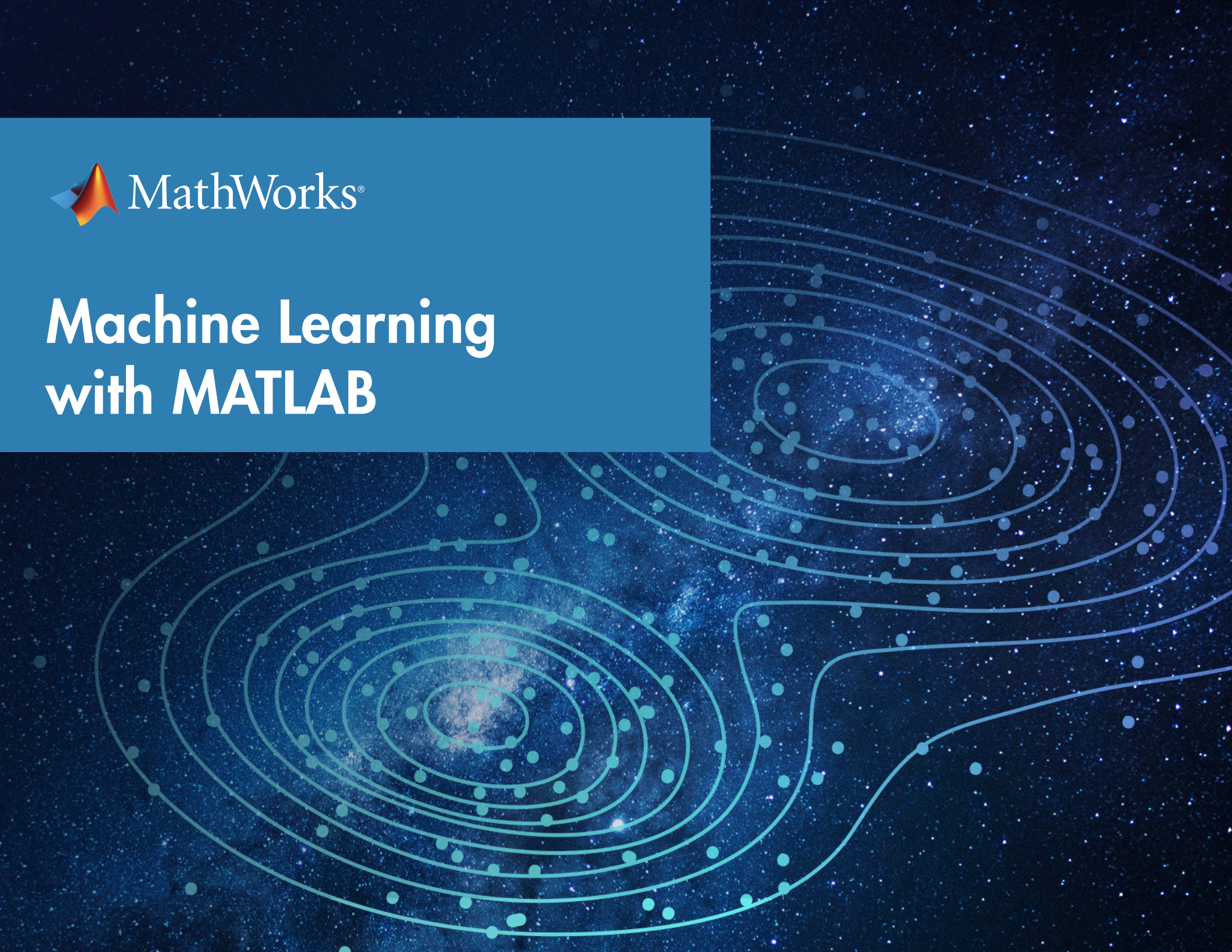

MathWorks®
Machine Learning
with MATLAB

목차

1. 개요
2. 시작하기
3. 비지도 학습 적용
4. 지도 학습 적용


1부: 개요

머신 러닝이란?

머신 러닝은 사람과 동물에게 자연스러운 일, 즉 경험을 통해 학습하는 것을 컴퓨터가 수행할 수 있도록 가르칩니다. 머신 러닝 알고리즘은 미리 결정된 방정식을 모델로 의존하지 않고 계산 방법을 사용하여 데이터에서 직접 정보를 “학습”합니다. 학습할 수 있는 샘플 수가 증가함에 따라 알고리즘 성능이 적절하게 향상됩니다.

더 많은 데이터, 더 많은 질문, 더 나은 답변

머신 러닝 알고리즘은 통찰력을 생성하고 더 나은 의사결정과 예측을 수행하도록 도와주는 자연 패턴을 데이터에서 찾습니다. 자연 패턴은 의료 진단, 주식 거래, 에너지 부하 예측 등에서 중요한 결정을 내리기 위해 매일 사용됩니다. 미디어 사이트는 머신 러닝에 의존하여 수백만 개의 옵션을 살펴보고 노래 또는 동영상 추천을 제공합니다. 소매업체는 머신 러닝을 통해 고객의 구매 행동에 대한 통찰력을 얻습니다.

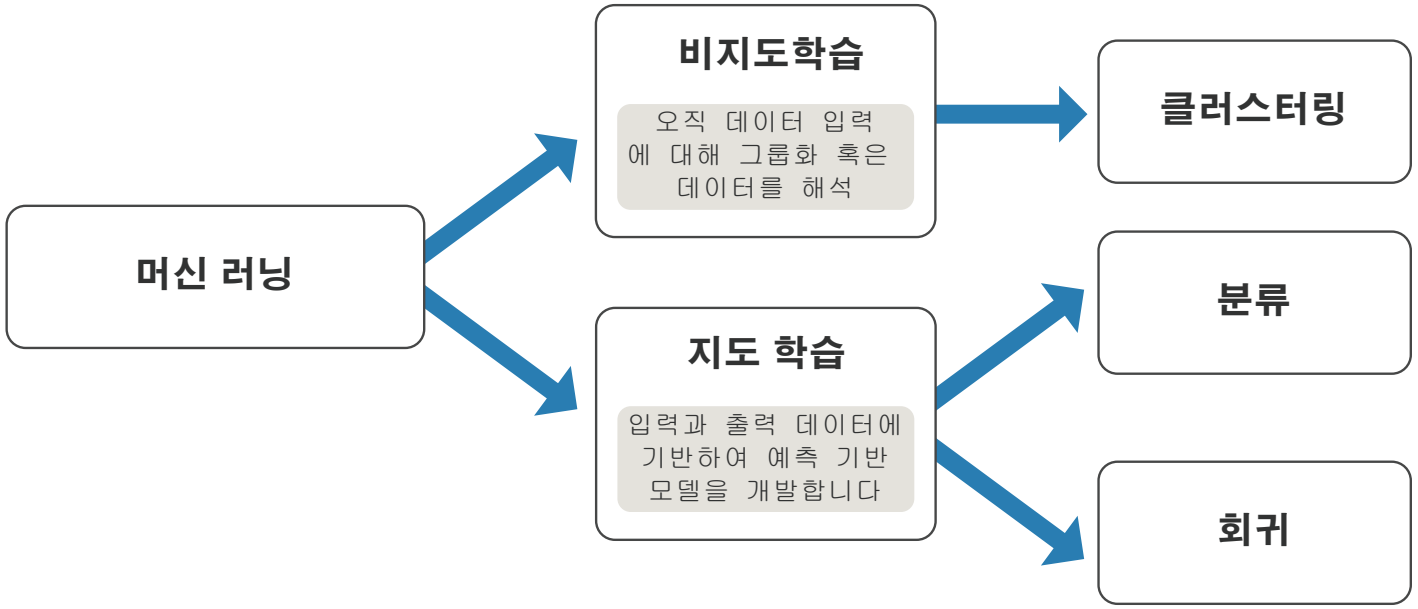
실제 응용 분야

빅데이터가 부상하면서 다음과 같은 영역에서 문제를 해결하는 데 머신 러닝이 특히 중요해졌습니다.

- 계산 금융 - 신용 평가 및 알고리즘 트레이딩
- 이미지 프로세싱 및 컴퓨터 비전 - 얼굴 인식, 얼굴 인식, 동작 인식, 객체 인식
- 계산 생명공학 - 종양 감지, 신약 발견, DNA 염기 서열 (DNA Sequence)
- 에너지 생산 - 가격 및 부하 예측
- 자동차, 항공 및 제조 - 예측 기반 유지 보수 시스템 개발
- 자연어 처리

머신 러닝의 작동 방식

머신 러닝은 두 가지 유형의 기법을 사용합니다. 즉, 지도 학습 기법은 미래 출력을 예측할 수 있도록 알려진 입력 및 출력 데이터를 기반으로 모델을 훈련하고, 비지도 학습 기법은 입력 데이터에서 숨겨진 패턴이나 고유 구조체를 찾습니다.



지도 학습

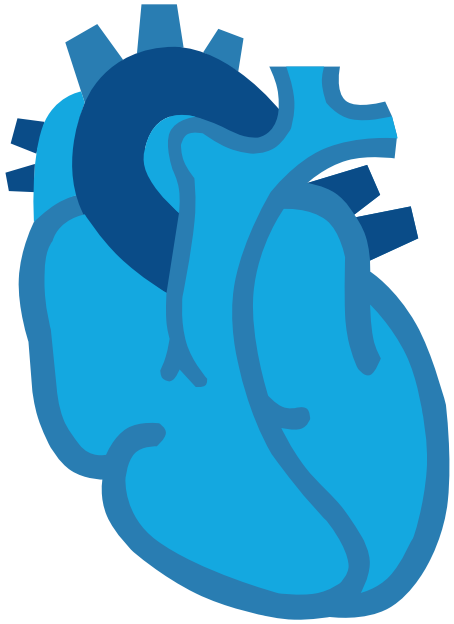
지도 학습의 목적은 불확실성이 있을 때 증거를 기반으로 예측을 하는 모델을 작성하는 것입니다. 지도 학습 알고리즘은 알려진 입력 데이터셋 및 해당 데이터에 대한 알려진 출력을 사용하고 새 데이터에 대한 응답을 위해 합리적인 예측을 생성하도록 모델을 훈련합니다.

지도 학습 분류 및 회귀 기법을 사용하여 예측 모델을 개발합니다.

- **분류 기법은** 이메일이 진짜 또는 스팸인지 여부, 종양이 악성 또는 양성인지 여부 등의 개별 응답을 예측합니다. 분류 모델은 입력 데이터를 범주로 분류합니다. 일반적인 응용 분야에는 의료 이미징, 음성 인식, 신용 평가 등이 있습니다.
- **회귀 기법은** 온도 변화 또는 전력 수요 변동 등의 연속 응답을 예측합니다. 일반적인 응용 분야에는 전기 부하 예측, 알고리즘 트레이딩 등이 있습니다.

지도 학습을 사용하여 심장마비 예측

임상의들이 누군가 1년 내에 심장마비를 일으킬지 여부를 예측하고자 한다고 가정해 보십시오. 임상의들은 연령, 체중, 키, 혈압을 비롯하여 이전 환자들에 대한 데이터를 보유하고 있으며, 이전 환자들에게 1년 내에 심장마비가 나타났는지 여부를 알고 있습니다. 따라서 문제는 기존 데이터를 새로운 사람이 1년 내에 심장마비를 일으킬지 예측할 수 있는 모델에 결합하는 것입니다.

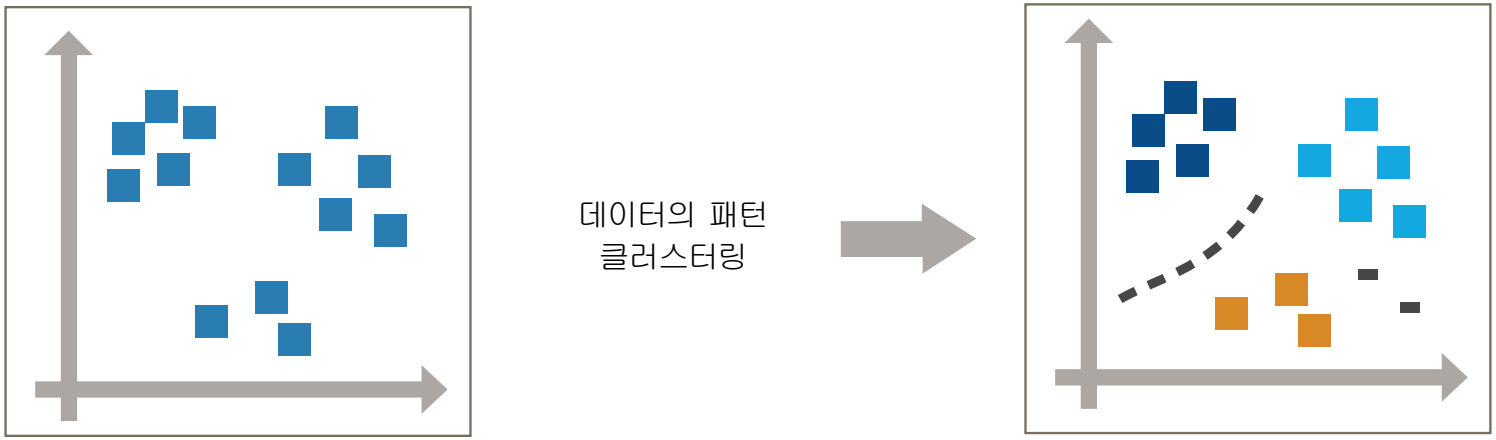


비지도 학습

비지도학습은 데이터에서 숨겨진 패턴이나 고유 구조체를 찾습니다. 그러한 패턴이나 구조체는 분류된 응답 없이 입력 데이터로 구성된 데이터셋에서 추론하는 데 사용됩니다.

