머신러닝 모델 종류

특징과 훈련 방법을 중심으로

(보고서용)

,

2018045214

최준희

List of contents

1. 지도 학습 모델

지도 학습이란, 데이터와 라벨 (데이터의 정보)를 함께 제공하여 학습을 시키는 방법이다. 지도 학습 중에서도 입력 받은 데이터가 어떤 분류에 해당되는지 예상하여 나누는 방식을 분류 (Classification) 이라 하며, 연속적인 데이터를 이용해 그에 대한 함수를 추론해 앞으로의 값을 예측하는 방식을 회귀 (Regression) 이라 한다.

1-1. Regression

회귀 본석을 할 때, **가장 먼저 고려해야 하는 것은 선형성과 비 선형성을 구분하는 것이다.** 선형성 이란 하나의 함수가 여러 개의 파라미터의 성질을 만족시키는 것을 말한다. 여기서 주의할 점은 선형과 비 선형을 구분할 때 **독립 변수와 종속 변수의 관계를 기준으로 나누면 안되며**, 선형과 비 선형을 **결정하는 것은 변수가 아니라, 회귀 계수(=Model parameter)**이다. 일반적으로 사용하는 회귀 모델은 선형 회귀 모델이며, 비 선형 회귀 모델의 예로는 딥러닝이 있다.

선형 회귀 모델은 회귀 계수간의 관계가 비교적 직관적이기 때문에 각 조건의 영향력을 해석하기가 비교적 쉬우며, **모든 조건을 선형 결합으로 표현**하기 때문에 모델링 대상의 데이터가 선형 결합으로 이루어지지 않은 경우엔 정확한 모델을 만들 수 없는 표현력의 한계가 있다.

두 번째로 고려해야 하는 것은 종속 변수이다. 종속 변수가 하나인 모델은 univariate regression model 이라 부르며, 종속 변수가 2개 이상인 경우에는 multivariate regression model 이라 부른다.다변량 회귀 모델은 종속 변수간에 서로 상관관계가 있는지, 종속변수가 서로 다른 종속변수의 독립변수 역할을 수행하는등 종속 변수의 관계와 조건에 따라 다양한 회귀 모델로 나뉘며 비선형 데이터셋 에서도 사용할 수 있는 모델도 있다.

1-1-1 Linear Regression

선형 회귀 모델은 변수들의 모델 파라미터의 가중치, 그리고 편향을 이용하여 예측하는 모델이다. 교재에서는 선형 회귀의 훈련을 가장 보편적으로 사용되고 있는 RMSE (평균 제곱근 오차)를 이용하며, RMSE를 최소화 하는 모델 파라미터를 설정해야 한다. 여기서, RMSE 보다 MSE (평균 제곱 오차)를 최소화 하는 것이 같은 결과를 내면서 더 간단하다. 오차를 최소화 하는 모델 파라미터를 찾기 위한 해석적인 방법이 있는데, 이를 정규방정식이라 한다.

(교재 내용 정리) 정규방정식을 사용하여 비용함수를 최소화 하는 파라미터를 계산한다. Numpy의 선형대수 모듈 (np.linalg)에 있는 inv() 함수를 이용해 역행렬을 계산하고, dot() method 를 통해 행렬 곱셈을 한다. ScikitLearn에서 선형 회귀를 수행하는 것은 더 간단한데, sklearn의 linear_model 모듈에 있는 LinearRegression 함수를 이용하면 된다. (?) 이 함수는 비용함수를 최소화 하는 모델 파라미터를 유사역행렬 (Moore-Penrose)을 사용하여 값을 구한다.

유사역행렬은 특잇값 분해 (singular value decomposition, SVD)라 부르는 표준 행렬 ㅂㄴ해 기법을 사용하여 계산된다. 이 기법은 정규방정식을 계산하는 것보다 훨씬 효율적이며, 데이터셋이 극단적인 경우도 처리할 수 있다. Scikitlearn의 LinearRegression 클래스의 복잡도는 O(n^2)이다.

1-1-2. Gradient Descent

선형 회귀 모델에서 사용하는 알고리즘 중 경사 하강법은 비용 함수를 최소화 하기 위해 반복하여 모델 파라미터를 조정한다. 학습률과 손실함수의 기울기를 이용하여 가중치를 업데이트 하면서 계산하는 것이다. 손실 함수의 기울기가 0이 되는 지점이 비용 함수의 최솟값이 된다. 이 때 중요한 파라미터는 스텝의 크기로, learning rate 하이퍼 파라미터로 결정하는데 이 값을 적절히 조절해야 한다. 모든 비용 함수가 2차 함수의 형태를 띄는 것은 아니기 때문에 이 파라미터가 너무 크거나 작으면 기울기의 최솟값을 구하기 어렵기 때문이다. 일반적으로 (데이터 셋이 극단적인 경우나 특성의 스케일이 매우 다른 경우가 아닌) MSE 비용 함수는 Convex function이기 때문에 지역 최솟값이 없고, 하나의 전역 최솟값을 가지게 되기 때문에 경사 하강법이 전역 최솟값에 가깝게 접근할 수 있음을 보장한다. 적절한 학습률을 찾기 위해 그리드 탐색을 사용하는데, 시간이 너무 오래 걸리는 모델을 제어하기 위해 반복 횟수를 제한한다. 일반적인 경사 하강법 (Batch Gradient Descent)은 훈련 세트가 커지면 적절히 반복 횟수를 제한해도 매우 오랜 시간이 걸리기 때문에, 확률적 경사 하강법을 사용하기도 한다. 매 스텝에서 한 개의 샘플 데이터를 선택하고, 그 샘플에 대한 기울기를 계산한다. 매 반복마다 다뤄야 할 데이터의 수가 현저히 줄어들기 때문에 시간은 절약되지만, 배치 경사 하강법에 비해 비용함수의 최솟값에 접근이 불안정하다.

