물리학의 많은 부분과 화학 전체에 대한 수학적 이론에 필요한 기본적인 물리적 법칙은 명백히 널리 알려져 있다. 그렇지만 어려운 점은 이러한 법칙의 정확한 적용이 문제를 풀기에는 너무 복잡한 방정식으로 이어진다는 것이다. 그러므로 양자 역학을 적용하는 근사적이고 실용적인 방법이 개발되어야 한다.

Introduction

이 튜토리얼은 커널 기반 머신 러닝 방법을 사용하여 양자 화학의 수치 결과에 대한 짧고 실용적인 소개를 제공한다. 소개, 핵심 아이디어 설명 및 컨텍스트 제공, 커널 리지 회귀의 예제를 사용하여 커널 기반 머신 러닝을 도입한 이론 부분, 원자화 에너지를 예측하는 방법에 대한 실제 사례를 제공하는 실용적인 부분, 작은 유기 분자(▷)로 표시된 연습은 주제에 대해 독자가 즉각적인 실제 경험을 하도록 만든다. 알고리즘은 쉬운 구현을 위해 의사 코드로 제공된다. Supporting Information 은 Mathematic(이것은 솔루션과 작업할 데이터 세트가 있는 기록수단이다)에서 사용되는 루틴의 기본 구현을 제공한다. 표 1 은 사용된 두 문자어 및 표기법의 용어집을 제공한다.

Why QM/ML models?

원자 규모에 대한 물질 이론인 양자역학은 슈뢰딩거의 방정식의 수치 해법을 통해 시스템의 많은 특성을 계산할 수 있다. 그렇다면 왜 재료 과학, 유기 화학 또는 생화학에서 더 많은 문제가 해결되지 않았을까? 가장 큰 문제는 필요한 계산 작업이다. 계산 작업이 시스템 크기에 따라 빠르게 증가하므로 실험 1 과 일치하는 수치 솔루션은 소형 시스템에만 국한된다. 또한 이러한 상황은 근사를 필요로 한다. 이를 해결하기 위해 BornOppenheimer approximation 과 같은 개념적 차원과 수치 적 차원에서 많은 근사치가 개발되어 슈뢰딩거의 방정식을 근사적으로 풀 수 있는 다양한 접근법이 개발되었다(표 2). 점근적 런타임에서 차이점의 중요성을 가정하려면 시스템의 크기를 두 배로 늘려야 한다. 그리고 런타임과 결합된 클러스터 방법의 경우 런타임은 밀도가 128 배만큼 증가하지만 런타임 기능을 갖는 밀도 함수 이론의 경우에는 몇 배가되면 컴퓨팅 리소스가 부족해질 것이다.

따라서 대규모 시스템 또는 많은 수의 소형 시스템에 대해 보다 높은 수준의 양자역학 방법의 정확성과 폭 넓은 적용 가능성이 필요할 때는 어떻게 해야 할까? 선형 스케일링 양자역학 방법이 우수한 점근적 런타임에 대한 지역성을 이용하여 대안을 제공할 수 있겠지만 이것을 모든 시스템에 적용할 수는 없으며, 또한 이는 사전 요소를 필요로 한다. 다른 방법은 QM 솔루션의 빠르고 정확한 근사를 위해 머신 러닝을 이용하는 것이다

인공 지능의 하위 분야인 머신러닝은 성능이 데이터와 함께 향상되는 알고리즘을 연구한다(경험으로부터 학습을 하는 것이다). 그것의 주요 관심사는 예측 또는 분석을 위한 데이터에서 규칙성(무작위성)을 체계적으로 식별하고 활용하는 것이다. 머신러닝은 뇌 컴퓨터 인터페이스, 추천 시스템, 로봇 공학, 화학(cheminformatics)과 같은 다양한 분야에서 성공적으로 적용되어 왔다. 널리 알려진 알고리즘으로는 인공 신경 네트워크와 지원 벡터 머신이 있다. 머신러닝은 다양한 설정에서 사용할 수 있다. 이 듀토리얼은 감독된 학습 문제에 초점을 맞춘다. 학습하는 문제는 n 개의 참조쌍 f 의 학습 집합이 주어지면 x(입력)에서 y(레이블)로의 매핑(함수)을 학습한다. 양자역학에서 그러한 문제의 예는 ab initio 분자 역학이다. 이 분자역학은 잠재적인

에너지 표면을 배우고, 시스템 구성을 에너지로 매핑한다. 또한 퀘도가 없는 밀도 함수 이론과 전자 밀도에서 운동 에너지에 이르는 지에 대해서도 배운다. 그리고 분자 특성 예측을 통해 분자를 특성 값에 매핑한다. 그 핵심의 대부분은 입력 간의 유사성에 대한 개념에 기반하여 고차원 공간에서 종종 데이터 포인트 사이를 보간법으로 푸는 것이다. 이와 같이, 머신러닝은 본질적으로 귀납적이며, 데이터 중심적이고, 고도로 경험적이다. 일부 머신러닝의 방법중에서도 특히 커널 기반 학습과 같은 현대의 방법은 강한 이론적 토대를 가지고 있다. 이러한 의미에서 머신러닝은 원칙적으로 경험적이다.

The key idea

QM/ML 모델은 머신러닝을 사용하여 일련의 양자역학 계산에서 중복을 활용하여 참조 계산 사이사이에 보간한다. 일련의 유사한 시스템에 대해 동일한 양자역학 계산을 실행하면 이러한 계산에는 시스템의 유사성으로 인해 중복 정보가 포함된다(동일한 계산이 약간 다른 입력에 대해 반복되어 상관 출력을 산출함). 예를 들어 분자 동역학 시뮬레이션(구조 변화)을 실행하고, 공통 scaffold 또는 치환기(화학 공간의 변화)를 갖는 일련의 분자에 대한 특성을 계산하는 것이다. 이러한 중복은 양자역학 계산 중 일부만 수행하고 나머지 양자역학 계산을 수행하여 나머지 시스템에 대한 대략적인 솔루션을 얻을 수 있다. 이 접근법의 유용성은 근사값 및 이를 구하는 데 드는 계산 비용으로 인해 발생하는 오류에 달려 있다.