클러스터링은 가장 일반적인 비지도학습 기법입니다. 이 기법은 탐색적 데이터 분석을 통해 데이터에서 숨겨진 패턴이나 그룹을 찾는 데 사용됩니다.

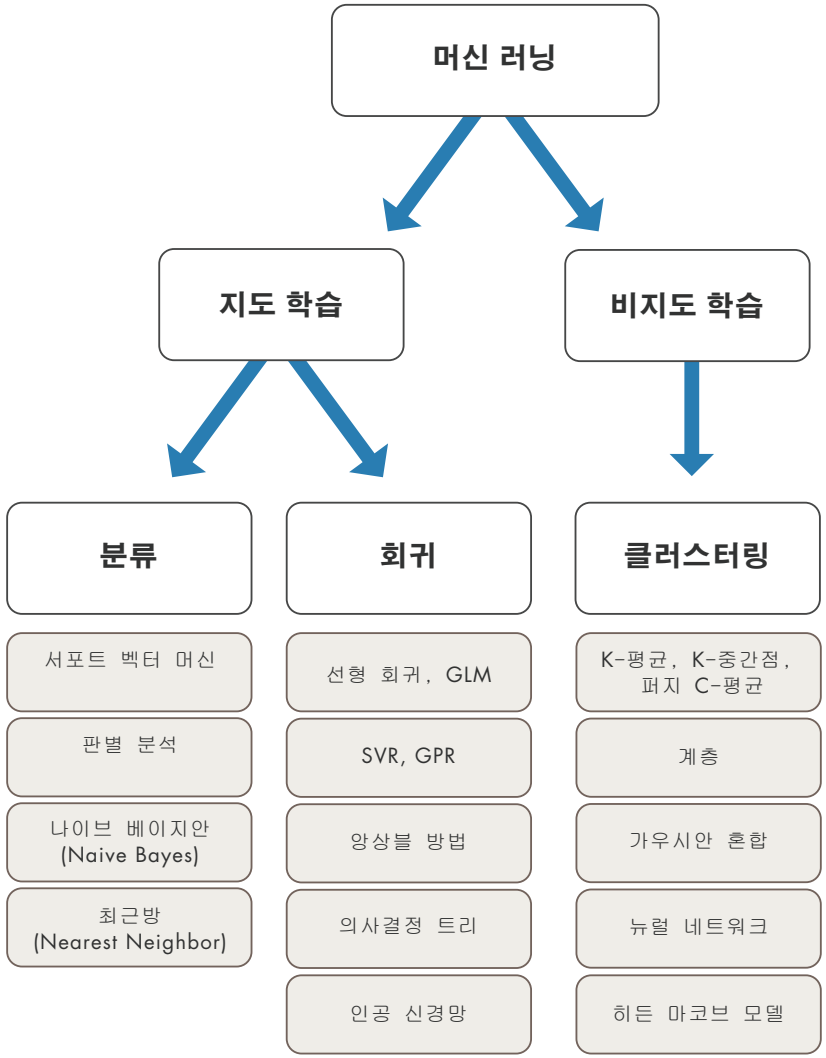
클러스터링 응용 분야에는 유전자 서열 분석, 시장 조사, 객체 인식 등이 있습니다.



어떤 알고리즘을 사용할지 어떻게 결정합니까?

수십 개의 지도 학습 및 비지도학습 알고리즘 알고리즘이 있고 각 알고리즘에는 다양한 학습 접근법이 사용되기 때문에 적합한 알고리즘을 선택하는 일은 매우 어려운 일처럼 보일 수 있습니다.

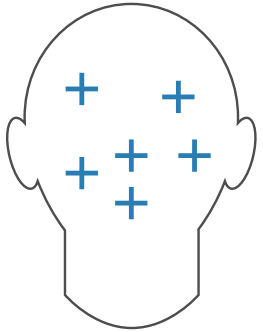
최상의 방법이나 모든 상황에 맞는 알고리즘은 없습니다. 적합한 알고리즘을 찾는 것은 어느 정도는 시행착오 과정이라 할 수 있습니다. 경험이 많은 데이터 과학자조차도 시도해보지 않으면 알고리즘이 적합한지 여부를 알 수가 없습니다. 하지만 알고리즘 선택은 작업 중인 데이터의 크기와 유형, 데이터에서 얻으려는 통찰력, 이 통찰력을 사용하는 방식에 따라서도 달라집니다.



머신 러닝을 언제 사용해야 합니까?

대용량 데이터와 많은 변수가 관련되어 있지만 기존 공식이나 방정식이 없는 복잡한 작업이나 문제가 있을 경우 머신 러닝을 사용해 보십시오. 예를 들어 다음과 같은 상황을 처리해야 할 경우 머신 러닝이 좋은 옵션입니다.

- 얼굴 인식, 음성 인식의 경우처럼 직접 작성하는 규칙과 방정식이 너무 복잡한 상황
- 거래 기록에서 사기를 감지하는 경우처럼 작업 규칙이 지속적으로 바뀌는 상황
- 자동화된 트레이딩, 에너지 수요 예측, 쇼핑 추세 예측의 경우처럼 데이터 특징이 계속 바뀌고 프로그램을 조정해야 하는 상황



고객 활용 사례

예술 작품을 분석할 수 있는 알고리즘 작성

러트거스 대학교의 예술 및 인공 지능 실험실 연구원들은 컴퓨터 알고리즘이 사람처럼 쉽게 그림을 스타일, 장르, 예술가별로 분류할 수 있는지 확인하고자 했습니다. 연구원들은 먼저 그림 스타일을 분류하기 위해 시각적 특징을 식별했습니다. 개발된 알고리즘은 비전문가인 일반인을 능가하여 60%의 정확도로 데이터베이스에서 그림 스타일을 분류했습니다.

연구원들은 스타일 분류에 유용한 시각적 특징(지도학습)이 예술적 영향(비지도학습)을 확인하는 데 사용될 수도 있다는 가설을 세웠습니다.

그들은 Google 이미지를 기반으로 훈련된 분류 알고리즘을 사용하여 특정 객체를 식별했으며, 550년의 기간에 걸쳐 서로 다른 예술가 66명이 그린 그림 1,700점 이상을 대상으로 알고리즘을 테스트했습니다. 이 알고리즘은 디에고 벨라스케스의 “교황 이노센트 10세의 초상”이 프랜시스 베이컨의 “교황 이노켄티우스 10세의 초상화 연구”에 미치는 영향을 비롯하여 관련 작품을 쉽게 식별했습니다.

고객 활용 사례

대형 건물의 HVAC 에너지 사용 최적화

사무실 건물, 병원, 기타 대형 상업 건물의 HVAC(난방, 환기 및 냉방) 시스템은 변화하는 날씨 패턴, 가변적인 에너지 비용 또는 건물의 열 특성을 고려하지 않기 때문에 비효율적인 경우가 많습니다.

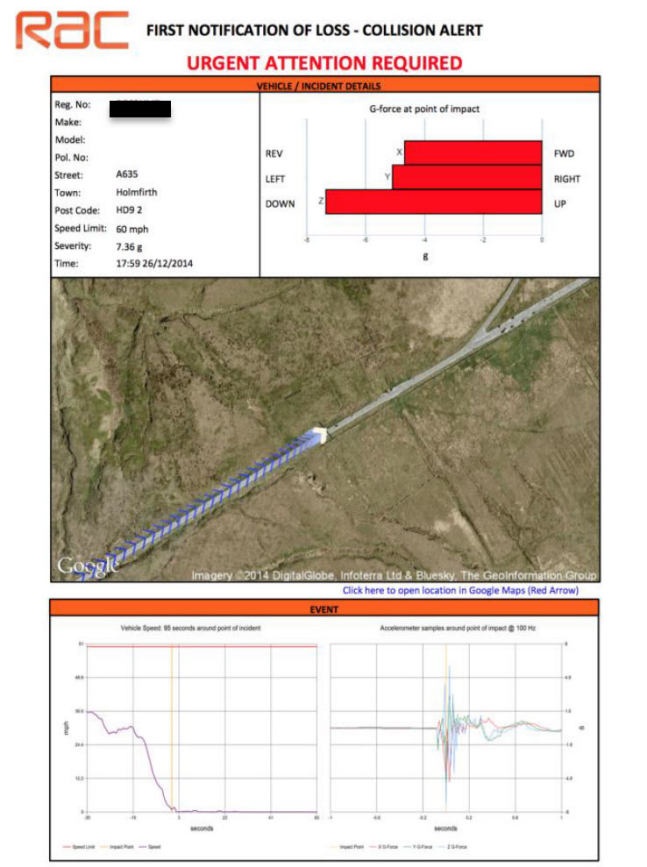
Building IQ의 클라우드 기반 소프트웨어 플랫폼은 이러한 문제를 해결합니다. 이 플랫폼은 고급 알고리즘과 머신 러닝 방법을 사용하여 전력계, 온도계, HVAC 압력 센서에서 수집된 수 기가바이트의 정보와 날씨 및 에너지 비용을 지속적으로 처리합니다. 특히, 머신 러닝을 사용하여 데이터를 세분화하고 난방 및 냉방 프로세스에 대한 가스, 전기, 증기 및 태양열의 상대적 기여도를 결정합니다. Building IQ 플랫폼은 정상 운영 시 대형 상업 건물의 HVAC 에너지 소비량을 10%~25% 절감합니다.

고객 활용 사례

저속 차량 충돌 감지

8백 만 이상의 회원을 보유한 RAC는 영국의 가장 큰 자동차 단체 중 하나로서, 개인/비즈니스 운전자에게 긴급출동 서비스, 보험 및 기타 서비스를 제공합니다.

도로 사고에 신속히 대응하고, 추돌 사고를 줄이고, 보험비를 낮추기 위해 RAC는 고급 머신 러닝 알고리즘을 사용하여 저속 충돌을 감지하고 과속 방지턱이나 포트홀 운행과 같이 더 일반적인 운전 이벤트와 이러한 이벤트를 구분하는 온보드 충돌 감지 시스템을 개발했습니다. 독립적인 테스트에서 RAC 시스템은 테스트 충돌 감지 시 92% 정확도를 나타냈습니다.



추가 정보

자세히 살펴볼 준비가 되셨습니까? 다음 리소스를 통해 머신 러닝 방법, 예제, 도구를 자세히 알아보십시오.

시청 자료

머신 러닝으로 간편하게 34:34

센서 데이터 분석을 위한 신호 처리 및 머신 러닝 기법 42:45

읽기 자료

머신 러닝 블로그 게시물: 소셜 네트워크 분석, 텍스트 마이닝, 베이지안 추론 등

머신 러닝 과제: 최상의 모델 선택과 과적합 방지

살펴보기

MATLAB 머신 러닝 예제

머신 러닝 솔루션

Classification Learner 앱을 사용한 데이터 분류

2부: 시작하기

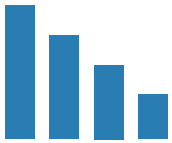
한번에 수행되지 않는 머신 러닝

머신 러닝에서는 시작부터 완료까지 한번에 수행되는 경우는 거의 없으며, 계속해서 새로운 방법을 반복적으로 시도하게 됩니다. 이 장에서는 체계적인 머신러닝 워크플로우를 설명하고 각 주요 작업에 대해 살펴봅니다.

머신 러닝 과제

대부분의 머신 러닝 과제는 데이터 처리 및 적합한 모델 찾기에 관련됩니다.

데이터의 형태와 크기는 다양합니다. 실제 데이터셋은 정리되지 않고, 불완전하며, 형식이 다양할 수 있습니다. 간단한 숫자 데이터만 있을 수도 있지만 경우에 따라 센서 신호, 텍스트, 카메라의 이미지 스트리밍 같은 여러 데이터 유형을 결합할 수도 있습니다.



데이터 전처리에는 전문 지식과 도구가 필요할 수 있습니다. 예를 들어 객체 감지 알고리즘을 훈련할 특징을 선택하려면 이미지 처리에 대한 전문 지식이 필요합니다. 데이터 유형에 따라 다른 전처리 접근법이 필요합니다.

데이터를 피팅할 최적 모델을 찾는 데는 시간이 걸립니다. 적합한 모델을 선택하는 것은 균형을 잡는 작업입니다. 유연성이 높은 모델은 노이즈일 수 있는 사소한 차이를 모델링하여 데이터를 오버피팅하는 경향이 있습니다. 반면, 단순 모델은 너무 많은 가정을 포함할 수 있습니다. 모델 속도, 정확성 및 복잡성 간에는 항상 상충관계가 존재합니다.

