잡는 일이기도 합니다.

- 훈련 속도
- 메모리 사용량
- 새 데이터에 대한 예측 정확도
- 투명성 또는 해석 가능성(알고리즘에서 예측이 생성되는 이유를 쉽게 이해할 수 있는 정도)

가장 일반적으로 사용되는 분류 및 회귀 알고리즘을 자세히 살펴보겠습니다.

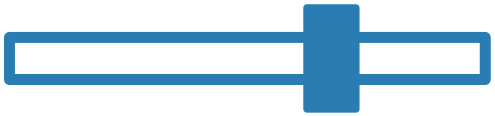
더 큰 훈련 데이터셋을 사용하면 새 데이터에 대해 일반화가 잘 수행되는 모델이 생성되는 경우가 많습니다.

훈련 속도

메모리 사용량

예측 정확도

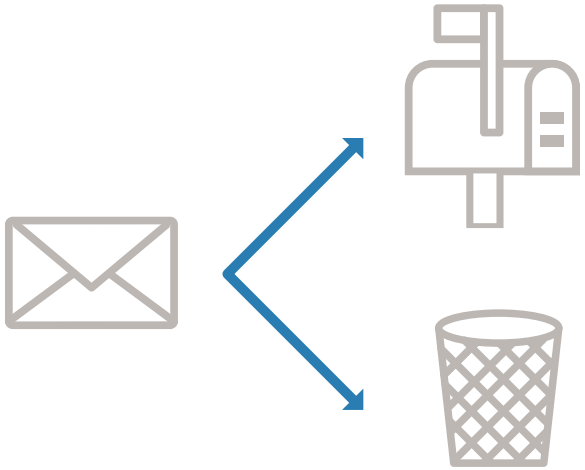
해석 가능성



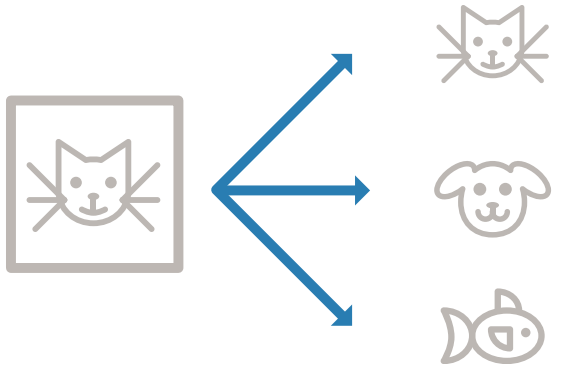
이진 계층 대 다계층 분류

분류 문제를 처리하는 경우 먼저 문제가 이진인지 또는 다계층인지 확인합니다. 이진 분류 문제의 경우 단일 훈련 또는 테스트 항목(인스턴스)을 두 개의 클래스로만 나눌 수 있습니다(예: 이메일이 진짜인지, 스팸인지 여부를 확인하려는 경우). 다계층 분류 문제의 경우 3개 이상으로 나눌 수 있습니다(예: 이미지를 개, 고양이 또는 기타 동물로 분류하도록 모델을 훈련하려는 경우).

다계층 분류 문제에는 더 복잡한 모델이 필요하기 때문에 일반적으로 더 어렵다는 점을 유념하십시오.



로지스틱 회귀 등의 특정 알고리즘은 이진 분류 문제에 맞게 특별히 설계되었습니다. 훈련 중에는 이러한 알고리즘이 다계층 알고리즘보다 더 효율적인 경향이 있습니다.



일반적인 분류 알고리즘

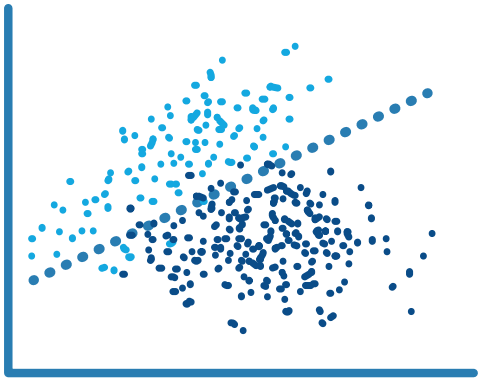
로지스틱 회귀

작동 방식

한 클래스에만 속하는 이진 응답의 확률을 예측할 수 있는 모델을 피팅합니다. 단순성 때문에 로지스틱 회귀는 일반적으로 이진 분류 문제의 시작점으로 사용됩니다.

최적 사용...