Related work

ab initio 전위 에너지 표면(PES)의 보간은 컴퓨터가 처음 머신러닝에 사용된 지난 세기 중반까지 거슬러 올라가는 오랜 역사를 가지고 있습니다. 이 주제는 다른 분야에서 힘 필드의 매개 변수화, 사용된 기능적 형태의 주요한 차이점, Cheminformatics 과 관련이 있으며, 정량적 구조 활동/특성 관계(QSAR/QSPR)와 특히 더 관련이 있다 이것은 실험적 결과 측정치가 보간되고 더 조악한 시스템 표현에 사용된 참조값의 큰 불확실성과 사용법이 주된 차이점이다.

1990 년대 초에는 인공 신경망(ANN)이 단일 시스템의 PES 보간에 사용되기 시작했으며 이후, 대규모 분자 역학 시뮬레이션을 위한 강력한 도구로 개발되었다. 여기에 Shepard 보간, 3 차원 스플라인, 움직이는 최소 제곱 및 상징적 회귀까지 다른 많은 접근법이 사용되었다. 커널 학습 프레임 워크의 도입과 공식화를 예견하고 정규화와 같은 개념은 Rabitz 에 의해 초기에 사용되었다. 다른 시스템, 예를 들어 분자 특성 평가에 대한 양자역학 결과 간의 보간은 상관 에너지와 결합 해리 엔탈피를 예측하기 위해 ANN 을 사용하여 대략 10 년 후에 시작되었고 이후 지원 벡터 머신과 같은 다른 방법도 사용되었다. 지난 시간동안 대용량 데이터 세트의 공개와 적용을 발견했는데 발견된 예시로는 다극(multipoles), PES 보간과 같은 가우스 프로세스 회귀 (GPR)와 같은 커널 기반 머신러닝 방법의 도입이 있고 퀘도 자유 밀도 함수 이론에 대한 운동 에너지에 전자 밀도를 매핑하고 표면을 분할하는 전이 상태 이론을 최적화하는 것과 같은 QM/ML 모델의 새로운 응용 또한 발견된 예시들 중 하나이다.

QM/ML 모델의 가장 중요한 측면 중 하나는 분자 또는 주기적 시스템이 보간을 위해 수치로 표현되는 방식이다. 대칭 함수 ad hoc descriptor, 방사형 기저 함수의 푸리에 확장, 원자 위치의 부드러운 중첩, 그리고 Coulomb 행렬을 포함하여 다양한 표현이 사용되어 왔다.

Machine Learning

머신러닝의 핵심에서, 대부분의 머신러닝은 데이터 포인트 간의 보간 역할을 한다. 그렇다면, 머신러닝은 그저 그럴듯하게 끼워 맞춘 단어일까? 그렇지 않다. 머신러닝은 솔루션을 위해 다양한 공식 문제와 알고리즘을 포함한다. QM/ML 모델의 경우 가장 중요한 문제는 관리 대상 학습 문제인 회귀 분석이다. 화학적 맥락에서 관련된 또 다른 문제의 예로는 치수 감소, 감독되지 않은 학습 문제, 반응 좌표(집단 변수) 식별과 관련된 것이 있다.

회귀는 관리 대상 학습 문제이다. Inputs $x \in X$ 와 해당 출력 $y_i \in Y$ 로 구성된 n 개 관측치(교육 데이터) 집합이 주어진다면, 새로운 inputs x 에 대한 label y 를 예측할 수 있고 이것은 일반 다중 회귀 분석이다. 모든 관측을 기입한 것은 독립적이고 동일하게 분포된 것으로 가정되고 이러한 가정은 실무상 자주 위반된다. 유한한 표본의 가치로부터 기능을 배우는 것(많은 교육 데이터와 호환이 가능한 것)의 문제는 유일한 솔루션이 없고, 추가적인 가정이 이루어져야 한다. 머신러닝에서는 보통 부드러움을 가정하고 이는 정규화로 이어진다. 본질적으로 가장 간단한 것을 고르게 된다(Occam's razor).

Learning with kernels

이 튜토리얼의 초점을 주요점인 커널 기반 머신러닝 방법이 1990 년대에 도입된 이래에 널리 대중화되고 있다. 그들은 강력한 이론적 기초와 그들의 주된 경쟁상대인 인공신경보다 설정에 있어서 보다 더 쉬운 경향을 보인다. 인공신겨에 대한 소개는 후자의 기입될 내용을 참조하면 된다. 튜토리얼에 있어 필요한 유일한 개념은 여기 소개되어 있다. 레퍼런스를 참조하길 바란다.

커널 기반 머신러닝의 주요 개념은 시스템적인 측면에서 선형 머신러닝 알고리즘의 비선형성을 유도하는 것이다. 이것은 입력값을 보다 높은 차원으로 매핑하는 방식으로 수행되고, 공간 및 선형 알고리즘을 적용한다. 이러한 접근은 두가지 즉각적인 문제가 있는데, 올바른 매핑을 찾아내는 것과 계산의 복잡함이다.

한 예로 d 도의 모든 배열 단량체의 공간에 x 를 매핑하는, R 에서 R 로 매핑하는 것을 고려해보자. 매핑된 p 의 크기는 입력된 공간의 p 의 크기에 다항식적으로 의존한다. P 가 120 이고, d 가 3 인 특징공간의 크기는 이미 1728000 이다. 가우시안 커널을 기반으로 하는 다른 매핑들은 무한한 치수의 특징적인 공간으로 들어간다. 그러한 매핑에서 특징 공간의 명시적 계산은 계산적으로 가능하지 않다.