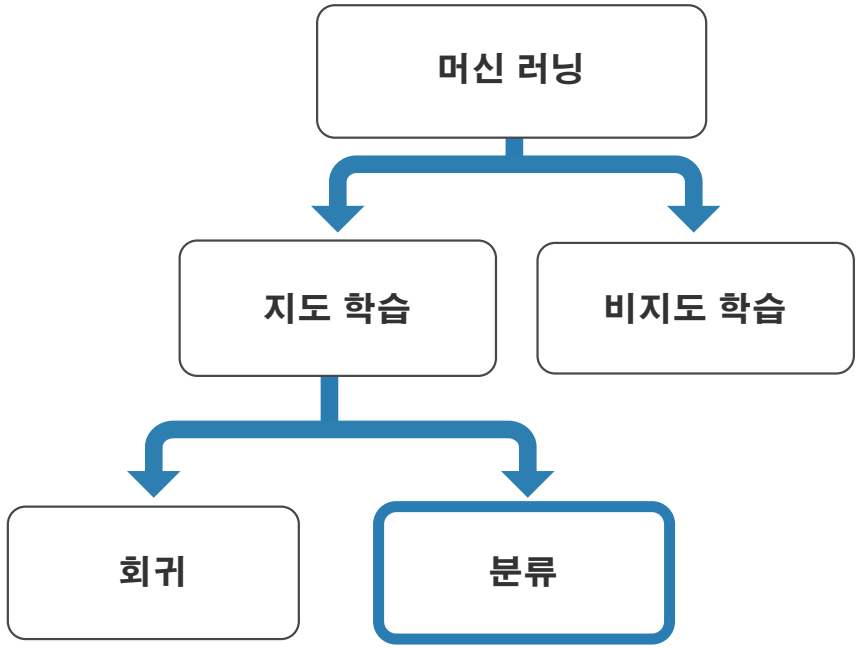
너무 어려운 작업처럼 들리십니까? 낙담하지 마십시오. 시행착오가 머신 러닝의 핵심이라는 점을 기억하십시오. 한 접근법이나 알고리즘이 작동하지 않으면 다른 방법을 시도하면 됩니다. 하지만 체계적인 워크플로가 있다면 순조롭게 출발할 수 있습니다.

시작하기 전에 고려할 질문

모든 머신 러닝 워크플로우는 세 가지 질문으로 시작됩니다.

- 어떤 유형의 데이터를 사용하고 있습니까?
- 데이터에서 어떤 통찰력을 얻고 싶습니까?
- 그러한 통찰력을 어떻게, 어디에 적용하겠습니까?

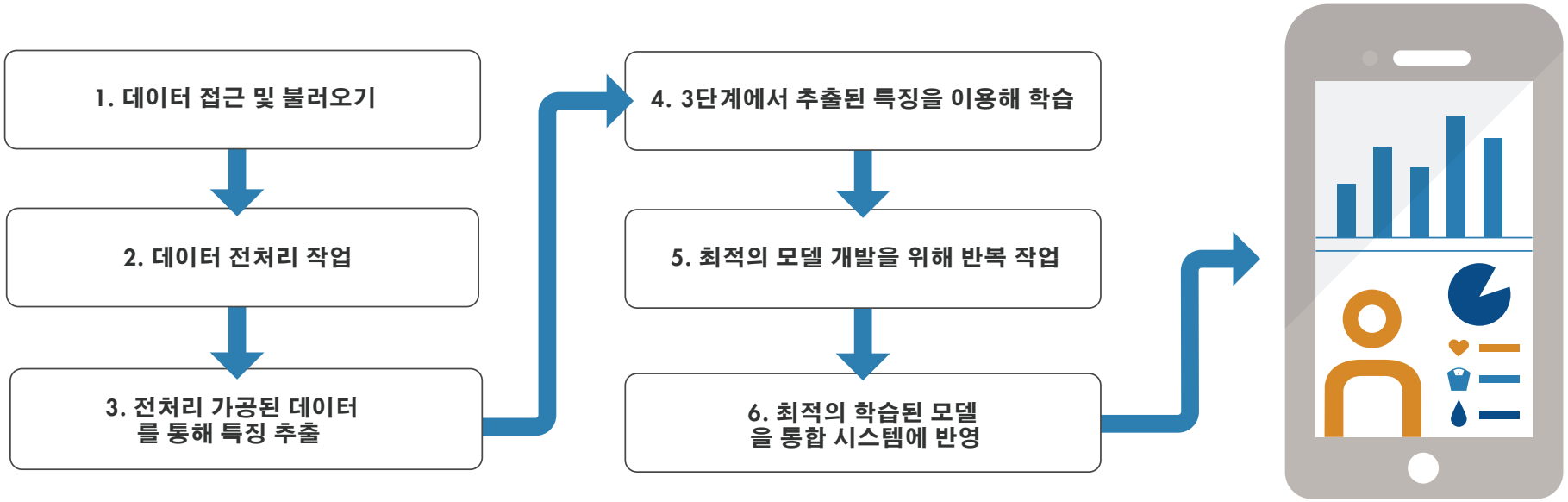
이러한 질문에 대한 답변을 통해 지도 학습을 사용할지 또는 비지도학습을 사용할지 결정할 수 있습니다.



온도나 주가 같은 연속 변수의 미래 가치 등에 대한 예측을 작성하거나 웹캠 비디오 화면에서 자동차 제조업체를 식별하는 등의 분류를 수행하도록 모델을 훈련해야 할 경우 지도 학습을 선택합니다.

데이터를 탐색해야 하고 데이터를 클러스터로 분할하는 것과 같이 좋은 내부 표현을 찾도록 모델을 훈련하려는 경우 비지도학습 선택합니다.

워크플로우 한눈에 보기



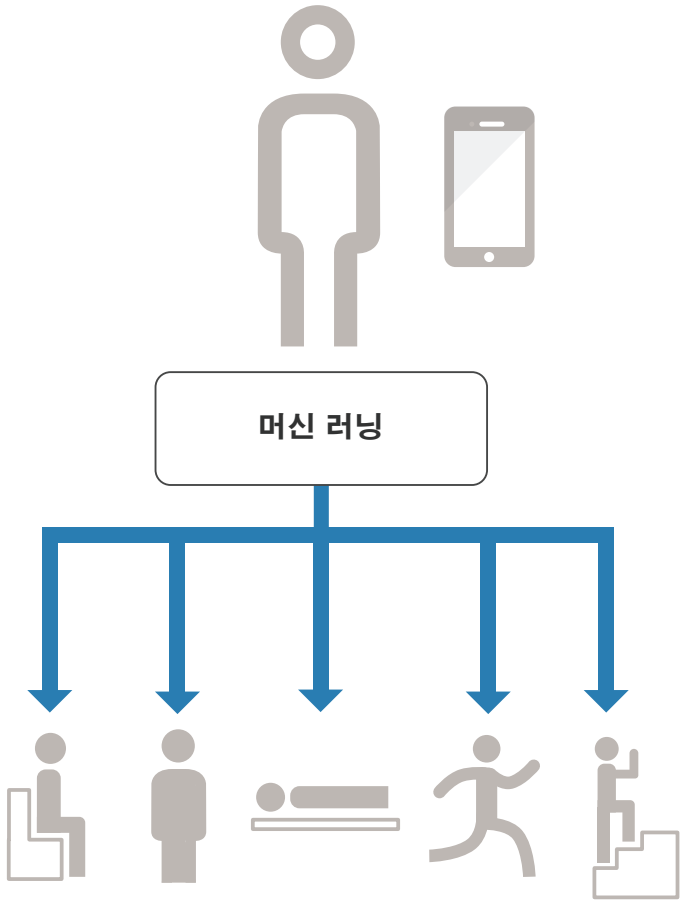
다음 섹션에서는 실례로 건강 상태 모니터링 앱을 사용하여 단계를 자세히 살펴보겠습니다. 전체 워크플로는 MATLAB®에서 완료됩니다.

신체활동 분류기 개발을 위한 학습

이 예제는 휴대 전화의 건강 상태 모니터링 앱 개발을 위한 예제입니다. 입력 데이터는 3차원 위치 및 가속 센서 데이터로 구성됩니다. 출력은 걷기, 서기, 달리기,계단 오르기, 눕기로 구성되었습니다.

입력 데이터를 이용해 신체 활동 분류기 모델 개발을 위해 학습을 수행합니다. 여기서는 분류기 개발이 목적이기 때문에 지도 학습을 적용하겠습니다.

개발된 분류기는 App개발에 활용되어 매일 사용자의 신체 활동을 추적할 수 있습니다.



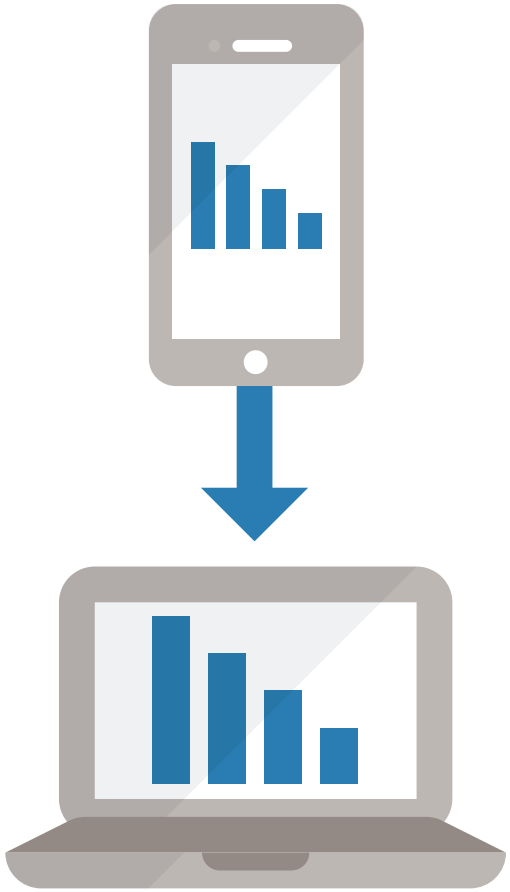
1 1단계: 데이터 로드

위치 및 가속도 센서 데이터를 불러오기 위해 다음을 수행합니다.

1. 전화기를 들고 앉아 전화기의 데이터를 기록하고 "앉기"라는 텍스트 파일에 저장합니다.
2. 전화기를 들고 서서 전화기의 데이터를 기록하고 "서기"라는 두 번째 텍스트 파일에 저장합니다.
3. 분류할 각 활동에 대한 데이터를 수집할 때까지 단계를 반복합니다.

레이블이 지정된 데이터 세트를 텍스트 파일에 저장합니다. 텍스트나 CSV 등의 플랫 파일 형식은 쉽게 작업하고 데이터를 간단히 가져올 수 있습니다.

머신 러닝 알고리즘은 노이즈와 중요한 정보의 차이를 알려줄 만큼 스마트하지 않습니다. 훈련에 데이터를 사용하기 전에 데이터가 정리되고 완전한지 확인해야 합니다.



2 2단계: 데이터 전처리

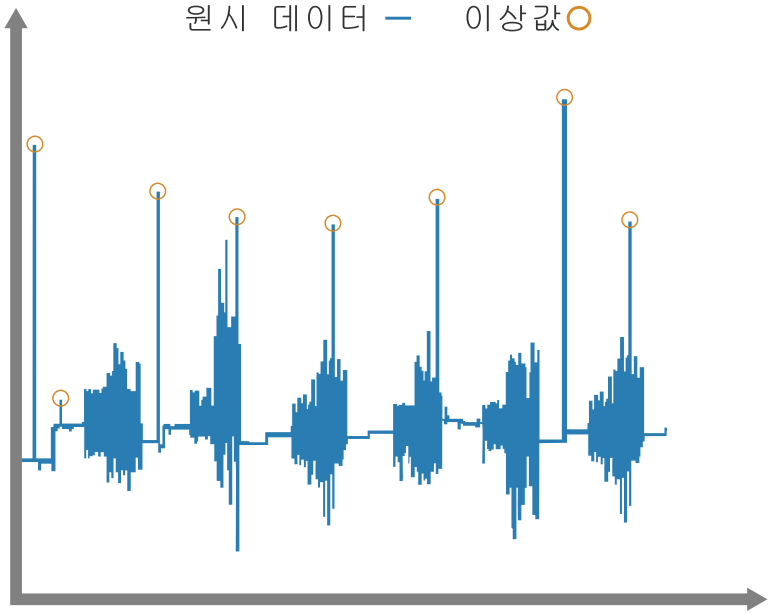
데이터를 MATLAB로 가져오고 각각 레이블이 지정된 세트를 플로팅합니다.
데이터를 전처리하려면 다음을 수행합니다.

1. 나머지 데이터 외부에 있는 이상값 데이터 포인트를 검색합니다.

이상값을 무시할지 여부 또는 이상값이 모델에서 고려해야 할 현상을 나타내는지 여부를 결정해야 합니다. 이 예제에서는 이상값을 무시해도 됩니다(데이터를 기록하는 동안 의도치 않게 이동한 것으로 확인됨).

2. 누락된 값이 있는지 확인합니다(기록 중에 연결이 끊어져 데이터가 손실되었을 수 있기 때문).

누락된 값을 그냥 무시할 수 있지만 이렇게 하면 데이터 세트 크기가 감소합니다. 또는 다른 샘플의 비슷한 데이터를 보간하거나 사용하여 누락된 값을 근삿값으로 대체할 수 있습니다.



활동 추적 데이터의 이상값.