- 데이터를 명확히 단일 선형 경계로 구분할 수 있는 경우
- 더 복잡한 분류 방법을 평가하기 위한 기준선으로 사용



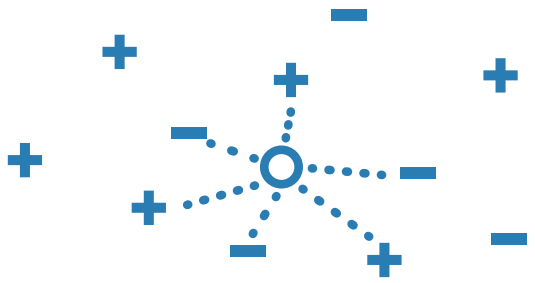
kNN(k-Nearest Neighbor)

작동 방식

kNN은 데이터셋 내에서 최근방(Nearest Neighbor)의 클래스를 기반으로 객체를 범주화합니다. kNN 예측은 서로 가까이 있는 객체가 비슷하다고 가정합니다. 최근방(Nearest Neighbor)을 찾는 데는 유클리드, 도시 구획, 코사인, Chebychev 등의 거리 메트릭이 사용됩니다.

최적 사용...

- 벤치마크 러닝 규칙을 설정하기 위해 단순 알고리즘이 필요한 경우
- 훈련된 모델의 메모리 사용량이 중요한 문제가 아닌 경우
- 훈련된 모델의 예측 속도가 중요한 문제가 아닌 경우

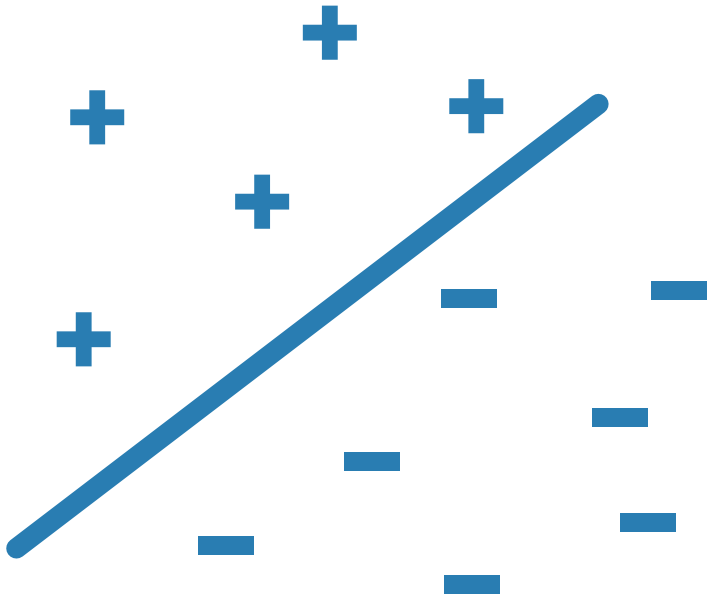


일반적인 분류 알고리즘 *계속*

SVM(서포트 벡터 머신)

작동 방식

한 클래스의 모든 데이터 포인트를 다른 클래스의 데이터 포인트와 구분하는 선형 결정 경계(초평면)을 찾아서 데이터를 분류합니다. SVM에 대한 최적의 초평면은 데이터가 선형적으로 구분 가능한 경우 두 클래스 사이에서 가장 큰 차이를 가진 초평면입니다. 데이터가 선형적으로 구분 가능하지 않은 경우 손실 함수를 사용하여 초평면의 잘못된 쪽에 있는 포인트에 페널티를 적용합니다. 경우에 따라 SVM은 커널 변환을 사용하여 비선형적으로 구분 가능한 데이터를 선형 결정 경계를 찾을 수 있는 상위 차원으로 변환합니다.



최적 사용...

- 클래스가 두 개만 있는 데이터의 경우(오류 정정 출력 코드라는 기법을 통해 다계층 분류에도 사용할 수 있음)
- 비선형적으로 구분 가능한 고차원 데이터의 경우
- 간단하고, 해석하기 쉽고, 정확한 분류기가 필요한 경우

일반적인 분류 알고리즘 *계속*

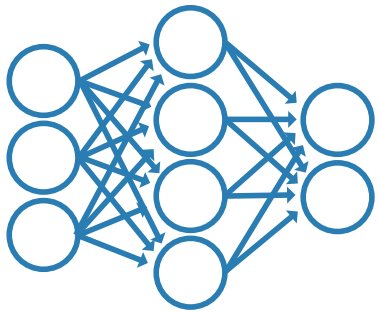
인공신경망

작동 방식

사람의 뇌에서 영감을 받은 뉴럴 네트워크는 연결성이 높은 뉴런 네트워크로 구성되어 입력을 원하는 출력에 연관시킵니다. 네트워크는 제공된 입력이 정확한 응답에 매핑되도록 연결 강도를 반복 수정하는 방식으로 훈련됩니다.

최적 사용...

- 매우 비선형적인 시스템을 모델링하는 경우
- 데이터가 점진적으로 사용 가능해지고 모델을 지속적으로 업데이트하려는 경우
- 입력 데이터에 예기치 않은 변경이 있을 수 있는 경우
- 모델 해석 가능성이 중요한 문제가 아닌 경우



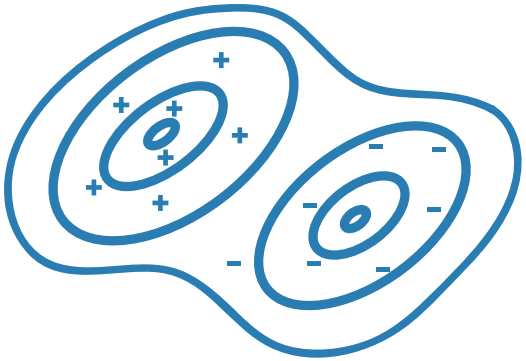
나이브 베이지안

작동 방식

나이브 베이지안 분류기는 클래스에 있는 특정 특징이 다른 특징의 존재에 관련되지 않는다고 가정합니다. 이 분류기는 새 데이터가 특정 클래스에 속할 가장 높은 확률을 기반으로 데이터를 분류합니다.