커널 트릭은 두 개의 관찰을 기반으로 이 문제에 대한 해결책을 제공한다. 첫째, 대부분의 선형 머신러닝 알고리즘은 입력 사이에 내부 제품만 사용하도록 다시 작성될 수 있다(이 알고리즘에는 표준, 각도 및 거리에 대한 정보, 즉 입력 간의 관계에 대한 정보가 포함되어 있다). 이렇게 하면 두 제품 간의 내부 제품 계산에 사용되는 기능 공간 벡터를 사용하는 임의 계산 문제가 줄어든다. 두번째로, 커널이라 불리는 기능은 입력 공간 벡터에서 작동하지만, 특징 공간의 내부 제품 평가와 동일한 결과를 산출한다. 즉, 기능 공간의 커널 기능을 평가해 기능 공간의 내부 제품 평가를 대체할 수 있다. 이 두 가지 관찰을 결합하면 특징 공간의 명시적 계산 문제를 잘 피할 수 있다.

Kernel functions

커널 함수는 동일한 결과를 제공하는 입력 공간의 계산을 통해 특징 공간의 계산을 대체할 수 있다. 이것은 길이, 각도 및 직교와 같은 기하학적 개념을 일반화하는 내부 제품을 통해 달성된다. 실제 벡터 공간 V 의 경우 V x V 함수는 해당 함수를 보유하는 경우(바로 이어지는 수식 참조), inner product 이다. V 의 쌍들을 inner product space 라고 부른다. 두 벡터(a, b)는 그들의 inner product 가 0 일때만 직교한다. Inner real space 에서 θ의 각도는 (1)의 수식과 같이 정의된다. <.,.>의 inner product 는 a^2 =<a,a>를 통해서 표준을 구성하는데 사용될 수 있다. 평행사변형에 한하여, 유클리드 표준은 그것의 본질을 만족하는 유일한 일반 Lp 이다.

커널은 일부 표준 공간 H 에서 내부 요소에 해당하는 함수다. 형식적으로, k : X*X – R 은 수식 (2)에 관한 맵 ⌀이 존재할 때만 커널이다. 노골적으로 ⌀나 H 가 충분한지는 꼭 알아야할 필요가 없다. 커널을 이용해서 고차원 기능 공간의 inner product 는 커널로 암시적으로 계산될 수 있다. 그리고 이것은 입력 공간의 값, 즉 피쳐 공간 치수로 인한 계산 복잡성 문제를 완화시켜준다. 커널은 양의 정의 속성으로 특정 지어진다. 실제 대칭 행렬 k 는 수식(3)일 경우에만 양수이다. K 는 c=0 일 때 평형이 발생하는 경우에만 양적으로 명확하다. 긍정적이고 엄격하게 확정된 행렬은 양의 준정부호와 양정치행렬이고도 하며, 문헌을 참조할 때는 이에 상응하는 주의를 기울여야 한다.

그램 행렬의 벡터 x1-xn 은 그들의 inner product K 이 관한 n*n 행렬 k 이다. 양정치 그램 행렬을 가진 행렬 k 는 양정 행렬이라고 불린다. Inner product 들은 수식 (4)로 인해서 양의 값이다. 그 반대로, 모든 양의 명확한 함수는 (커널 힐버트 공간과 무어 아론자른 정리를 재현함으로써) 일부 내부 제품 공간에서 내부 제품에 해당한다는 것을 보여줄 수 있다. 따라서 올바른 커널은 양의 정의로 특징지어진다. Eq(3) 이외의 매트릭스의 긍정적인 정의 기준은 f 및 모든 고유값이 음(양)이 아닌 경우에만 실베스터의 기준을 포함한다.

Specific kernels

선형 커널 k 는 동일한 input 과 특정 공간⌀(x)=x 에서 틀림없이 가장 간단한 커널이다. 선형 커널을 사용하면 원래 선형 알고리즘과 동등한 결과를 얻을 수 있다. 새로운 문제를 위해서, 이것은 시도된 첫번째 커널이다.

Δ Linear kernel yields the original mode 머신러닝 모델 f 를 기반한 선형 커널에서 원래 모델을 산출한다. Ai 가 회귀 계수인 경우, Xi ∈ R 는 교육 입력이며 X∈R 는 선형 커널의 경우 선형 회귀 모형 Eq8 이 생성됨을 나타낸다.

가우시안 커널(비선형 커널, 방사형 기본 함수 커널도 포함한다)은 비선형 커널 모델의 비선형 커널 모델(수식 5)을 위한 일반적인 기본 선택이다. Σ >0 이 존재하는 곳은 커널이 작동하는 길이 척도를 결정하는 초모수이다. 가우스 커널은 많은 문제에 대해 상당히 잘 수행되기 때문에 선형 커널을 따르기에 좋은 커널이다. 이것은 무한 치수 피쳐 공간으로 매핑된다. 가우스 커널의 동작을 이해하려면 제한 사례 수식(6)을 고려하면 된다.

첫 번째 경우, 모든 입력은 서로 직교하는 서로 다른 치수로 매핑되어 오버핏으로 이어진다. 두 번째 경우, 모든 입력이 단일 포인트로 매핑되어 언더스티어링으로 이어진다. Σ 의 중간값의 경우, 커널값은 [x-z], [x-z]를 위해 1 에 접근하는 0, [x-z]를 위한 0 에 의해서 결정된다. 따라서 입력 공간에 가까운 샘플은 피쳐 공간에서 상관 관계가 있는 반면 먼 샘플은 직교 하위 공간에 매핑된다. 이러한 방식으로 가우스 커널은 Σ 에

의존하는 크기값에 따라 크기가 달라지는 로컬 근사치로 표시될 수 있다. 세부적인 분석을 위해서는 ref75 를 참조하면 된다.