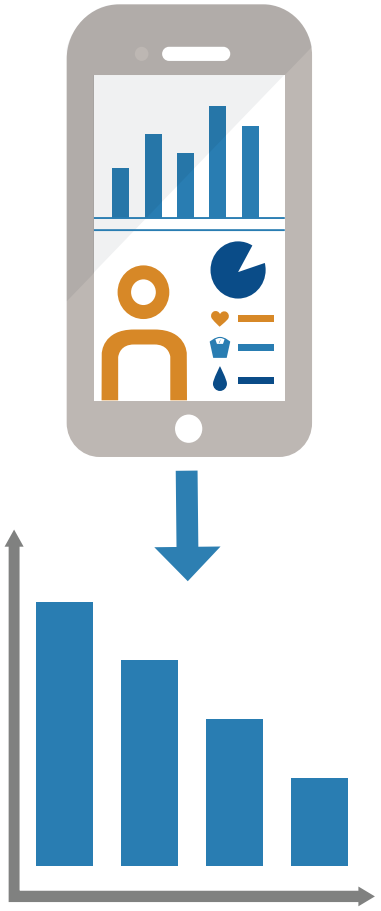
많은 응용프로그램에서 이상값은 중요한 정보를 제공합니다. 예를 들어 신용 카드 위조 감지 앱에서 이상값은 고객의 일상적인 구매 패턴을 벗어나는 구매를 나타냅니다.

2 2단계: 데이터 전처리 *계속*

3. 알고리즘이 전화기의 이동이 아니라 주체의 이동에 초점을 맞추도록 가속도계 데이터에서 중력 효과를 제거합니다. 일반적으로 바이쿼드 필터와 같은 간단한 하이패스 필터가 이를 위해 사용됩니다.

4. 데이터를 두 개의 세트로 나눕니다. 일부 데이터는 테스트용으로 저장하고(테스트 세트) 나머지는 모델 작성에 사용합니다(훈련 세트). 이를 홀드아웃이라고 하며 유용한 교차 유효성 검사 기법입니다.

모델링 프로세스에 사용되지 않은 데이터를 대상으로 모델을 테스트하여 알 수 없는 데이터에서는 모델이 어떻게 작동하는지 확인합니다.



3 3단계: 특징 도출

특징 도출(특징 엔지니어링 또는 특징 추출이라고도 함)은 머신 러닝의 가장 중요한 부분 중 하나입니다. 특징 도출은 원시 데이터를 머신 러닝 알고리즘에서 사용할 수 있는 정보로 전환합니다.

활동 추적기의 경우 가속도계 데이터의 빈도 콘텐츠를 캡처하는 특징을 추출하고자 합니다. 알고리즘에서는 이러한 특징을 기반으로 걷기(낮은 빈도)와 달리기(높은 빈도)를 구분할 수 있습니다. 선택한 특징이 포함된 새 테이블을 만듭니다.

특징 선택을 사용하여 다음을 수행합니다.

- 기계 학습 알고리즘의 정확도 개선
- 고차원 데이터 세트의 모델 성능 향상
- 모델 해석 가능성을 향상
- 오버피팅 방지

3 3단계: 특징 도출 *계속*

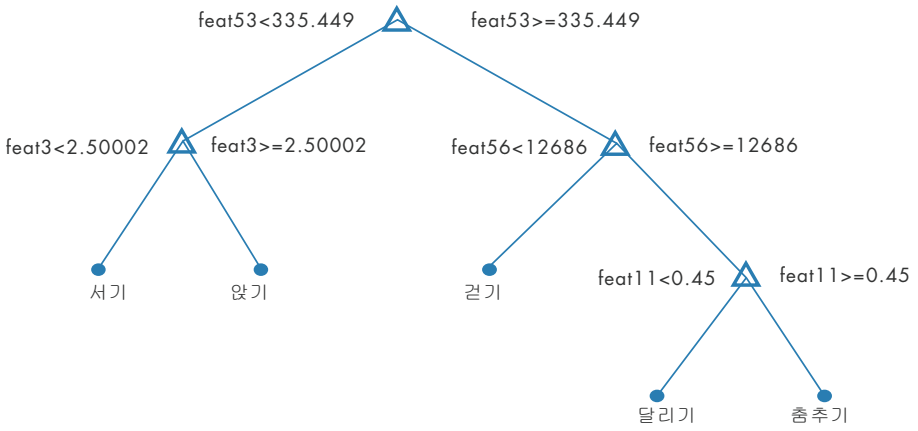
도출할 수 있는 특징의 수에는 제한이 없습니다. 하지만 다양한 유형의 데이터에 일반적으로 사용되는 많은 기법이 있습니다.

데이터 유형	특징 선택 작업	기법
센서 데이터	원시 센서 데이터에서 신호 속성을 추출하여 하이 레벨 정보 생성	**최대값 분석** – FFT를 수행하고 상위(Dominant) 빈도 식별 **펄스 및 전환 메트릭** – 상승 시간, 하강 시간, 정착 시간 등의 신호 특성 도출 **스펙트럼 측정** – 신호 출력, 대역폭, 중심 주파수, 중간 주파수 플로팅
이미지 및 비디오 데이터	모서리 위치, 해상도, 색상 등의 특징 추출	**시각적 단어 모음** – 모서리, 코너, 색상 영역 등의 로컬 이미지 특징에 대한 히스토그램 작성 **HOG(Histogram of Oriented Gradients)** – 로컬 그라디언트 방향에 대한 히스토그램 작성 **최소 고유값 알고리즘** – 이미지에서 코너 위치 감지 **모서리 감지** – 밝기가 급격히 바뀌는 지점 식별
트랜잭션 데이터	데이터에서 정보를 개선하는 도출된 값 계산	**타임스탬프 분해** – 타임스탬프를 일, 월 등의 성분으로 분해 **집계 가치 계산** – 특정 이벤트가 발생한 총 횟수 등의 상위 특징 작성

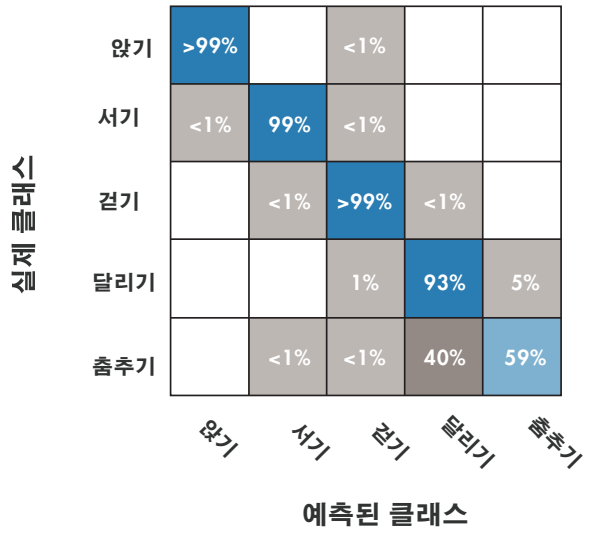
4 4단계: 모델 작성 및 훈련

모델 작성 시 간단한 것부터 시작하는 것이 좋습니다. 간단해야 더 빠르게 실행하고 더 쉽게 해석할 수 있습니다.

기본 의사결정 트리에서 시작합니다.



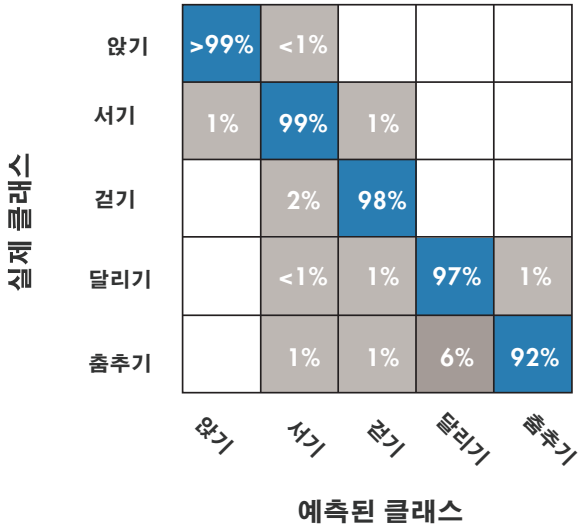
얼마나 잘 작동하는지 확인하기 위해 1단계에서 작성한 실제 클래스 레이블과 모델별로 수행된 분류를 비교하는 테이블인 오차 행렬을 플로팅합니다.



오차 행렬에 의하면 모델에 춤추기와 달리기를 구분하는 데 문제가 있음을 알 수 있습니다. 이 유형의 데이터에는 의사결정 트리가 작동하지 않을 수 있습니다. 몇 가지 다른 알고리즘을 시도해 보겠습니다.

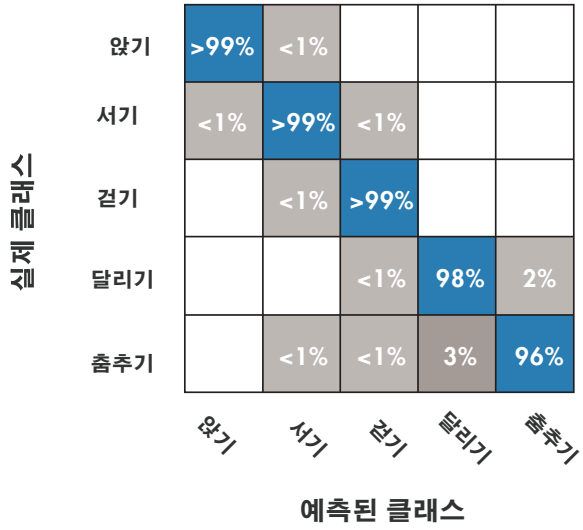
4 4단계: 모델 작성 및 훈련 *계속*

모든 훈련 데이터를 저장하고, 새 포인트를 훈련 데이터에 비교하고, “K” 최근접 포인트의 가장 빈번한 클래스를 반환하는 단순 알고리즘인 KNN(K-Nearest Neighbor)에서 시작합니다. 단순 의사결정 트리의 94.1% 정확도에 비해 이 모델은 98% 정확도를 제공합니다. 오차 행렬도 향상된 모습을 보입니다.



하지만 KNN은 예측을 하는 데 모든 훈련 데이터가 필요하므로 실행 시 상당한 양의 메모리가 사용됩니다.

여기서는 선형 판별 모델을 시도하지만 결과가 향상되지 않습니다. 마지막으로 다계층 SVM(서포트 벡터 머신)을 시도합니다. 다음과 같이 SVM이 매우 잘 작동하고 이제 99% 정확도를 얻습니다.



모델에 반복 수행하고 여러 알고리즘을 시도하여 목표를 달성했습니다. 이제 분류기에서 여전히 춤추기와 달리기를 안정적으로 구별할 수 없는 경우 모델을 개선할 여러 방법을 살펴보겠습니다.

5 5단계: 모델 개선

모델 개선에는 모델을 더 간단히 만들거나 복잡성을 추가하는 두 가지 방향이 있습니다.

간소화

먼저 특징 수를 줄일 수 있는 방법을 찾아봅니다. 널리 사용되는 특징 감소 기법으로는 다음이 있습니다.

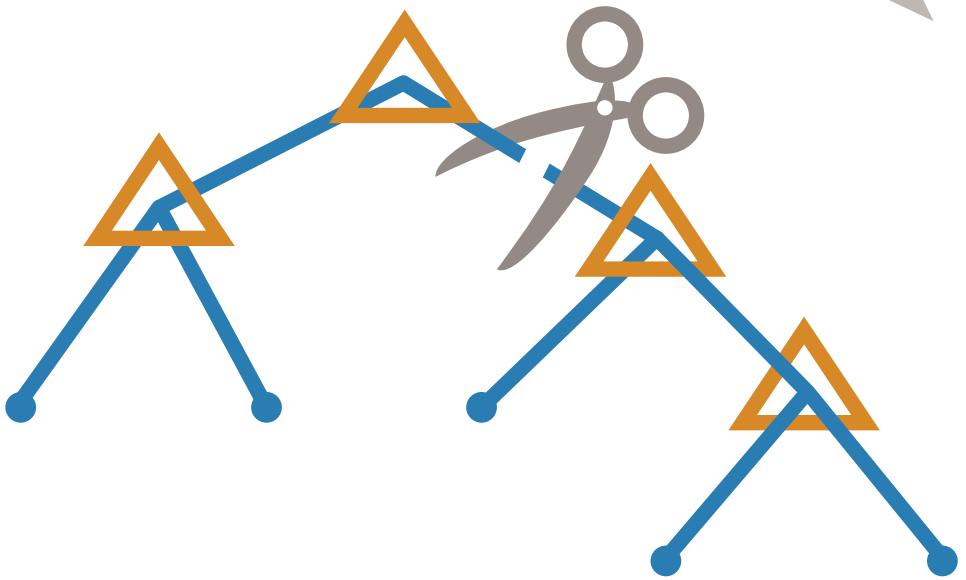
- 상관관계 행렬 – 상관관계가 깊지 않은 변수(또는 특징)를 제거할 수 있도록 변수 간 관계를 표시합니다.
- PCA(주성분 분석) – 원래 특징 간에 주요 차이를 캡처하는 특징 조합을 찾아 중복성을 제거하고 데이터셋에서 강력한 패턴을 도출합니다.
- Sequential Feature selection기법을 이용한 변수 축약 – 더 이상 성능 향상이 없을 때까지 모델에서 반복해서 변수를 줄입니다.