최적 사용...

- 많은 파라미터가 포함된 작은 데이터셋의 경우
- 해석하기 쉬운 분류기가 필요한 경우
- 금융 및 의료 응용프로그램에서 자주 발생하는 경우처럼 모델에 훈련 데이터에 없던 시나리오가 발생할 경우

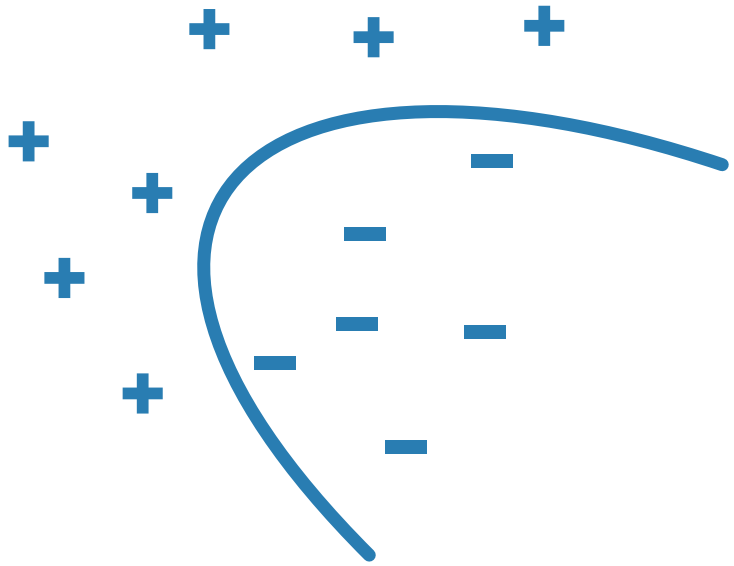


일반적인 분류 알고리즘 *계속*

판별 분석

작동 방식

판별 분석은 특징의 일차 결합을 찾아 데이터를 분류합니다. 판별 분석은 가우시안 분포를 기반으로 여러 클래스에서 데이터를 생성한다고 가정합니다. 판별 분석 모델 훈련에는 각 클래스에 대한 가우시안 분포의 파라미터를 찾는 작업이 포함됩니다. 분포 파라미터는 일차 또는 이차 함수가 될 수 있는 경계를 계산하는 데 사용됩니다. 이러한 경계는 새 데이터의 클래스를 확인하는 데 사용됩니다.



최적 사용...

- 해석하기 쉬운 단순 모델이 필요한 경우
- 훈련 중 메모리 사용량이 중요한 문제인 경우
- 빠르게 예측하는 모델이 필요한 경우

일반적인 분류 알고리즘 *계속*

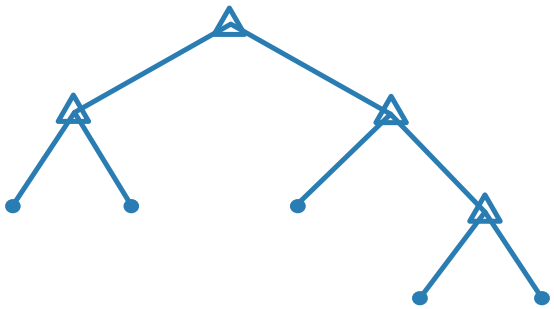
의사결정 트리

작동 방식

의사결정 트리를 사용하면 루트(시작) 에서 아래쪽 리프 노드까지 트리의 의사결정을 따르는 방식으로 데이터에 대한 응답을 예측할 수 있습니다. 트리는 예측자 값이 훈련된 가중치에 비교되는 분기 조건으로 구성됩니다. 분기 수와 가중치 값은 훈련 프로세스에서 결정됩니다. 추가 수정 또는 정리를 통해 모델을 간소화할 수 있습니다.

최적 사용...

- 해석하기 쉽고 빠르게 피팅되는 알고리즘이 필요한 경우
- 메모리 사용량을 최소화하기 위해
- 높은 예측 정확성이 필요하지 않은 경우



배그드(Bagged) 및 부스티드(Boosted) 의사결정 트리

작동 방식

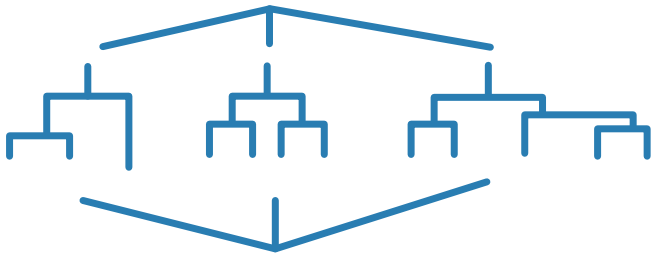
이러한 앙상블 방법에서는 여러 “더 약한” 의사결정 트리가 “더 강한” 앙상블로 결합됩니다.