라플라시안 커널은 여러 연구에서 분자 특성을 예측하는 부분에서 가우스 커널보다 더 잘 수행된다. 가우스 커널처럼, 그것(수식 7)은 지수적이지만 유클리드 표준 대신에 1-norm Z 를 사용한다.

figure 4 는 선형, 가우스 및 라플라시안 커널의 그림을 원점에 배치된 기준 함수와 유사하게 단일 변수 x(왼쪽 열)의 함수 k(0, x)로 나타내므로 모양에 대한 약간의 직관을 제공한다. 또한 figure 4 는 공분산 함수(오른쪽 열)로 해당 커널을 사용하여 확률적 프로세스에서 랜덤하게 샘플링된 기능을 보여준다. 이것들은 각 커널을 모델링할 수 있는 기능의 형태에 대한 직관을 제공한다.

Linear regression

Multiple linear regression. d 치수의 선형 모형은 이러한 치수의 선형 조합에 의해 제시되며, 각각 회귀 계수 베타(수식 8)에 의해 가중됩니다.

단, Eq. (8)에는 편향 용어 1b 가 없으므로 원점을 통과하는 기능만 모델링할 수 있다. 바이어스 용어를 사용하는 것은 입력과 라벨, 즉 x 와 y 대신 X-Xi 와 Y-Yi 를 중심으로 작업하는 것과 같다. 지금부터 이것은 한 케이스로 가정한다. (training set 는훈련 전에 뺄 수 있기 때문에 일반적인 손실 없이 예측을 위해 추가될 수 있다.)

이상적으로는 일반화 오류, 즉 새로운 입력의 평균 오류를 최소화하는 계수 베타를 찾는 것이 바람직하다. 단, 표본의 분포는 일반적으로 알려져 있지 않으며, 한정된 교육 세트(수식 9)에만 액세스할 수 있다. 여기서 X 는 xi 행이 있는 입력 행렬이다. 수식 9 는 특정 오류 또는 손실(즉, 제곱 오류)을 구성한다. 결과적인 볼록 최적화 문제 (수식 10)는 수식 11 의 역이 존재할 경우 경사도를 0 으로 설정하여 해결된다. 이 접근 방식의 한 가지 문제는 교육 데이터가 정확히 장착되어 있다는 것이다. 즉, 라벨 노이즈(다른 구현 또는 설정으로 인한 계산된 속성의 수치 편차 또는 양자역학 방법의 정확도를 초과하는 의미 없는 마지막 비트)도 정확하게 결합된다. 이러한 작은 차이를 정확하게 적용하면 큰 계수(bi)가 발생하여 교육 데이터에서 거의 취소되지만 새로운 입력에 큰 오류가 발생하는 경우가 많다. 이는 오버핏의 한 형태이다.

Ridge regression. 릿지 회귀 분석은 중첩을 방지하기 위해 정규화가 추가된 선형 회귀 분석입니다. 정규화는 회귀 계수를 서로로, 0 으로 줄여 앞에서 설명한 과적 효과를 완화시킵니다. 편향은 증가하지만 분산은 감소시킨다. 릿지 회귀 분석에서는 수식 10 에 페널티 용어를 추가하여 λ>0 범위 내의 정규화의 강도를 결정하는 하이퍼 파라미터, 수식 12 을 야기한다. 계수 벡터 베타 2 의 일반모델은 f 모델의 평활성과 복잡성과 관련이 있으며, λ 값이 클수록 더 부드럽고 단순한 모델이 된다. 수식 11 의 도출과 유사하며, 베타에 대한 해결은 (베타 =수식)을 산출하며, 여기서 도출되는 동일한 행렬을 나타낸다. 수식 11 과는 달리, 이의 반대는 언제나 λ>0 일때에만 존재한다. 수식 13 은 모델의 모든 매개변수 b 를 결정하지만, 하이퍼 파라미터 k 는 결정하지 않는다.

Δ Ridge regression solution.

Kernel ridge regression

커널 트릭을 Rackmount 회귀 분석에 적용하면 KRR(kernel rade regression)이라는 비선형 버전의 커널에 의해 비선형성이 결정된다. 알고리즘은 한 번만 도출하고 구현하면 되며, 그 후에 다른 커널과 함께 사용하면 효과적으로 다른 비선형 버전의 능선 회귀 분석을 제공할 것이다.

커널 트릭은 입력의 내부 곱에만 의존하는 선형 ML 알고리즘에 적용될 수 있다. "커널화 된"다른 알고리즘의 예로는 지원 벡터 머신, 주성분 분석, 가우스 프로세스 회귀 및 부분 최소 제곱이 있다.

커널 학습 알고리즘은 특징 공간 H 에서 암묵적으로 수행된다. 특징 벡터 2H 는 직접 접근할 수 없기 때문에(내부 제품 만이 가능함), 커널 모델은 수식 8 에서와 같이 차원에 대한 합으로 표현되지 않는다. 그러나 훈련 예제에 대한 요약으로서, 대수학 법칙(수식 14)은 이것이 항상 가능함을 보장한다.

대수 정리는 이것이 항상 가능하다는 것을 보장한다. 직관적으로 H 의 차원성은 높을 수 있지만, 솔루션은 계획된 훈련 데이터의 유한 범위에 존재하므로 유한한 표현이 가능합니다. 대응하는 볼록 최적화 문제는 H 가 H 에서의 f 의 일반, 즉 특징 공간에서의 선형 융기 회귀 모델의 복잡도이고, K∈ R…는 트레이닝 샘플들 간의 커널 행렬이다. 이전과 마찬가지로 그래디언트를 0 으로 설정하면 회귀 계수에 대한 분석 솔루션이 생성된다.

그림 5 는 길이 스케일 하이퍼 파라미터 r 의 역할을 나타내는 가우시안 커널을 사용한 KRR 모델의 예를 보여준다. 비록 σ 이 Eq.에서 정규화 항과 직접적으로 관련이 없지만, 수식 15 에서, 예측의 부드러움을 제어하고, 또한 효과적으로 규칙화한다.