다음으로 모델 자체를 줄이는 여러 방법을 살펴봅니다. 다음과 같은 방법이 있습니다.

- 의사결정 트리에서 분기 정리
- 앙상블 학습을 통해 모델의 불필요한 변수를 줄입니다.

좋은 모델에는 가장 예측 가능성이 높은 특징만 포함됩니다. 잘 일반화되는 단순 모델이 새 데이터에 대해 일반화하거나 잘 훈련될 수 없는 복잡한 모델보다 효율적입니다.

머신 러닝의 경우 기타 여러 계산 프로세스에서처럼 모델을 간소화하여 더 이해하기 쉽고, 더 강력하고, 더 계산 효율적인 모델을 만들 수 있습니다.



5 5단계: 모델 개선 *계속*

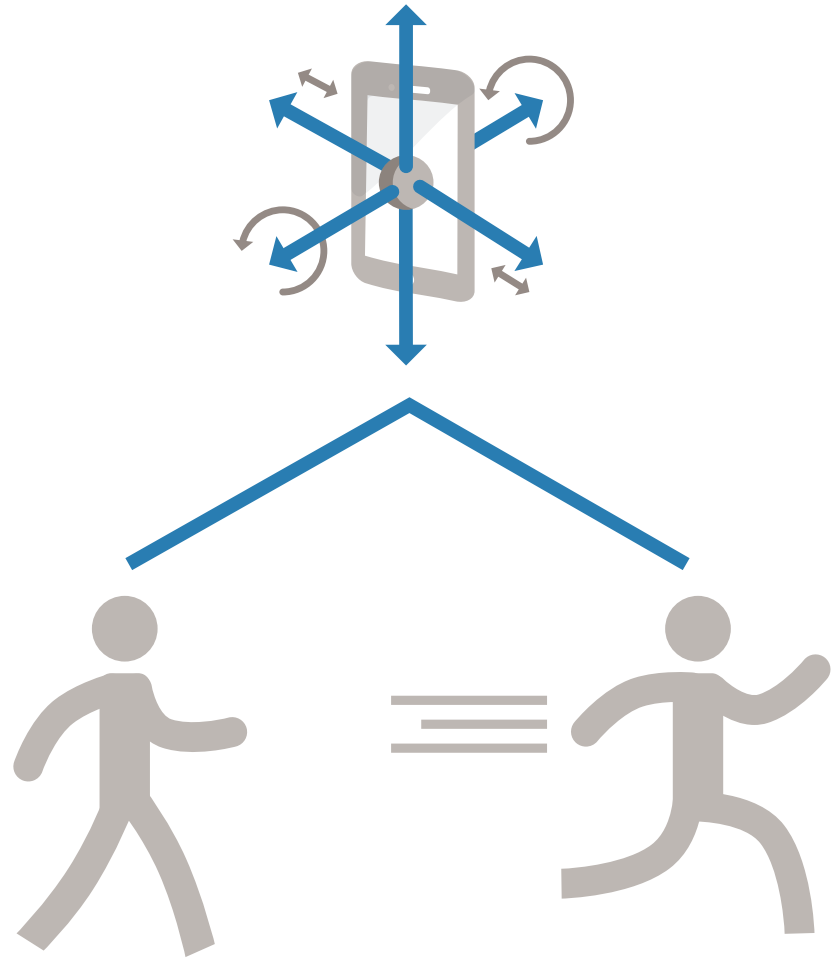
복잡성 추가

지나친 일반화로 인해 모델에서 춤추기와 달리기를 구별할 수 없다면 모델을 더 미세 조정하는 방법을 찾아야 합니다. 방법은 다음과 같습니다.

- 모델 조합 사용 – 간단한 여러 모델을 해당 모델이 단독으로 할 수 있는 것보다 데이터에서 추세를 더 잘 나타낼 수 있는 더 큰 모델로 병합합니다.
- 더 많은 데이터 추가 - 가속도 센서와 위치 센서 데이터를 살펴봅니다. 위치 센서 데이터는 활동 중에 휴대 전화의 방향을 기록합니다. 이 데이터는 여러 가지 활동에 대한 고유한 특징을 제공할 수 있습니다. 예를 들어 달리기에는 고유한 가속도 및 회전 조합이 있을 수 있습니다.

모델을 조정한 후 전처리 중에 준비한 테스트 데이터에 대해 성능을 검증합니다.

모델이 테스트 데이터에서 활동을 안정적으로 분류할 수 있다면 모델을 전화기로 이동하고 추적을 시작할 준비가 된 것입니다.



추가 정보

자세히 살펴볼 준비가 되셨습니까? 다음 리소스를 통해 머신 러닝 방법, 예제, 도구를 자세히 알아보십시오.

시청 자료

머신 러닝으로 간편하게 34:34

센서 데이터 분석을 위한 신호 처리 및 머신 러닝 기법 42:45

읽기 자료

지도학습 워크플로우 및 알고리즘

MATLAB 분석을 사용한 데이터 기반 통찰력: 에너지 부하 예측 사례 연구

살펴보기

MATLAB 머신 러닝 예제

Classification Learner 앱을 사용한 데이터 분류


3부: 비지도 학습 적용

비지도 학습을 고려하는 경우

비지도 학습은 데이터를 탐색하려고 하지만 아직 구체적인 목표가 없거나 데이터에 포함된 정보가 무엇인지 확실하지 않은 경우 유용합니다. 데이터의 차원을 줄이는 것도 좋은 방법입니다.

비지도 학습 기법

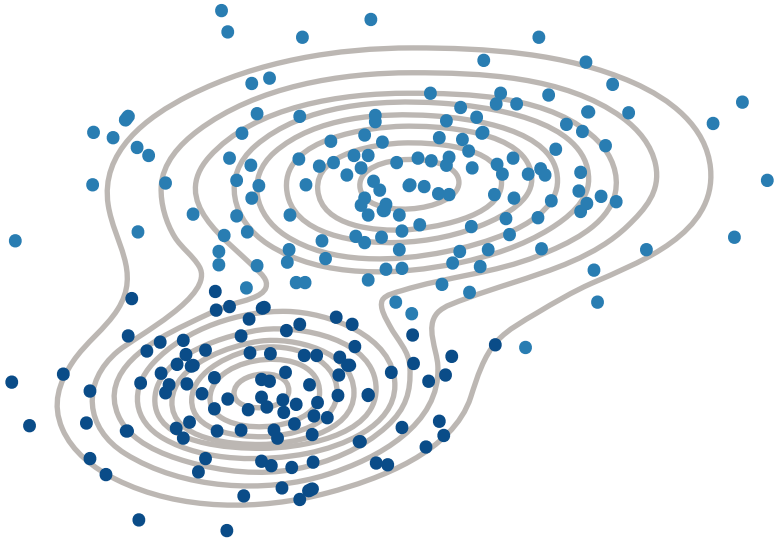
섹션 1에서 확인한 대로 대부분의 비지도 학습 기법은 클러스터 분석의 형태입니다.

클러스터 분석에서 데이터는 유사성 또는 공유 특성의 측정값을 기반으로 그룹으로 분할됩니다. 클러스터는 같은 클러스터의 객체는 매우 비슷하고 다른 클러스터의 객체는 뚜렷이 구별되도록 구성됩니다.

클러스터링 알고리즘은 다음 두 가지 큰 범주로 구분됩니다.

- 하드 클러스터링 - 각 데이터 포인트가 하나의 클러스터에만 속함
- 소프트 클러스터링 - 각 데이터 포인트가 두 개 이상의 클러스터에 속함

가능한 데이터 그룹을 이미 알고 있다면 하드 또는 소프트 클러스터링 기법을 사용할 수 있습니다.



데이터를 두 개의 클러스터로 분할하는 데 사용되는 가우시안 혼합 모델.

데이터 그룹화 방법을 잘 모르는 경우:

자기 조직화 특징 맵 또는 계층 클러스터링을 사용하여 데이터에서 가능한 구조체를 찾습니다.

클러스터 평가를 사용하여 지정된 클러스터링 알고리즘에 대한 “최적”의 그룹 수를 찾습니다.

일반적인 하드 클러스터링 알고리즘

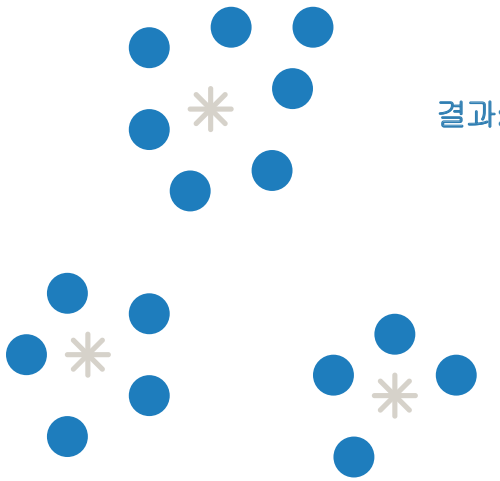
k-평균

작동 방식

데이터를 k개의 상호 배타적인 클러스터로 분할합니다. 포인트가 클러스터에 얼마나 잘 피팅되는지는 해당 포인트에서 클러스터 센터까지의 거리에 따라 결정됩니다.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 대용량 데이터 세트의 빠른 클러스터링을 위해



결과: 클러스터 센터

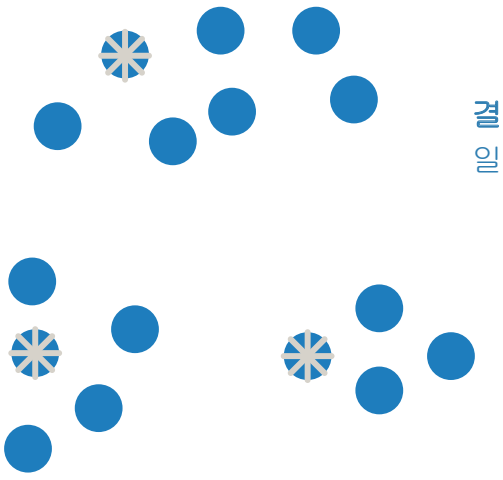
k-중간점

작동 방식

k-평균과 비슷하지만 클러스터 센터가 데이터의 여러 포인트와 일치해야 하는 요구 사항이 있습니다.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 범주형 데이터의 빠른 클러스터링을 위해
- 대용량 데이터 세트로 확장하기 위해



결과: 데이터 포인트와 일치하는 클러스터 센터

일반적인 하드 클러스터링 알고리즘 *계속*

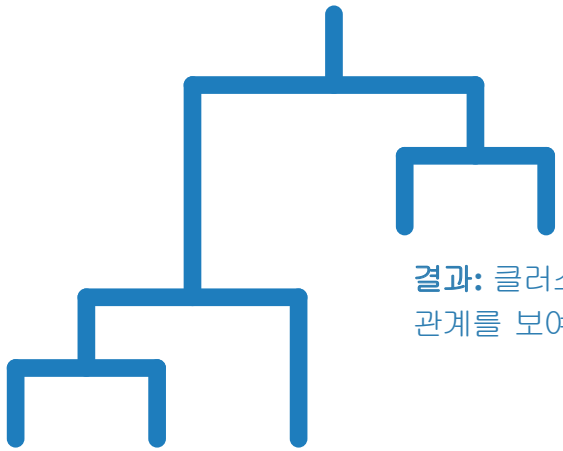
계층 클러스터링

작동 방식

포인트 쌍 간의 유사성을 분석하고 객체를 이진, 계층적 트리로 그룹화하여 중첩된 클러스터 세트를 생성합니다.

최적 사용...

- 데이터에 있는 클러스터 수를 미리 알지 못하는 경우
- 선택에 도움이 되도록 시각화를 원할 경우



결과: 클러스터 간 계층적 관계를 보여주는 덴드로그램

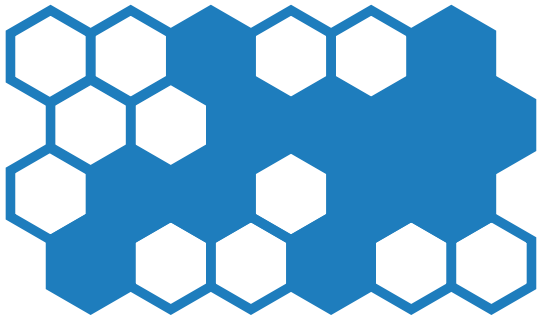
자기 조직화 맵

작동 방식

데이터셋을 토폴로지 보존 2차원 맵으로 변환하는 신경망 기반 클러스터링.

최적 사용...