배그드(Bagged) 의사결정 트리는 입력 데이터에서 부트스트랩(bootstrap) 되는 데이터에서 개별적으로 훈련된 트리로 구성됩니다.

부스팅(Boosting) 에는 반복적으로 “약한” 학습자를 추가하고 잘못 분류된 예제에 초점을 맞추도록 각 약한 학습자의 가중치를 조정하여 강력한 학습자를 생성하는 작업이 포함됩니다.

최적 사용...

- 예측자가 범주형(개별) 이거나 비선형적으로 동작하는 경우
- 모델 훈련에 필요한 시간이 중요한 문제가 아닌 경우



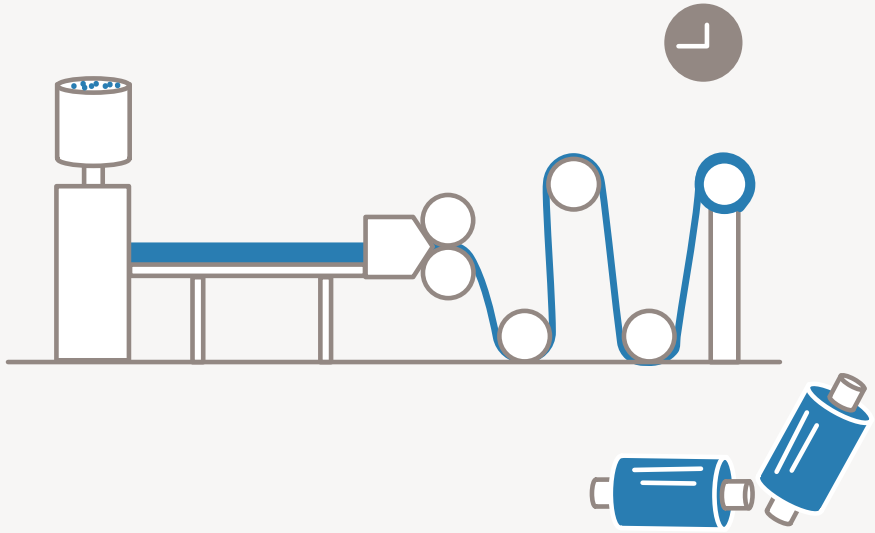
일반적인 분류 알고리즘 *계속*

사례: 제조 장비의 예측 유지관리

한 플라스틱 생산 공장이 연간 약 1천 8백만 톤의 플라스틱 및 박막 제품을 생산합니다. 공장의 작업자 900명은 하루 24시간, 연 365일 작업합니다.

기계 오류를 최소화하고 공장 효율성을 최대화하기 위해 엔지니어는 운영자가 시정 작업을 수행하고 심각한 문제 발생을 방지할 수 있도록 고급 통계와 머신 러닝 알고리즘을 사용하여 잠재적인 문제를 식별하는 상태 모니터링 및 예측 유지관리 응용프로그램을 개발합니다.

공장의 모든 기계에서 데이터를 수집, 정리, 기록한 후 엔지니어는 뉴럴 네트워크, kNN(k-Nearest Neighbor), 배그드(Bagged) 의사결정 트리, SVM(서포트 벡터 머신)을 비롯한 여러 머신 러닝 기법을 평가합니다. 각 기법 적용 시 엔지니어는 기록된 기계 데이터를 사용하여 분류 모델을 훈련하고 나서 모델의 기계 문제 예측 기능을 테스트합니다. 여러 테스트에 의하면 배그드(Bagged) 의사결정 트리의 앙상블이 생산 품질을 예측하기 위한 가장 정확한 모델입니다.



일반적인 회귀 알고리즘

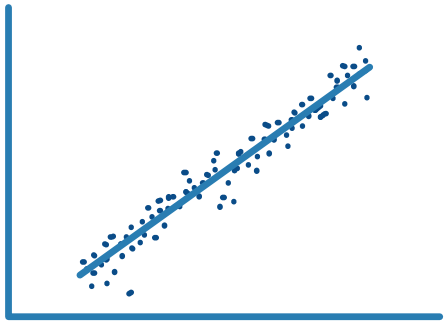
선형 회귀

작동 방식

선형 회귀는 연속 응답 변수를 하나 이상의 예측 변수에 대한 일차 함수로 설명하는 데 사용되는 통계 모델링 기법입니다. 선형 회귀 모델은 간단하게 해석하고 쉽게 훈련할 수 있으므로 보통 새 데이터셋에 피팅되는 첫 번째 모델입니다.

최적 사용...