Implementation

커널 기반 학습 알고리즘의 구현은 KRR 의 예에서 논의됩니다.

Basic considerations. 수식 14 와 17 을 고려하자. KRR 모델을 훈련하는 데 사용된 모든 정보는 훈련 데이터 간의 커널 평가 행렬 K 에 포함되어 있다. 마찬가지로, 새로운 입력을 예측하는 데 필요한 모든 정보는 훈련 대 예측 데이터의 커널 행렬에 포함된다. 따라서 커널은 알고리즘이 데이터를 보는 "렌즈"로 묘사될 수 있다. 커널 매트릭스는 커널 학습 알고리즘 구현을 위한 자연스러운 인터페이스 선택이다. 이것은 데이터에 필요한 필수 정보가 정확하게 포함되어 있다. 개별 커널 평가와 달리 빌딩 블록으로 커널 매트릭스에 의존하는 또 다른 이점은 이 전략이 매트릭스 및 벡터 제품과 같은 최적화된 절차의 호출이 거의 없는 고급 인터프리터 언어와 잘 작동한다는 것이다. 인터프리터 인텐시브 코드에서 많은 호출보다 바람직하다. 수치 라이브러리와 함께 하위 레벨 언어 (예:Fortran, C)를 사용하는 경우에도 이와 유사한 주장이 적용된다.

Kernel ridge regression. 식 (17)에서 회귀 계수 알파는 (K+ kI)…가 대칭이고 양의 확정적인 방정식의 선형 시스템을 푸는 것으로 얻어진다. Rasmussenand William 가 수치 안정성을 위해 제안한 한 가지 방법은 Cholesky 분해와 관련한 K=..I = U+ U 의 식을 사용하는 것입니다(여기서 U 는 위쪽 삼각형). 하나는 방정식의 두 선형 시스템 UUa=y 을 풀어내는데 처음으로 y 에 관한 식을 풀고 그 다음에 베타에 관한 식을 푼다. UT 는 하부 삼각형이고 U 는 상부 삼각형이므로, 이는 순방향 및 역방향 치환이라 불리는데 이 데이터에 대해 단지 2 개의 직선 패스를 필요로 한다.

UT 베타=y 의 경우, 수식 18 을 얻는다. Ua=베타의 경우는 순서가 바뀐다. 모델이 훈련되고 나면, 예측은 수식 14 에 의해 만들어진다. 새로운 input X 에 대한 예측은 계수의 벡터와 해당 커널 평가의 벡터 사이의 내적이다. 여러 예측 샘플 x1-xn 행렬을 예측하는 것은 x∈R 이 훈련 대 인풋이 Lij 인 커널 예측 훈련과 x1-xa 의 열을 사용하여 편리하게 표현되고 효율적으로 계산될 수 있다. 알고리즘 1 은 Cholesky 분해를 사용하여 KRR 모델을 훈련하기위한 의사 코드를 제공한다.

What about other methods?

양자역학/머신러닝 모델링에 대한 또 다른 인기있는 선택 사항은 가끔씩 Kriging 이라고 불리는 Gaussian process regression (GPR)과 인공 신경 네트워크(ANN)다. GPR 은 빈도주의적 KRR 의 Bayesian 등가물이며 예측 분산과 같은 다른 종소리와 호각이 다르지만 동일한 예측을 제공한다. GPR 에서, 커널은 입력들 사이의 공분산의 역할을 한다. 그림 6 은 GPR 의 기본 개념을 보여준다. KRR 및 GPR 을 포함하는 수식 14 형식의 모든 모델은 비모수 모델이며, 그들의 수는 트레이닝 데이터와 함께 증가한다. ANN 은 파라메트릭 방법이며 (우선적으로 네트워크 아키텍처가 선택된다면) 고정된 수의 매개 변수를 갖는다. ANN 에 대한 자세한 분석은이 자습서의 범위를 벗어난다. 자세한 내용은 Jorg Behler 의 양자역학/머신러닝 용 ANN 에 대한 리뷰 및 자습서를 참조하면 된다.

Model selection and performance estimation

Model selection. 머신러닝 모델의 하이퍼 파라미터를 어떻게 선택할까? 더 일반적으로, 어떻게 서로 다른 머신러닝 모델 중에서 하나를 선택할까? 데이터 세트가 주어진 후보 모델 세트에서 모델을 선택하는 문제를 모델 선택이라고 한다. KRR 에 대하여, 파라미터 a 가 결정된다 수식 17 을 통해, 훈련 세트가 주어지고 이것은 커널 k 와 정규화 하이퍼 파라미터 k 와 커널의 임의의 하이퍼 파라미터를 선택하고 데이터 셋에 따라 최적의 선택을 한다. 일반적인 지침 원리는 Occam'ㄴ Razor.이다. 목적을 위해 동일한 성능을 가진 모델 중에서 가장 단순한 것이 선호되어야 한다고 명시되어 있다. 모델 선택에 대한 많은 접근법이 여기에서 사용되며, 선택 기준으로서의 성능 추정에 초점을 둔다. 구체적인 예로서, 비슷한 성능이 주어지면 이 튜토리얼에 제시된 모델에 대해 (i) Gaussian 과 Laplacian 커널에 걸친 선형 커널, (ii) 라플라시안 커널 위의 가우스 (Gaussian) (iii) 높은 정규화 강점, (iv) (ii) - (iv)의 추정기의 평활도가 견적가에 대한 매끄러움으로 선호된다.

Estimating model performance. 이상적으로는 한가지 요소는 새로운 데이터에 대한 모델의 오류를 알고 싶어한다. 통계적 학습 이론에서 이것은 모델 f 의 위험성에 의해 측정된 다통계적 학습 이론에서 이것은 모델 f 의 위험에 의해 측정된다. 여기서 P 는 입력과 레이블의 합동 분포이고, L: Y*Y-R 은 예측의 오차를 측정하는 손실 함수이다. 수식 20 은 f 의 예상 오류입니다. 불행히도, P 는 거의 알려져 있지 않고 R 은 경험에 의한 위험으로 한정된 훈련 데이터로부터 추정되어야 한다.