- 고차원 데이터를 2차원 또는 3차원으로 시각화하기 위해
- 토폴로지(모양)를 보존하여 데이터의 차원을 추론하기 위해



결과:
저차원(일반적으로 2차원) 표현

일반적인 하드 클러스터링 알고리즘 *계속*

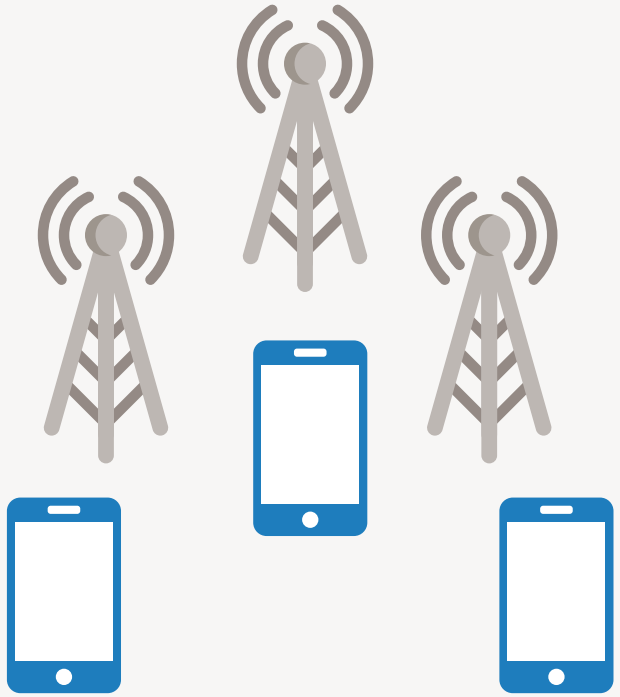
사례: 사이트 휴대 전화 기지국에 대한 k-평균 클러스터링 사용

한 휴대 전화 회사에서 가장 안정적인 서비스를 제공할 휴대 전화 기지국의 수와 배치를 알고자 합니다. 신호 수신을 최적화하기 위해 기지국은 사람들의 클러스터 내에 있어야 합니다.

워크플로우는 나중에 필요하게 될 클러스터 수에 대한 초기 추측(Initial Guess) 으로 시작됩니다. 이 추측을 평가하기 위해 엔지니어는 서비스를 기지국 3개 및 기지국 4개와 비교하여 각 시나리오에 대해 얼마나 잘 클러스터링할 수 있는지(즉, 기지국이 서비스를 얼마나 잘 제공하는지) 확인합니다.

전화기 한 대는 한 번에 한 곳의 기지국에만 신호를 보낼 수 있으므로 이러한 클러스터링은 매우 어렵습니다. k-평균은 데이터의 각 관측값을 공간 정위를 포함하는 객체로 처리하므로 팀에서는 k-평균 클러스터링을 사용합니다. k-평균 클러스터링은 각 클러스터 내의 객체가 가능한 한 서로에게 가깝고 가능한 한 다른 클러스터의 객체와 멀리 있는 파티션을 찾습니다.

알고리즘을 실행한 후 팀에서는 데이터를 클러스터 3개 및 4개로 분할한 결과를 정확히 확인할 수 있습니다.



일반적인 소프트 클러스터링 알고리즘

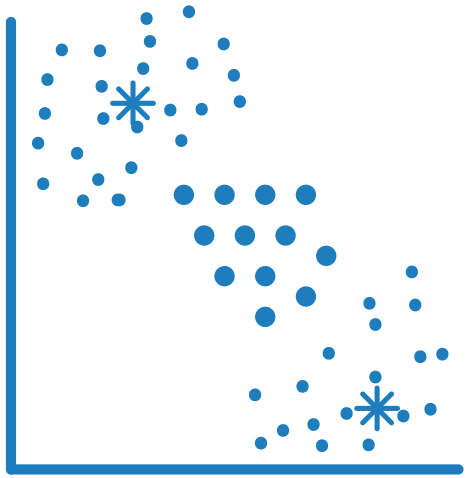
퍼지 c-평균

작동 방식

데이터 포인트가 둘 이상의 클러스터에 속할 수 있는 경우 파티션 기반 클러스터링.

최적 사용...

- 클러스터 수를 알고 있을 경우
- 패턴 인식을 위해
- 클러스터가 겹칠 경우



결과: 클러스터 센터 (k-평균과 유사), 그러나 포인트가 둘 이상의 클러스터에 속할 수 있도록 퍼지니스 포함

가우시안 혼합 모델

작동 방식

데이터 포인트가 특정 확률을 포함한 다양한 다변량 정규 분포에서 나오는 파티션 기반 클러스터링.

최적 사용...

- 데이터 포인트가 둘 이상의 클러스터에 속할 수 있는 경우
- 클러스터의 크기와 클러스터 내의 상관관계 구조체가 다양할 경우

결과: 포인트가 클러스터에 있을 확률을 제공하는 가우시안 분포 모델

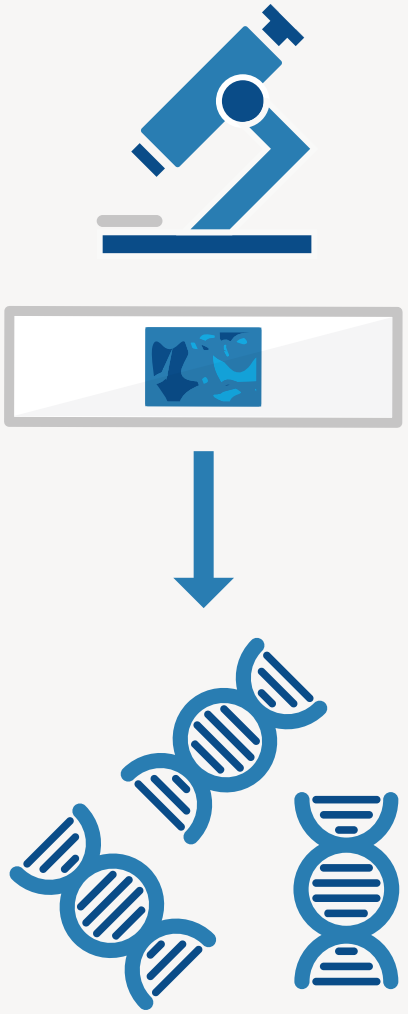
일반적인 소프트 클러스터링 알고리즘 *계속*

사례: 퍼지 c-평균 클러스터링을 사용하여 유전자 발현 데이터 분석

한 생물학자 팀은 정상 및 비정상 세포 분열에 관여한 유전자를 더 잘 이해하기 위해 마이크로어레이에서 유전자 발현 데이터를 분석하고 있습니다. (유전자가 단백질 생성과 같은 세포성 기능에 적극적으로 관여한 경우 유전자가 "발현"된다고 말합니다.)

마이크로어레이에는 두 개의 조직 표본에 기반을 둔 발현 데이터가 포함됩니다. 연구자들은 표본을 비교하여 유전자 발현의 특정 패턴이 암 확산에 관련되는지 확인하고자 합니다.

데이터를 전처리하여 노이즈를 제거한 후 데이터를 클러스터링합니다. 같은 유전자가 여러 생물학적 과정에 관여될 수 있으므로 단일 유전자는 하나의 클러스터에만 속하지 않을 가능성이 높습니다. 연구자들은 퍼지 c-평균 알고리즘을 데이터에 적용합니다. 그러고 나서 클러스터를 시각화하여 비슷한 방식으로 동작하는 유전자 그룹을 식별합니다.



차원성 감소를 통한 모델 개선

머신 러닝은 큰 데이터셋에서 패턴을 찾을 수 있는 효과적인 방법입니다. 하지만 데이터가 크면 더 복잡해집니다.

데이터셋이 커질수록 특징 수 또는 *차원*을 줄여야 하는 경우가 많아집니다.

사례: EEG 데이터 감소

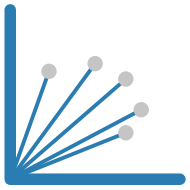
뇌의 전기 활동을 캡처하는 EEG(뇌전도) 데이터가 있다고 가정하고 이 데이터를 사용하여 미래에 있을 수 있는 발작을 예측하려고 합니다. 데이터는 각각 원래 데이터셋의 변수에 해당하는 수십 개의 리드를 사용하여 캡처되었습니다. 이러한 각 변수에는 노이즈가 포함됩니다. 예측 알고리즘을 더 강력하게 만들려면 차원성 감소 기법을 사용하여 더 적은 수의 특징을 도출합니다. 이러한 특징은 여러 센서에서 계산되므로 원시 데이터를 직접 사용한 경우보다는 개별 센서의 노이즈에 덜 민감합니다.

일반적인 차원성 감소 기법

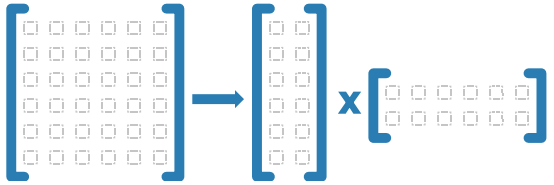
가장 일반적으로 사용되는 세 가지 차원성 감소 기법은 다음과 같습니다.

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.

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



음수 미포함 행렬 분해 - 모델 텀(term)이 물리적 수량과 같은 음이 아닌 수량을 표현해야 할 경우 사용됩니다.



주성분 분석 사용

많은 변수를 포함하는 데이터셋에서 변수 그룹은 보통 함께 이동합니다. PCA는 적은 수의 새 변수가 대부분의 정보를 캡처하도록 원래 변수의 일차 결합을 통해 새 변수를 생성하여 이 정보 중복을 활용합니다.

각 주성분은 원래 변수의 일차 결합입니다. 모든 주성분은 서로 직각이므로 중복된 정보가 없습니다.

사례: 엔진 상태 모니터링

엔진의 여러 센서에 대한 측정값이 포함된 데이터셋이 있습니다 (온도, 압력, 배출 등). 많은 데이터가 정상 상태의 엔진에서 나오는 데이터이지만, 센서에서 유지관리가 필요할 경우의 엔진에서 나오는 데이터도 캡처했습니다.

개별 센서를 살펴봐도 분명한 이상 증상을 확인할 수 없습니다. 하지만 PCA를 적용하면 센서 측정값에 있는 대부분의 변형이 적은 수의 주성분을 통해 캡처되도록 이 데이터를 변환할 수 있습니다. 원시 센서 데이터를 확인하는 것보다 이러한 주성분을 검사하면 정상 및 비정상 상태 엔진을 더 쉽게 구분할 수 있습니다.

인자 분석 사용

데이터셋에 겹치는(서로 종속된) 측정 변수가 포함될 수 있습니다. 인자 분석을 사용하면 모델을 다변량 데이터에 피팅하여 이런 종류의 상호 종속성을 예측할 수 있습니다.

인자 분석 모델에서 측정 변수는 더 적은 수의 눈에 띄지 않는 (잠재적) 인자에 의존합니다. 각 인자는 여러 변수에 영향을 미칠 수 있으므로 공통 인자라고 알려져 있습니다. 각 변수는 공통 인자의 일차 결합에 종속된다고 가정합니다.

사례: 주가 변동 추적

100주 동안 10개 회사에 대한 주가의 퍼센트 변화가 기록되었습니다. 10개 회사 중 4개는 기술 회사, 3개는 금융 회사, 나머지 3개는 유통 회사입니다. 같은 부문의 회사들의 주가가 경제 조건이 바뀌면서 함께 달라진다는 가정은 합리적인 것 같습니다. 인자 분석은 전제를 뒷받침할 정량적 증거를 제공할 수 있습니다.

음수 미포함 행렬 분해

이 차원 감소 기법은 특징 공간의 저차수 근사법을 기반으로 합니다. 음수를 포함 하지 않는 행렬에 대한 특징 수를 줄일수 있습니다.

사례: 텍스트 마이닝

여러 웹 페이지의 어휘와 스타일 차이를 살펴보려 한다고 가정해 보겠습니다. 각 행이 개별 웹 페이지에 해당하고 각 열이 단어 (“the”,”a”,”we” 등)에 해당하는 행렬을 만듭니다. 데이터는 특정 페이지에서 특정 단어가 나타나는 횟수입니다.

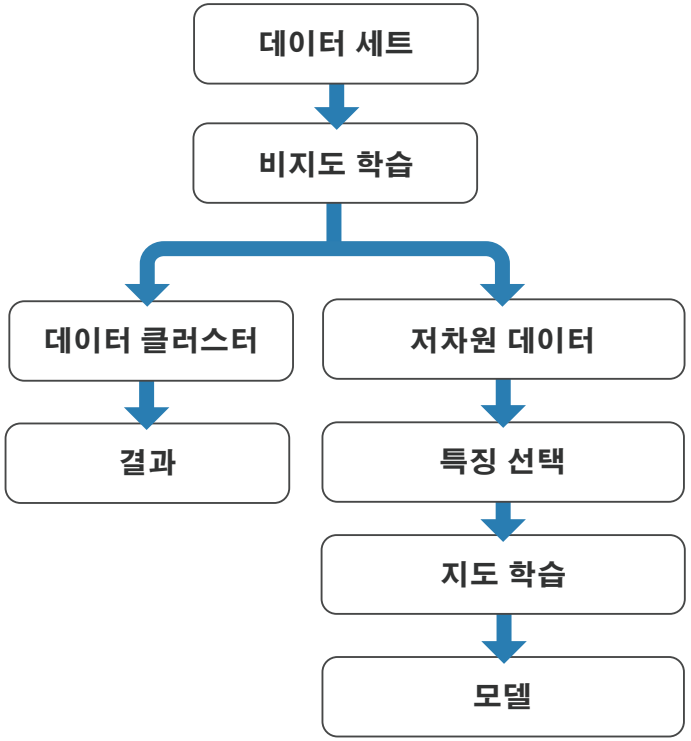
영어에는 수백만 개 이상의 단어가 있으므로 음수 미포함 행렬 분해를 적용하여 개별 단어가 아니라 하이 레벨 개념을 나타내는 임의 개수의 특징을 만듭니다. 이러한 개념을 사용하면 발언, 뉴스, 교육 콘텐츠, 온라인 소매 콘텐츠를 더 쉽게 구별할 수 있습니다.