- 해석하기 쉽고 빠르게 피팅되는 알고리즘이 필요한 경우
- 다른 더 복잡한 회귀 모델을 평가하기 위한 기준선으로 사용



비선형 회귀

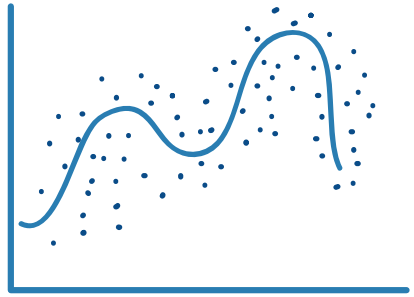
작동 방식

비선형 회귀는 실험 데이터에서 비선형 관계를 설명하도록 도와주는 통계 모델링 기법입니다. 비선형 회귀 모델은 보통 모델이 비선형 방정식으로 설명되는 모수적 모델로 간주됩니다.

"비선형"은 파라미터의 비선형 함수인 피팅 함수를 나타냅니다. 예를 들어 피팅 파라미터는 b_0, b_1 및 b_2입니다. 방정식 $y = b_0+b_1x+b_2x^2$는 피팅 파라미터의 선형 함수이지만, $y = (b_0x^{b1})/(x+b_2)$는 피팅 파라미터의 비선형 함수입니다.

최적 사용...

- 데이터에 강력한 비선형 추세가 있고 선형 공간으로 쉽게 변환할 수 없는 경우
- 사용자 지정 모델을 데이터에 피팅하기 위해



일반적인 회귀 알고리즘 *계속*

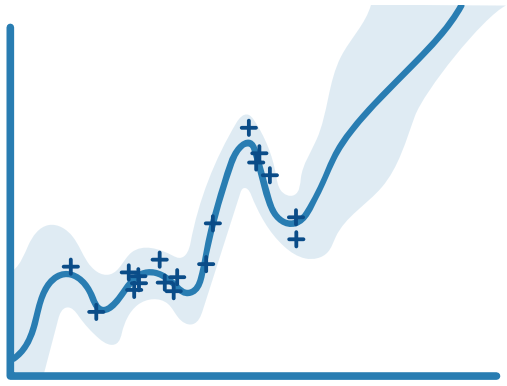
가우시안 프로세스 회귀 모델

작동 방식

GPR(가우시안 프로세스 회귀) 모델은 연속 응답 변수 값을 예측하는 데 사용되는 비모수적 모델입니다. 이러한 모델은 공간 분석 분야에서 불확실성이 있을 경우 보간을 위해 널리 사용됩니다. GPR은 크리깅이라고도 합니다.

최적 사용...

- 지하수 분포에 대한 수리지질학적 데이터 같은 공간 데이터를 보간하기 위해
- 자동차 엔진 같은 복잡한 설계를 쉽게 최적화할 수 있는 대리 모델로 사용



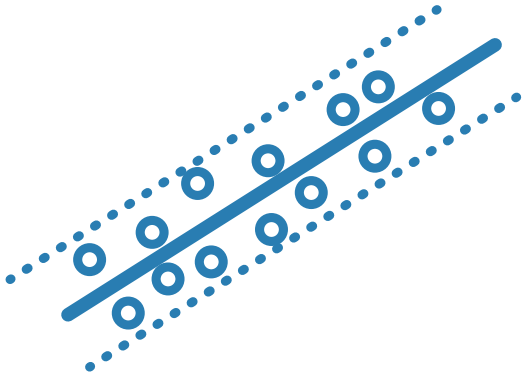
SVM 회귀

작동 방식

SVM 회귀 알고리즘은 SVM 분류 알고리즘처럼 작동하지만, 연속 응답을 예측할 수 있도록 수정됩니다. 데이터를 구분하는 초평면을 찾는 대신, SVM 회귀 알고리즘은 오류에 대한 민감성을 최소화하기 위해 가능한 한 작은 파라미터 값을 사용하여 측정된 데이터에서 작은 값만큼 벗어나는 모델을 찾습니다.

최적 사용...

- 고차원 데이터의 경우(많은 예측 변수가 있음)



일반적인 회귀 알고리즘 *계속*

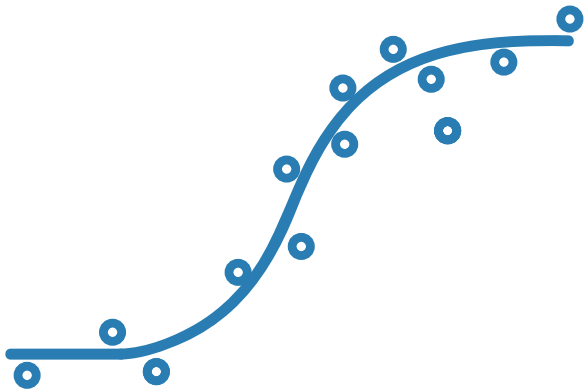
일반화된 선형 모델

작동 방식

일반화된 선형 모델은 비선형 모델에서 선형 방법을 사용하는 특별한 경우입니다. 이 모델에는 입력의 일차 결합을 출력의 비선형 함수(링크 함수)에 피팅하는 작업이 포함됩니다.