훈련 세트의 경험적 오차를 최소화하는 모델 f 는 보통 나쁜 선택이다. f 가 충분히 복잡하다면 훈련 데이터를 학습하고 새로운 데이터로 일반화하지 않는다(룩업 테이블과 유사하게 훈련 세트 및 기타 모든 입력의 실패를 완벽하게 재생산할 것이라고 예상됨). 이를 오버 피팅이라고 하며, 카운터 오버 피팅에 대한 한 가지 접근법은 모델 복잡성에 관한 용어를 추가하여 정규화를 사용하는 것이다. 이는 앞에서 소개 한 정규화 된 회귀이다.

Choosing hyperparameters. 앞의 논의는 수식 13 과 수식 17 의 회귀 계수 β 와 α 의 선택에 대한 이론적 근거를 제공한다. 그렇다면 이들을 어떻게 선택할까? 간단한 전략은 홀드 아웃 세트 (유효성 검사 세트)를 사용하는 것이다. 이것은 시작 부분에 설정되어 있고 훈련에 사용되지 않는 예제 집합{xi,yi}^n 이다. 하이퍼 파라미터θ에 대한 가능한 값의 세트 ⌀가 주어지면, 그것은 최적의 하이퍼 파라미터를 찾기 위해 홀드 아웃 세트에 대한 경험적 위험을 최적화한다. 충분하지 않은 데이터를 사용하여 충분히 큰 홀드 아웃 세트를 설정하지 못하면 교차 검증 또는 부트 스트래핑과 같은 방법을 사용할 수 있다. 이렇게 하면 교육 데이터가 다른 방식으로 반복적으로 분할되어 데이터를 효과적으로 재사용 할 수 있다. 홀드 아웃 세트, 교차 유효성 검사, 부트 스트랩 및 유사한 통계 유효성 검사 절차의 사용을 이해하는 핵심은 모델 유효성 확인의 황금률을 따르는 것이다. 훈련 중에 사용된 데이터에 대한 모델 성능을 예측하지 않도록 하자.

L, f, ⌀에 따라 수식 23 을 풀면 어려울 수 있다. 그리드 검색은 구현이 간단하지만 계산적으로 요구되며 최대 2 ~ 3 개의 하이퍼 파라미터로 제한된다. 즉, 각각의 하이퍼 매개 변수에 대한 일련의 시험 값이 주어지면 이 값의 모든 조합에 대해 그리드 ⌀1* ⌀2...을 설정한다. 각 그리드 엔트리에 대해, 해당 하이퍼 파라미터를 사용하여 트레이닝 세트에서 모델을 훈련시키고 홀드 아웃 세트의 성능을 평가한다. 이에 따라 최적의 홀드 아웃 설정 성능을 가져오는 하이퍼 매개 변수가 선택된다. 그리고 실제로는 대수 그리드가 자주 사용된다.

Error statistics. 어떤 통계가 보고되어야 할까? 기본 선택은 RMSE 가 수식 10 및 수식 15 의 최적화 기준의 일부인 평균 제곱 오차(RMSE) 및 평균 절대 오차(MAE, 허용되지 않는 에러)를 포함하고 MA 는 모델의 오차의 평균 크기를 포함한다. 유용하고 널리 사용되는 또 다른 통계치는 피어슨의 상관 계수(제품 - 순간 상관 계수)의 제곱이다. A1…ak, b1…bk 을 두 개의 랜덤 변수 A, B 에서 추출한 샘플이라고 가정하다. 여기서 var 와 covar 는 분산과 공분산을 나타낸다. 선형 모델 (한 변수는 레이블 y 역할을 하고 다른 변수는 f(x)예측 역할을 함)의 경우 R^2∈[0,1]은 모델에서 설명하는 레이블 분산의 비율로 해석할 수 있다. 단일 숫자로 전체 데이터 세트의 성능을 요약하면 항상 정보가 손실된다. 따라서 모델 개발 중에 산점도 (예 : 그림 9 참조) 또는 히스토그램의 형태로 오류 분포를 살펴 보는 것이 좋다.

Predicting Atomization Energies

이 튜토리얼의 실제적인 부분에서는 다음과 같은 시나리오를 가정해보자. 7k 개의 작은 유기 분자의 데이터 세트가 주어지며 이론의 밀도 기능 수준에서 원자화 에너지를 예측하라는 요청을 받았다고 가정해보자. 당신의 컴퓨팅 리소스를 사용하면 1k 참고 계산을 할 수 있다. 그렇다면 전체 데이터 세트의 원자화 에너지를 얼마나 정확하게 예측할 수 있을까?

이 섹션에서는 이전에 소개된 방법론을 사용하여 이 질문에 대답하는 방법을 단계별로 설명한다. 설정은 머신러닝을 사용하여 원자화 에너지를 예측하는 최근의 연구를 모델로 한다. 보조 정보에는 (i)분무 에너지가 있는 7k 개의 작은 유기 분자의 데이터 세트로, 훈련 세트, 모델 및 성능 추정치의 선택을 후 향적으로 평가, (ii) Wolfram 언어로 작성된 제시된 알고리즘의 기본 참조 구현, (iii) 연습 문제에 대한 해결책을 제공하는 참고요소가 있다. 이것들을 모두 함께 구현하면 실험을 당장 시작할

수 있다. 다른 프로그래밍 환경을 선호하는 경우, 제시된 알고리즘을 구현하는 것은 간단하고 좋은 연습을 제공해야할 것이다.