다음 단계

이 섹션에서는 비지도 학습을 위한 하드 및 소프트 클러스터링 알고리즘을 더 자세히 살펴보고, 데이터에 적합한 알고리즘을 선택할 경우의 몇 가지 팁을 제공하고, 데이터셋에서 특징 수를 줄이면 모델 성능이 어떻게 향상되는지 설명했습니다. 다음 단계에서는 다음을 수행합니다.

- 비지도 학습이 최종 목표일 수 있습니다. 예를 들어 시장 조사를 수행하는 동안 웹 사이트 동작에 따라 소비자 그룹을 구분하여 대상으로 지정하고자 할 경우 거의 확실하게 클러스터링 알고리즘을 통해 찾고 있는 결과를 얻을 수 있을 것입니다.
- 반면, 비지도 학습을 지도 학습의 전처리 단계로 사용하려고 할 수 있습니다. 예를 들어 클러스터링 기법을 적용하여 더 적은 수의 특징을 도출하고 해당 특징을 분리기 훈련을 위한 입력으로 사용합니다.

섹션 4에서는 지도 학습 알고리즘 및 기법을 살펴보고 특징 선택, 특징 감소, 파라미터 튜닝을 통해 모델을 개선하는 방법을 알아보겠습니다.



추가 정보

자세히 살펴볼 준비가 되셨습니까? 다음 비지도 학습 리소스를 살펴보십시오.

클러스터링 알고리즘 및 기법

k-평균

K-평균 및 계층 클러스터링을 사용하여 데이터의 자연 패턴 찾기

K-평균 및 자기 조직화 맵을 사용하여 유전자 클러스터링

K-평균 클러스터링을 사용한 색상 기반 분할

계층 클러스터링

연결 기반 클러스터링

아이리스 클러스터링

자기 조직화 맵

자기 조직화 맵을 사용하여 데이터 클러스터링

퍼지 C-평균

퍼지 C-평균 클러스터링을 사용하여 의사 임의 데이터 클러스터링

가우시안 혼합 모델

가우시안 프로세스 회귀 모델

가우시안 분포 혼합으로부터 데이터 클러스터링

소프트 클러스터링을 사용하여 가우시안 혼합 데이터 클러스터링

가우시안 혼합 모델 튜닝

이미지 처리 예: 가우시안 혼합 모델을 사용한 자동차 감지

차원성 감소

PCA를 사용한 미국 도시 내 생활의 질 분석

인자 분석을 사용한 주가 분석

음이 아닌 행렬 분해

음이 아닌 행렬 분해 수행

차감 클러스터링을 사용한 교외 통근 모델링


4부: 지도 학습 적용

지도학습을 고려해야 하는 경우

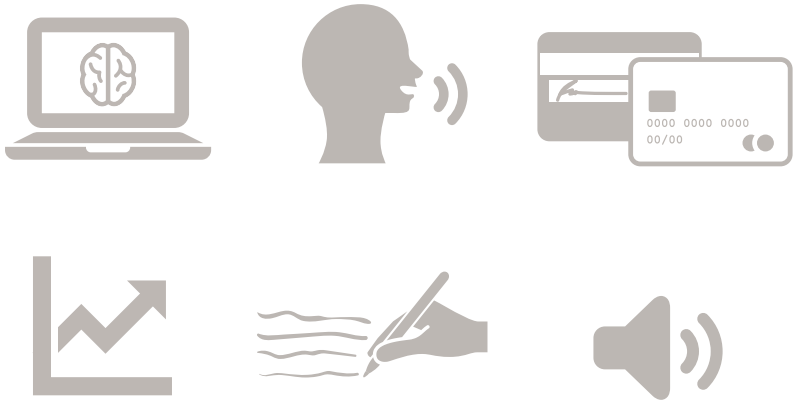
지도학습 알고리즘은 알려진 입력 데이터 세트(훈련 세트) 및 알려진 데이터에 대한 응답(출력)을 사용하고 새 입력 데이터에 대한 응답을 위해 합리적인 예측을 생성하도록 모델을 훈련합니다. 예측하려고 하는 출력에 대한 기존 데이터가 있는 경우 지도학습을 사용합니다.

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.

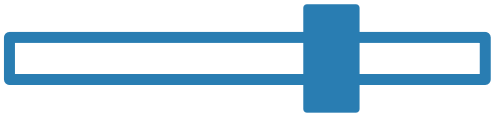
더 큰 훈련 데이터셋을 사용하면 새 데이터에 대해 일반화가 잘 수행되는 모델이 생성되는 경우가 많습니다.

훈련 속도

메모리 사용량

예측 정확도

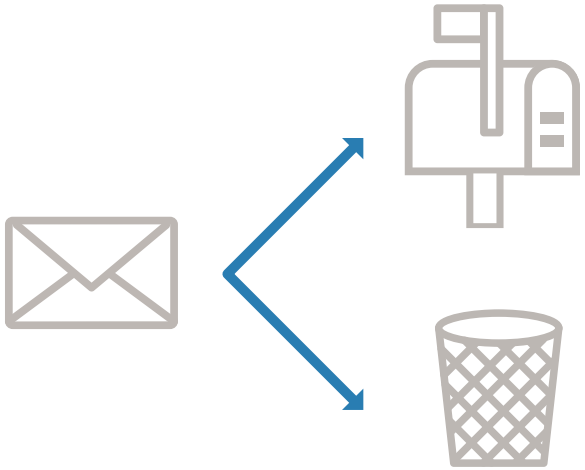
해석 가능성



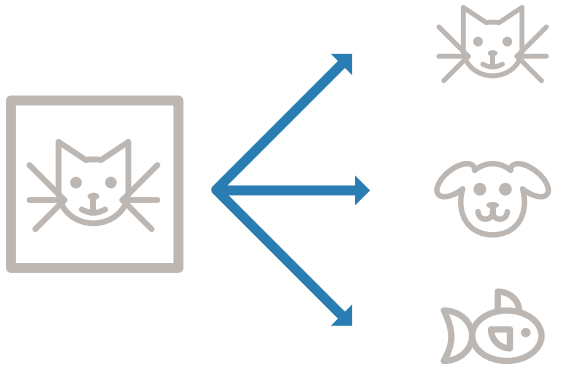
이진 계층 대 다계층 분류

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

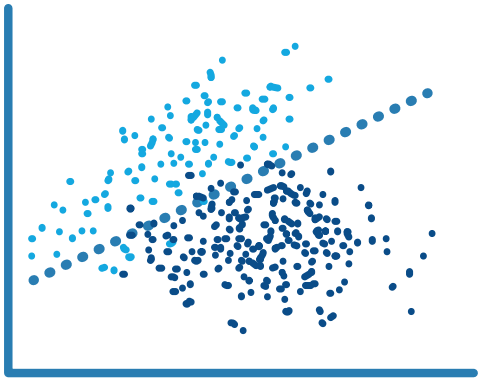
로지스틱 회귀

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



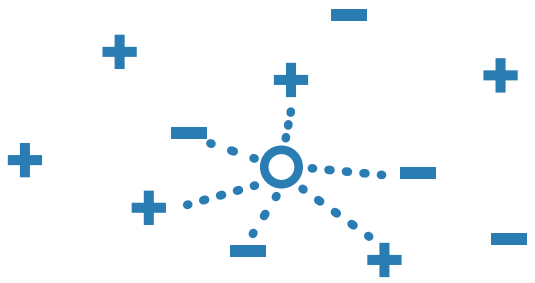
kNN(k-Nearest Neighbor)

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

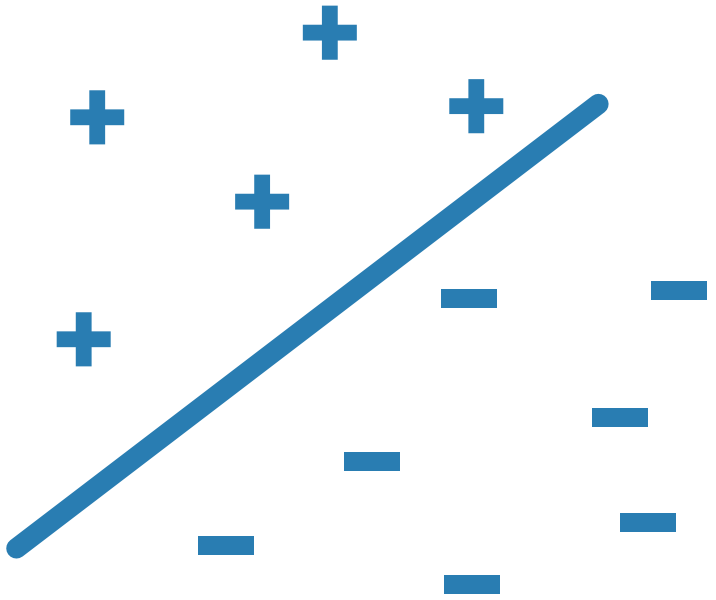


일반적인 분류 알고리즘 *계속*

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*

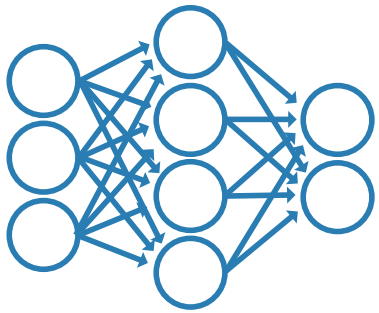
인공신경망

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



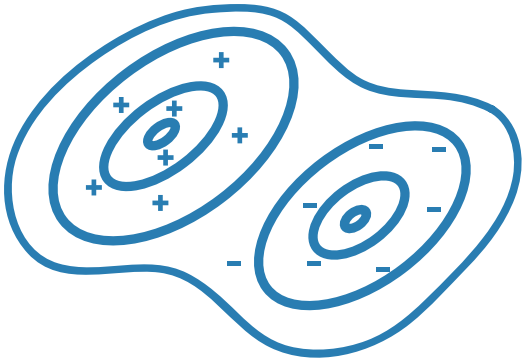
나이브 베이지안

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

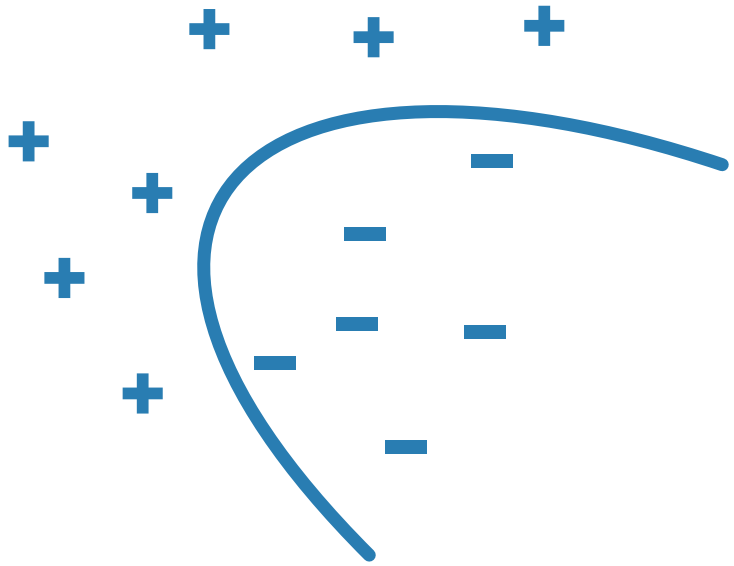


일반적인 분류 알고리즘 *계속*

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*

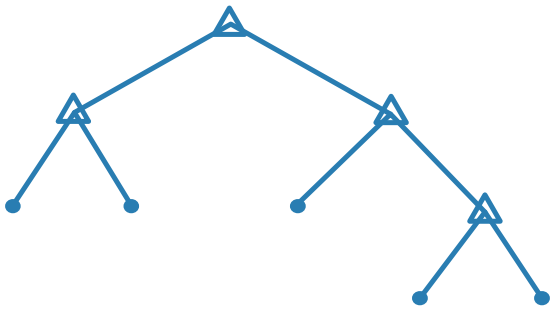
의사결정 트리

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

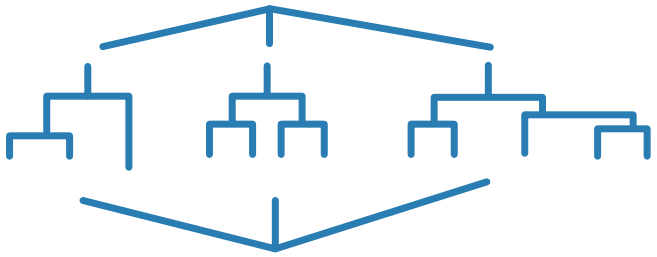
이러한 앙상블 방법에서는 여러 “더 약한” 의사결정 트리가 “더 강한” 앙상블로 결합됩니다.