최적 사용...

- 항상 양수로 예측되는 응답 변수와 같은 비정규 분포가 응답 변수에 있는 경우



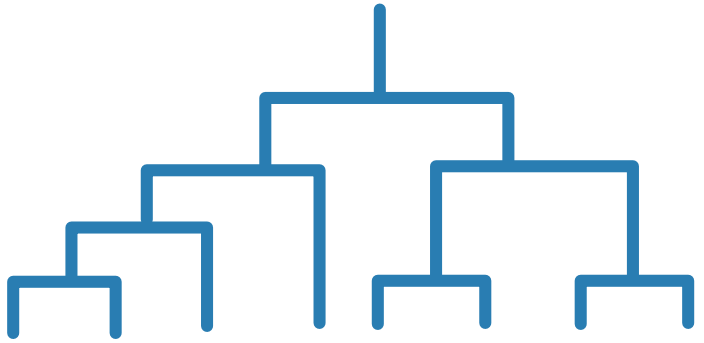
회귀 트리

작동 방식

회귀 의사결정 트리는 분류 의사결정 트리와 비슷하지만, 연속 응답을 예측할 수 있도록 수정됩니다.

최적 사용...

- 예측자가 범주형(개별) 이거나 비선형적으로 동작하는 경우

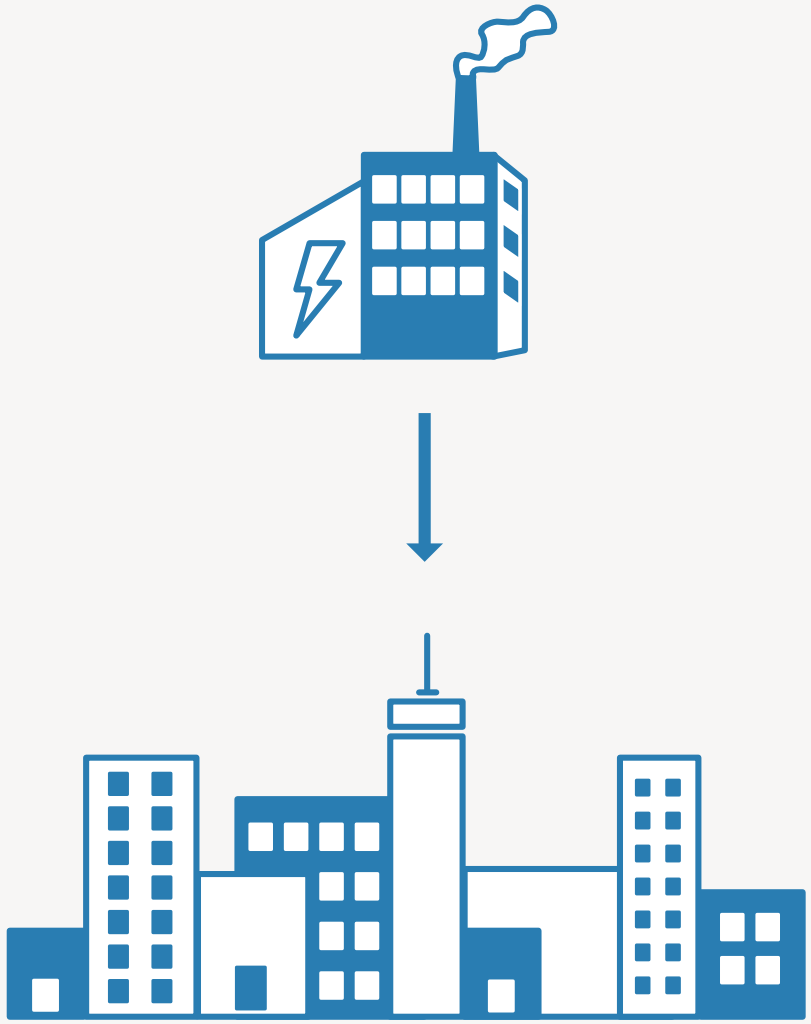


일반적인 회귀 알고리즘 *계속*

사례: 에너지 부하 예측

한 대규모 가스/전기 회사의 가스/전기 분석가가 다음 날의 에너지 수요를 예측하는 모델을 개발했습니다. 이 모델을 사용하여 전력망 운영자는 리소스를 최적화하고 발전소 생산을 예약할 수 있습니다. 각 모델은 전력 소비 및 가격 기록 데이터, 날씨 예측, 그리고 최대 출력, 효율성, 비용, 발전소 긴급 출장에 영향을 미치는 모든 작업 제약 조건을 비롯하여 각 발전소에 대한 파라미터를 확인하기 위해 중앙 데이터베이스에 액세스합니다.

분석가는 테스트 데이터 세트에 대한 낮은 MAPE(절대 백분율 오차 평균)를 제공한 모델을 검색했습니다. 여러 유형의 회귀 모델을 시도한 후 시스템의 비선형 동작을 캡처하는 능력 때문에 뉴럴 네트워크가 가장 낮은 MAPE를 제공한 것이 확인되었습니다.