Dataset

데이터 세트에는 생성된 데이터베이스 GDB 에서 추출한 7k 개의 작은 유기 분자가 포함되며, 여기에는 힘 필드 완화 기하학 및 DFT 원자화 에너지가 있다(자세한 내용은 부록 참조). 분자는 원소 H, C, N, O, S 로 구성되며, 최대 7 개의 비 -H 원자가 있다(표 3). 머신러닝 모델은 힘 필드 최소치에서 분자의 기하학을 DFT 최소치의 에너지로 매핑한다. 따라서 머신러닝 모델은 기하 구조의 변화를 보완해야 한다. 이것이 필요한 이유를 보려면 예측에 대한 입력으로 DFT 최소값을 사용하는 것을 고려해보자. 새로운 분자의 경우, 구조를 완화하기 위해 DFT 계산을 수행해야 한다. 이것이 머신러닝 모델이 대체해야 할 계산식이다.
Δ 이 자습서의 지원 정보를 다운로드하여 데이터 세트를 얻자. xyz 파일에서 분자 구조 및 원자화 에너지를 로드하도록 하자. 파일은 확장된 XYZ 형식이다(그림 7).
Δ 표 3 을 재생산하기 위해 각 분자의 비 -H 원자 수를 세보자.
Δ 일부 분자를 시각화해보자.

Training set. 첫 번째 단계는 원자화 에너지를 "계산"하기 위해 1k 분자를 선택하는 것이다(여기서는 데이터 세트가 제공되기 때문에, 이 에너지 시나리오를 모델로 사용하도록 허용된 레이블이다). 큰 균질 데이터 세트의 경우, 무작위로 트레이닝 세트를 그리는 것으로 충분하다. 그러나 표 3 은 제공된 데이터 세트가 비 수소 원자의 수와 관련하여 동질성이 없음을 보여준다. 신뢰할 수 있는 예측을 위해 4 개 이하의 비-H 원자를 갖는 분자의 예가 너무 적기 때문에 이들 모두가 훈련 세트에 포함된다(양자역학 계산을 수행하는 것과 같다). 나머지 941 개의 분자는 5 개 이상의 비-H 원자를 갖는 모든 분자에서 무작위 추출되며, 원자화 에너지와 관련이 있는 것으로 알려진 크기로 층을 형성한다. 트레이닝 세트에 포함되지 않은 모든 분자는 예측 세트에 할당된다.

Δ Create a training set. 4 개 이하의 비-H 원자를 갖는 모든 k 분자를 선택하자. 나머지 분자를 원자 수로 정렬하고 모든 7102k/10002k 번째 원자를 선택하여 적절하게 반올림하자. 선택된 분자를 트레이닝 세트에 할당하고 다른 모든 분자는 예측 세트에 할당하자.

Hold-out set. 모델 선택에서 커널 및 하이퍼 파라미터의 선택을 위해, 트레이닝 세트를 분할할 필요가 있다. 이는 교차 검증을 사용하는 등 여러 가지 방법으로 수행할 수 있다. 여기에는 남아있는 것들을 적절한 훈련 세트로 사용하여 100 개의 분자의 홀드 아웃 세트를 따로 설정하는 간단한 접근법을 제공하기에 충분한 데이터가 있다. 홀드 아웃 세트는 성능을 예측하는 데 사용되며, 6k 예측 세트에 대한 프록시 역할을 하며 가능하면 가능한 (분배의 관점에서) 가깝게 닮아야 한다. 그러므로, 그것은 4 개 이하의 비-H 원자를 갖는 분자를 포함하지 않아야 하고, 원자의 수에 의해 계층화되어야 한다.

Δ Create a hold-out set. 훈련 세트에서 원자 수로 계층화된 5 개 이상의 비-H 원자를 갖는 100 개의 분자를 선택하고 이를 홀드 아웃 세트에 지정하자. 나머지 900 개의 분자가 적절한 훈련 세트를 구성한다.

Representation

쿨롱 행렬은 요소 유형과 분자 내 내부 거리를 인코딩하는 간단하면서도 효과적인 행렬 표현이다. 이는 분무 에너지의 예측을 위한 머신러닝 알고리즘에 분자를 수치로 나타내기 위해 참조[45,58-62]에서 사용되었다. Zi 는 원자 i 의 원자 번호(핵전하)이고, Ri 는 원자 단위(Bohr 반경 a0)의 위치이다. M 은 대칭이며 분자 내에 원자가 있는 행과 열과 동일한 개수를 가진다. 직관적으로 M 의 각 행(및 열)은 원자에 해당하며 분자의 나머지와 상호 작용하는 방식을 인코딩한다. 대각선을 벗어난 원소들은 내부의 거리를 스케일링하여 기하학을 인코딩하는 반면, 주 대각 원소는 거리가 원자가 0 인 곳에서 원소 유형을 인코딩하는 자유 에너지에 적합하다.

양자역학/머신러닝 표현은 이러하다. (i) 재산, 특히 번역, 회전 및 핵 순열을 변화시키지 않는 변환에 대한 불변성, (ii) 다른 변형에 관한 변형, (iii) 고유함, (iv) 연속적이며 이상적으로는 분화가 가능함. 수식 26 은 분자, 즉, f{(Zi; Ri)}는 M 에서부터 병진 및 회전까지 재구성될 수 있지만, 원자의 재 색인 생성에는 변하지 않는다. 한 가지 해결책은 행과 열을 동시에 순열로 바꾸어 행 기준을 내림으로써 쿨롱 행렬을 정렬하는 것이다. 이렇게 하면 핵 순열에 대한 불변성이 보장되지만 정렬이 변경되는 불연속성이 도입된다.