정리하면, 일반적인 데이터 셋 에서는 배치 경사 하강법을 통해 서서히 비용함수의 최솟값에 접근해 나가는게 유리하며, 비용함수가 매우 불규칙한 데이터 셋에서는 확률적 경사 하강법이 더 유리하다.

1-1-3. Polynomial Regression

가지고 있는 데이터가 매우 복잡한 형태를 가지고 있다면, 비선형 데이터를 학습하는 데 선형 모델을 사용할 수도 있다. 각 특성의 거듭제곱을 새로운 파라미터로 추가하고, 그 특성을 포함한 데이터셋에 선형 모델을 훈련시키는 것이다. 즉 선형 모델의 훈련에 새로운 데이터를 추가해 나가는 것이 다항 회귀 방식이다. 즉, 파라미터가 여러 개일 때 다항 회귀를 통해 파라미터 간의 관계를 찾을 수 있다. 예를 들어 두 개의 파라미터 a, b가 있을 때 degree = 3을 적용하면 a^2, a^3, b^2, b^3 뿐만 아니라, ab, a^2b, ab^2와 같은 교차항도 추가가 되기 때문이다. 단, 파라미터의 수가 너무 많으면 당연히 교차항은 기하 급수적으로 늘어나기 때문에 주의해야 한다.

(교재 내용) 그림 4-14를 보면 고차 다항 회귀 모델과 선형 회귀 모델의 결과를 볼 수 있다. 다항 회귀를 사용하게 되면, 고차항 일수록 훈련 데이터셋에 과대적합하는 경향이 있고, 이는 다르게 보면 고차항일수록 더 정확해진다는 것이다. (과하게 정확하면 과대적합) 이를 조절하기 위해 교차 검증을 사용했는데, 또 다른 방법으로는 학습 곡선을 살펴보는 것이다. 이 그래프는 훈련 세트와 검증 세트의 모델 성능을 훈련 세트 크기의 함수로 나타낸다. 이 그래프를 생성하기 위해서는 훈련 세트에서 크기가 다른 서브 세트를 만들어 모델을 여러 번 훈련시키면 된다. 학습 곡선에 대한 2가지 그래프 유형은 교재를 참고하면 된다.

1-1-4 Ridge, Lasso, elasticnet regression

과대 적합을 감소시키는 방법은 모델을 적절히 규제하는 것이다. 다항 회귀 모델을 규제하는 간단한 방법은 다항식의 차수를 감소시키는 것이다. 선형 회귀 모델에서는 보통 모델의 가중치를 제한하는 방식을 사용한다.

- Ridge Regression (Tikhnonov Regularization)

릿지 회귀는 규제가 추가된 선형 회귀다. 비용 함수에 규제항이 추가되며, 이는 학습 알고리즘을 데이터에 맞게 조절하는 것 뿐만 아니라, 모델의 가중치가 가능한 작게 유지되도록 해준다. 규제항은 훈련하는 중에만 추가된다. 이때 하이퍼 파라미터 알파의 값을 통해 규제 정도를 조절할 수 있는데 알파의 값이 0이면 릿지 회귀는 선형 회귀와 완전히 동일해 진다.

- LASSO Regression (Least Absolute Shrinkage and Selection Operator)

라쏘… 는 어려워서 이번 에 정리하지 못했습니다 ㅠㅠ

- Elastic net

엘라스틱 넷은 릿지 회귀와 라쏘 회귀를 절충한 모델이다. 규제항은 릿지와 회귀의 규제항을 단순히 더하고, 혼합 정도는 혼합 비율 r을 사용해 조절한다. R이 0이면 릿지 회귀와 같고, R이 1이면 라쏘 회귀와 같아진다.

일반적으로 평범한 선형 회귀는 규제가 없으므로, 약간의 규제가 있는 것이 대부분의 경우에 좋으므로 최대한 사용을 피해야 한다. 보통 릿지 회귀를 사용하지만, 쓰이는 특성이 몇 개 정도로 생각되면 라쏘나 엘라스틱 넷을 사용하는 것이 좋다. 이 모델들은 불필요한 특성의 가중치를 0으로 만들어 준다. 특성 수가 훈련 샘플 수보다 많거나, 특성 몇 개가 강하게 연관되어 있을 때는 보통 라쏘보다는 엘라스틱 넷을 사용하는 것이 좋다.

1-1-5. Early stopping

검증 에러가 최솟값에 도달하면 훈련을 중지하는 것. 그림 4-20을 보면 된다.

1-1 요약

데이터셋을 분석하여 각 변수들간의 상관관계를 파악한 뒤, 적절한 전처리를 거치고, 몇 몇 모델을 가중치를 조절하면서 비용 함수를 최소화 할 수 있는 모델과 가중치를 통해 accuracy가 높은 결과를 구해야 한다…

Logistic Regression

Softmax Regression (multinomial logistic regression)