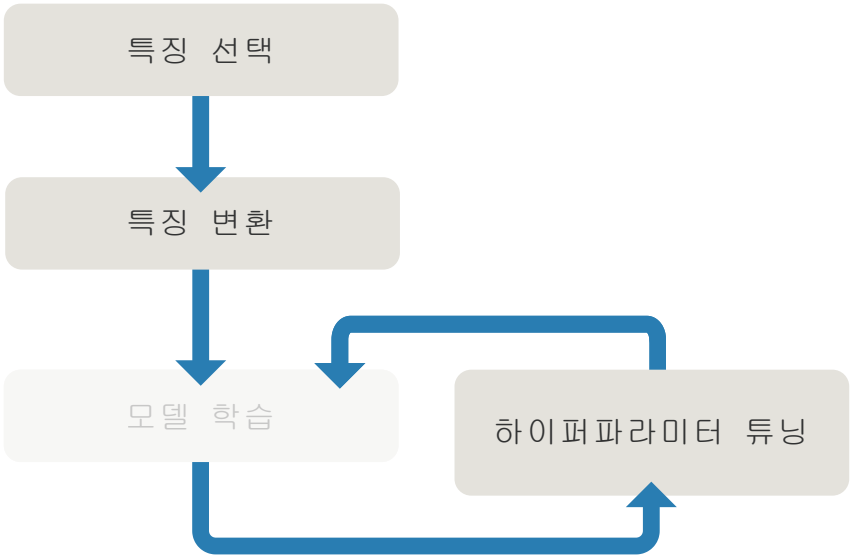
모델 개선

모델 개선은 정확성과 예측력을 높이고 오버피팅을 방지하는 것을 의미합니다(모델이 데이터와 노이즈를 구분할 수 없는 경우). 모델 개선에는 특징 엔지니어링(특징 선택 및 변환) 및 하이퍼파라미터 튜닝이 포함됩니다.

특징 선택: 데이터 모델링 시 최적의 예측력을 제공하는 가장 관련성이 높은 특징 또는 변수를 식별합니다. 이는 모델에 변수를 추가하거나 모델 성능을 개선하지 않는 변수를 제거하는 것을 의미할 수 있습니다.

특징 변환: 주성분 분석, 음이 아닌 행렬 분해, 인자 분석과 같은 기법을 사용하여 기존 특징을 새 특징으로 전환합니다.

하이퍼파라미터 튜닝: 최적의 모델을 제공하는 파라미터 세트를 식별하는 프로세스입니다. 하이퍼파라미터는 머신 러닝 알고리즘이 모델을 데이터에 피팅하는 방법을 제어합니다.



특징 선택

특징 선택은 머신 러닝에서 가장 중요한 작업 중 하나입니다. 고차원 데이터를 처리하거나 데이터셋에 많은 특징과 제한된 수의 관찰이 포함된 경우에 특히 유용합니다. 특징 수를 줄이면 저장 및 계산 시간도 단축되고 결과를 더 쉽게 이해할 수 있습니다.

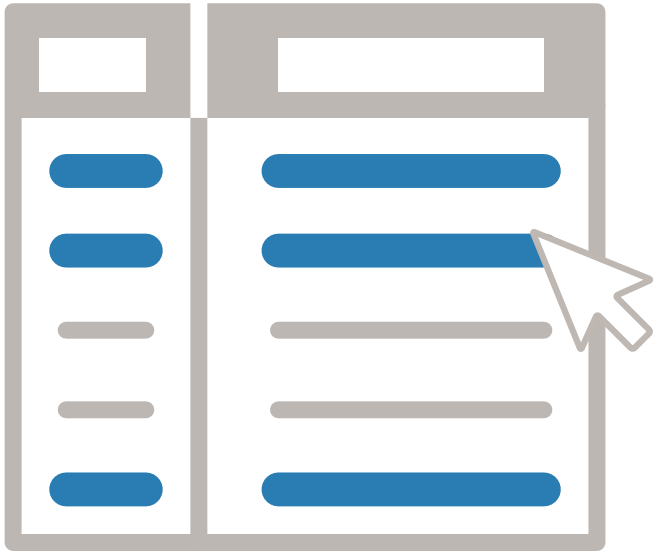
일반적인 특징 선택 기법은 다음과 같습니다.

단계적 회귀: 예측 정확성에 개선이 없을 때까지 순차적으로 특징을 추가하거나 제거합니다.

순차적 특징 선택: 예측 변수를 반복해서 추가하거나 제거하고 각 변경 사항이 모델 성능에 미치는 영향을 평가합니다.

정규화: 가중치(계수)를 0으로 줄이는 방식으로 축소 추정자를 사용하여 중복 특징을 제거합니다.

NCA(Neighborhood Component Analysis): 더 낮은 가중치를 가진 특징이 무시될 수 있도록 출력 예측 시 각 특징에 포함된 가중치를 찾습니다.