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



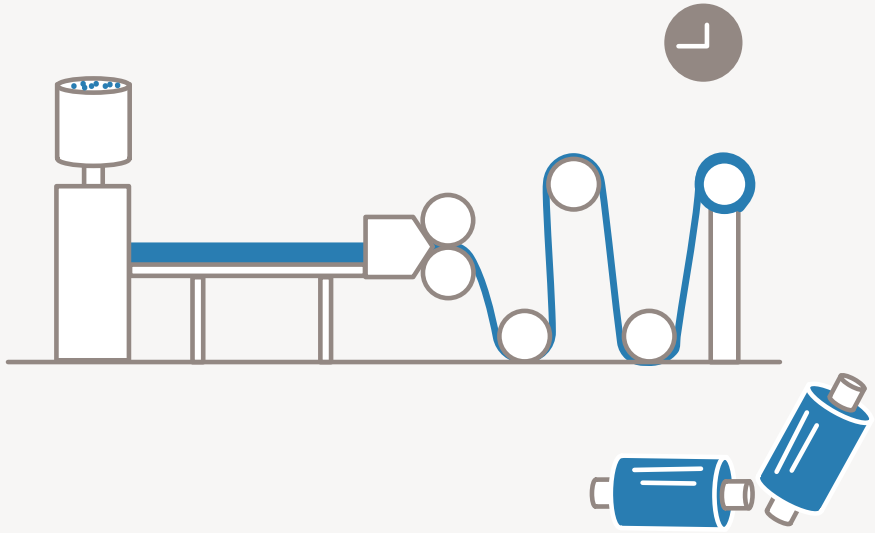
일반적인 분류 알고리즘 *계속*

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

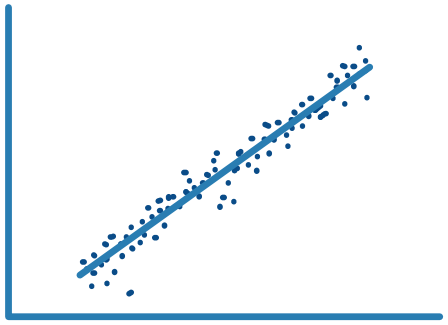
선형 회귀

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

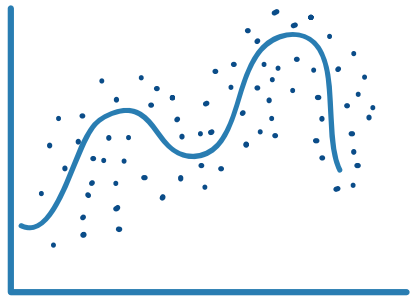
작동 방식

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*

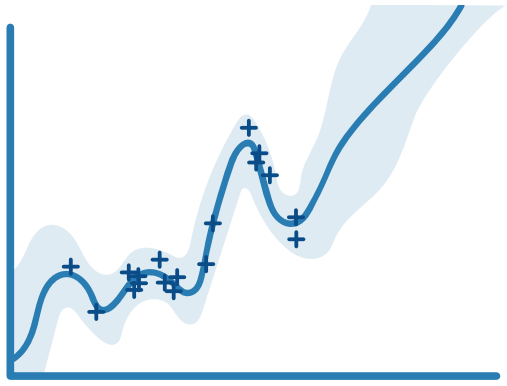
가우시안 프로세스 회귀 모델

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



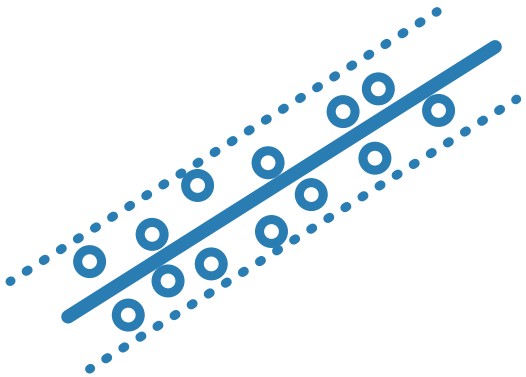
SVM 회귀

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*

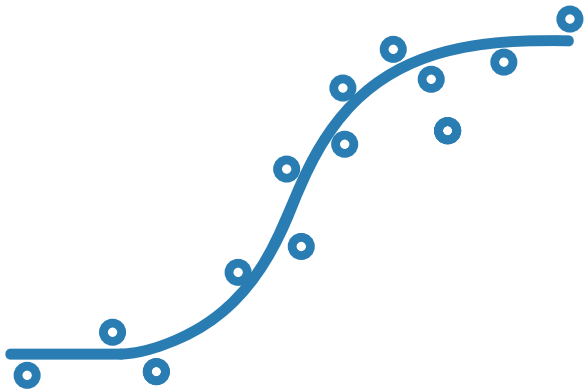
일반화된 선형 모델

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



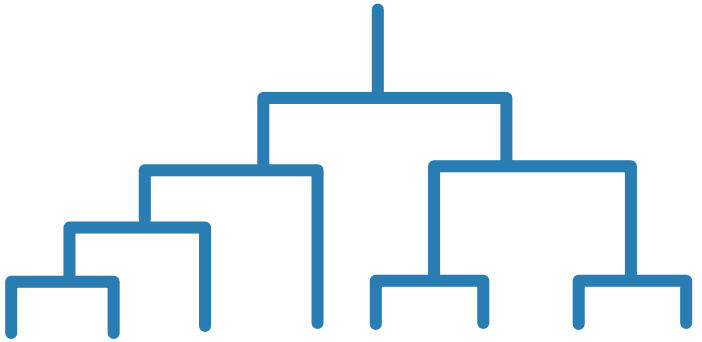
회귀 트리

작동 방식

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

최적 사용...

- 예측자가 범주형(개별) 이거나 비선형적으로 동작하는 경우

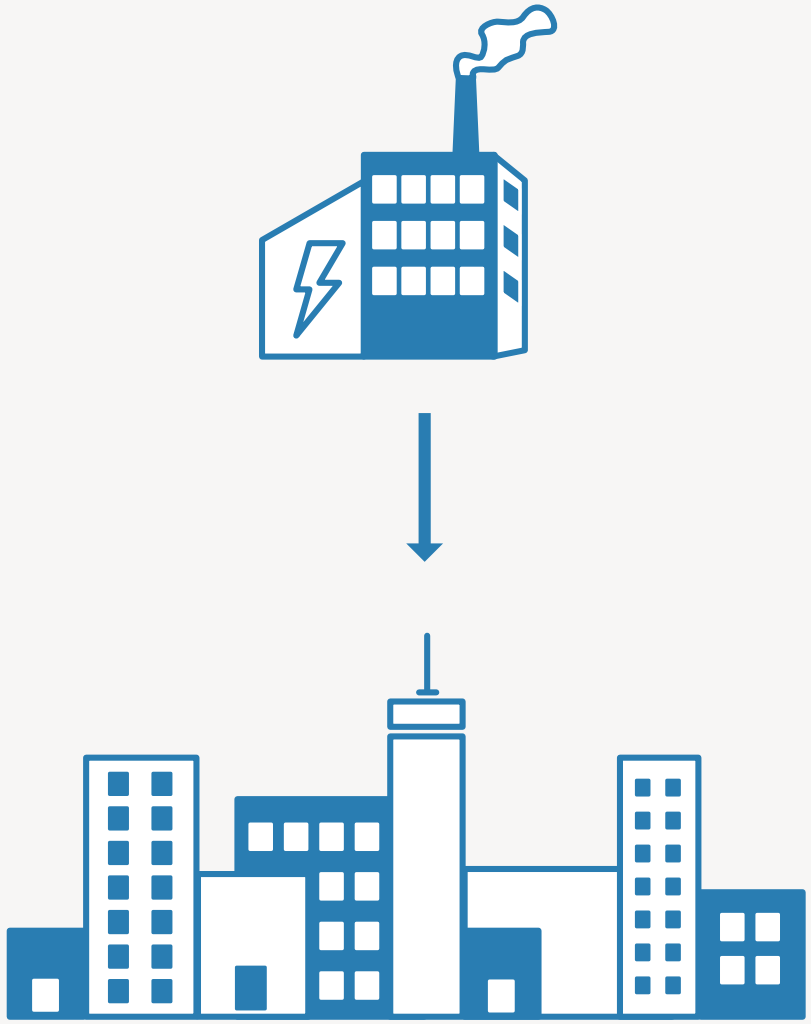


일반적인 회귀 알고리즘 *계속*

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.



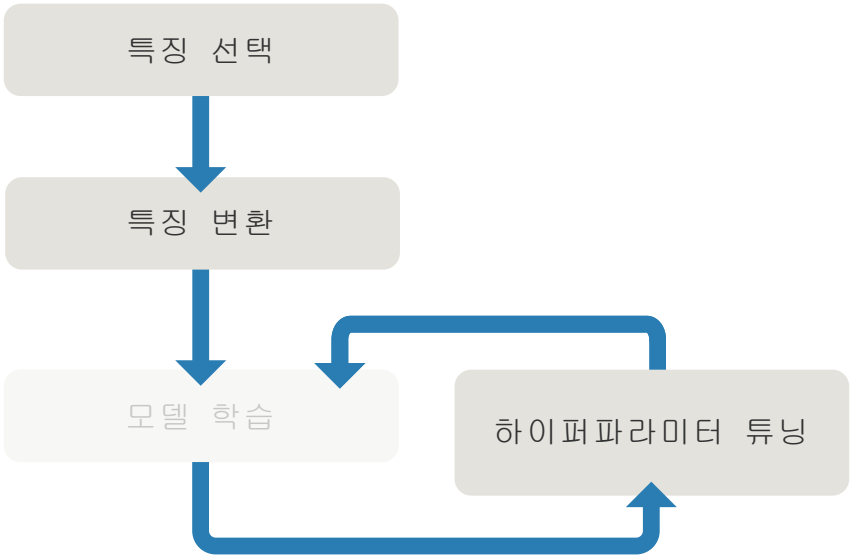
모델 개선

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.

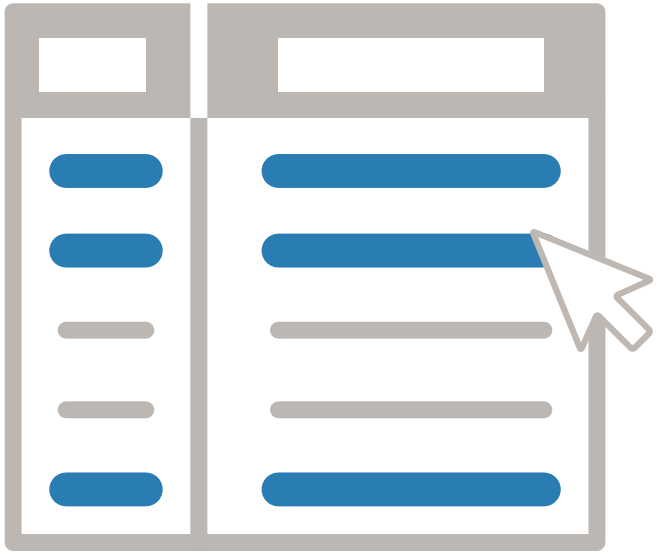
일반적인 특징 선택 기법은 다음과 같습니다.

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.



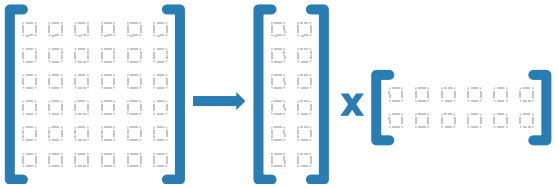
모델은 훈련하기 위해 선택하는 특징과 거의 비슷합니다.

특징 변환

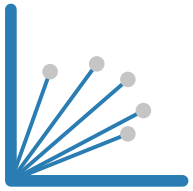
특징 변환은 차원성 감소의 형태입니다. 섹션 3에서 살펴본 대로 가장 일반적으로 사용되는 세 가지 차원성 감소는 다음과 같습니다.

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.

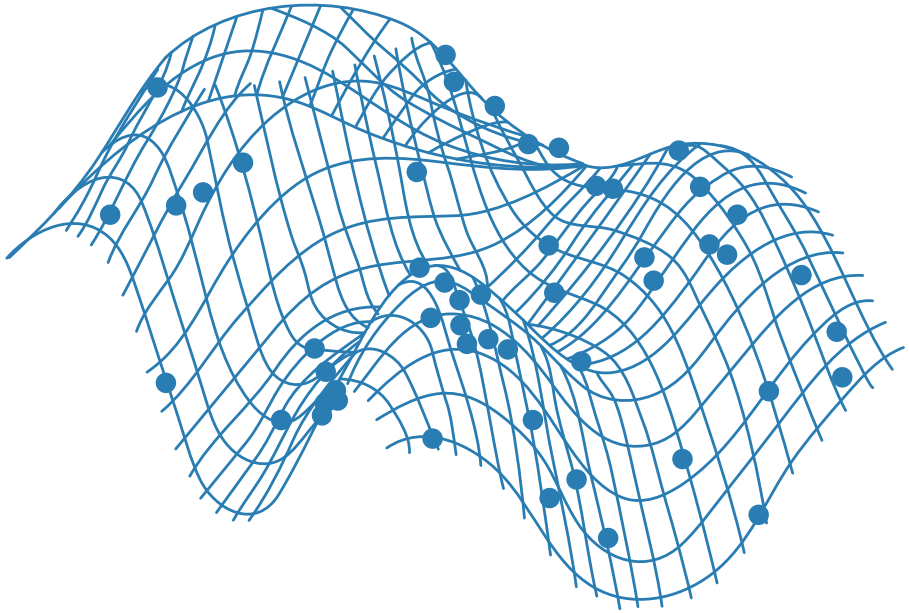


하이퍼파라미터 튜닝

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