Δ Compute Coulomb matrices. 데이터 세트의 모든 분자에 대해 수식 26 을 계산하자. 각 행렬에 대해 행의 표준을 계산하고 행과 열을 동시에 치환하여 정렬하여 표준에 따라 내림차순으로 정렬해보자. 그리고 각각의 행렬을 오른쪽과 아래쪽에 0 으로 채워보자. 이와 같이 하면 모든 행렬은 같은 크기를 갖는다. 23*23 가 데이터 세트의 분자 당 최대 원자 수이다. 각 행렬의 대각선을 포함하여 비 중복 삼각형 부분만 유지하고 276 차원의 열 벡터로 다시 정렬하자.

Model

Basic model. 가우시안 커널을 더해 KRR 을 사용하여 쿨롱 행렬을 소개한 출판물의 모델로 시작해보자. 주어진 하이퍼 파라미터들에 대해, 이것은 커널 매트릭스 K(L, 알고리즘 1), 및 성능 평가를 요구한다.

Δ KRR with Gaussian kernel for given hyperparameters 하이퍼 매개 변수 λ 및 σ에 대한 값을 선택하자. X1…x900 은 적절한 훈련 세트의 쿨롱 행렬을 나타내고, y1…y900 은 상응하는 원자화 에너지를 나타낸다. 수식 5 를 사용하여 K 를 계산해보자. 그런 다음 알고리즘 1 을 사용하고 K; y, k 를 사용하여 회귀 계수 알파를 계산하자. 홀드 아웃 세트 입력 x1…x100 에 대한 예측을 위해 커널 행렬 L 을 계산한 다음, 알고리즘 1 을 사용하여 예측 ~ f 를 얻어보자. 마지막으로 ~ f 와 ~y 를 사용하여 성능 통계 RMSE, MAE, 1-R^2 를 계산해보자.

Grid search 이제 그리드 검색 이제 기본 모델 구축이 완료되었고 그리드 검색을 수행하여 두 개의 하이퍼 매개 변수 λ 및 σ을 결정해보자. 시퀀스 a, a+ c, a+ 2c, … ,

b 를 나타내는 a, b, c 가 있는 곳에서 적절한 훈련 세트의 모델을 훈련시키고 홀드 아웃 세트에서 그 성능을 평가하자.

Δ KRR with Gaussian kernel and grid search for hyperparameters 수식 27 에 하이퍼 매개 변수의 각 조합에 대한 모델을 작성해보자(이전 연습과 동일). 그림 8 을 재현하기 λ 및 σ의 함수로 성능을 추정한다. 최적의 하이퍼 매개 변수를 결정하면 된다.
Results 그림 8(제일 윗 줄)은 하이퍼 매개 변수의 함수로서 홀드 아웃 설정 성능을 보여준다. λ=10^-6.5 와 σ=273 의 하이퍼 파라미터의 최적 선택은 표 4 의 수치 결과를 산출해준다. 이는 전체 데이터 세트를 예측하기위한 성능 추정이기도 하다. 모델이 나머지 6k 분자에서 실제로 얼마나 잘 수행되었는가?
Δ 최적화된 하이퍼 매개 변수 값을 사용하여 전체 1k 트레이닝 세트에서 모델을 훈련시키고 예측 세트에서 6k 분자를 예측한다.
표 4 에서 예측 집합의 성능은 보류 집합에서 예측 한 것보다 낫다. 새로운 데이터에 대한 오류를 과소 평가하는 것이 아니라 일반적으로 과대 평가하는 것이 바람직하다. 그림 9 (왼쪽)는 모든 예측의 산점도를 보여준다.
라플라시안 커널 가우스 커널을 사용하는 모델의 성능은 원래의 출판물에서 보고된 성능에 가깝다. 이것이 더 잘할 수 있을까? 커널 학습 모델을 개선하는 편리한 방법은 문제에 보다 적합한 다른 커널을 사용하는 것이다. 후속 출판물에서 Laplacian 커널은 종종 Coulomb 행렬을 표현으로 사용하는 모델에서 더 잘 수행되었다.
Δ Gaussian 커널 대신 수식 7 의 Laplacian 커널을 사용하여 그리드 탐색을 포함하여 이전 모델 구축 연습을 반복해보자. 동일한 교육, 예측, 적절한 교육 및 보류 설정을 사용하면 된다.
결과값 RMSE, MAE 및 R2 의 경우, 하이퍼 매개 변수의 최적 조합은 λ=10^-12 와 σ=10^3.6 이며 성능은 그림 8 (하단), 그림 9 (오른쪽) 및 표 4 에 나와 있다. 가우시안 커널과 비교시 상당한 향상, 그리고 하이퍼 파라미터, 특히 k 에 대한 민감도 감소, 홀드 아웃 세트의 성능 추정과 예측 데이터의 성능 간의 더 긴밀한 일치에 주목해보자.

What Next?

요약하자면 튜토리얼의 목적은 독자에게 커널 기반 기계 학습에 대한 기본적인 이해와 전산 양자 화학과 함께 사용하는 것을 말한다. 핵심 아이디어는 머신러닝을 사용하여 양자역학 참조 계산을 보간하여 수십 배에 이르는 상당한 계산 비용을 절감하는 것이다. 이를 위해, 결정적 요인은 보간 오차의 제어다. 즉, 머신러닝 근사는 양자역학 기준에 대한 대리로서 작용할 수 있을 정도로 충분히 근접해야 한다.
추가로 읽을만한 간단한 자습서는 필요에 따라 포괄적이지 않을 수 있다. 적용 범위 또는 동적 모델 재교육(즉석에서 학습하는 것)과 같은 많은 고급 주제가 다루어지지 않았다. 머신러닝/양자역학 모델에 관한 현대 연구의 횡단면은 이 자습서가 출판된 International Journal of Quantum Chemistry 의 특별 호에서 찾을 수 있다. 커널 능선 회귀에 대한 자세한 내용은 Hastie 외 저서 5.8 절을 참조하면 된다. 커널 학습 방법에 대한 좀 더 심층적인 처리를 위해서는 주제에 관한 최근 교과서, Scholkopf and Smola 의 고전 서적 또는 Hofman 의 검토를 참조하면 된다.