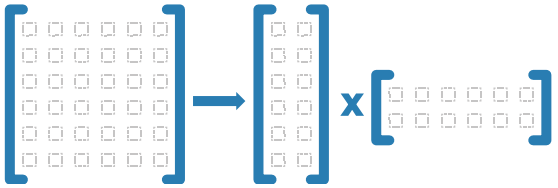
모델은 훈련하기 위해 선택하는 특징과 거의 비슷합니다.

특징 변환

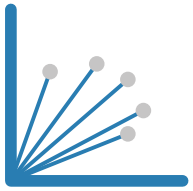
특징 변환은 차원성 감소의 형태입니다. 섹션 3에서 살펴본 대로 가장 일반적으로 사용되는 세 가지 차원성 감소는 다음과 같습니다.

PCA(주성분 분석): 처음 몇 개의 주성분을 통해 고차원 데이터셋에서 대부분의 차이 또는 정보가 캡처되도록 데이터에 대한 선형 변환을 수행합니다. 첫 번째 주성분이 가장 큰 차이를 캡처하고, 이어서 두 번째 주성분이 그 다음 차이를 캡처합니다.

음수 미포함 행렬 분해: 모델의 성분이 물리량과 같은 음수가 아닐 경우에 사용됩니다.



인자 분석: 데이터셋에서 변수 간 기본 상관관계를 식별하여 더 적은 수의 눈에 띄지 않는 잠재적인 인자 또는 일반적인 인자 측면에서 표현을 제공합니다.

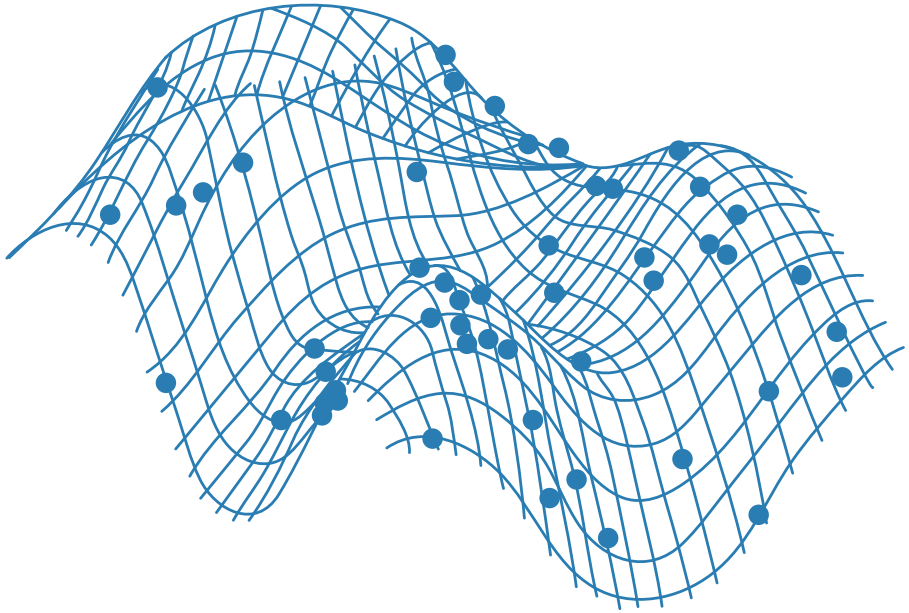


하이퍼파라미터 튜닝

대부분의 머신 러닝 작업처럼 파라미터 튜닝은 반복적인 프로세스입니다. 먼저 결과의 “최적 추측”을 기반으로 파라미터를 설정합니다. 목표는 최적 모델을 생성하는 “최적의 가능한” 값을 찾는 것입니다. 파라미터를 조정하고 모델 성능이 향상됨에 따라 어떤 파라미터 설정이 효과적이고 어떤 파라미터를 계속 튜닝해야 하는지 알 수 있습니다.

세 가지 일반적인 파라미터 튜닝 방법은 다음과 같습니다.

- 베이지안 최적화(Bayesian Optimization)
- 그리드 검색
- 증감 기반 최적화



잘 튜닝된 파라미터를 사용한 단순한 알고리즘은 보통 부적절하게 튜닝된 복잡한 알고리즘보다 더 나은 모델을 생성합니다.

추가 정보

자세히 살펴볼 준비가 되셨습니까? 다음 머신 러닝 방법, 예제, 도구를 살펴보십시오.

지도학습 시작하기

분류화

MATLAB를 활용한 머신 러닝: 분류 시작하기

초급 분류 예제

베이지안 브레인 티저(Bayesian Brain Teaser)

대화형 방식으로 의사결정 트리 살펴보기

서포트 벡터 머신

KNN(k-Nearest Neighbor) 분류

앙상블 분류기 학습

배그드(Bagged) 의사결정 트리를 사용하여 유전자 발현 데이터에서 종양 클래스 예측

회귀

선형 회귀

일반화된 선형 모델이란?

회귀 트리

자동차의 연료 소비율을 예측하도록 회귀 앙상블 학습

특징 선택

고차원 데이터를 분류하기 위한 특징 선택

93114v00