Machine learning

Hyeonho Lee

2018년 11월 5일

Contents

0 machine learning

인공지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 하는 알고리즘과 기술을 개발하는 분야를 말한다. 가령, 기계 학습을 통해서 수신한 이메일이 스팸인지 아닌지를 구분할 수 있도록 훈련할 수 있다.

기계 학습의 핵심은 표현(representation)과 일반화(generalization)에 있다. 표현이란 데이터의 평가이며, 일반화란 아직 알 수 없는 데이터에 대한 처리이다. 이는 전산 학습 이론 분야이기도 하다. 다양한 기계 학습의 응용이 존재한다. 문자 인식은 이를 이용한 가장 잘 알려진 사례이다.

알고리즘 유형은 지도학습, 비지도학습, 강화학습으로 크게 나뉘고 지도학습과 비지도학습의 중간인 준지도학습도 있다.

0.1 Supervised learning(지도학습)

훈련 데이터(Training Data)로부터 하나의 함수를 유추해내기 위한 기계 학습(Machine Learning)의 한 방법이다. 훈련 데이터는 일반적으로 입력 객체에 대한 속성을 벡터 형태로 포함하고 있으며 각각의 벡터에 대해 원하는 결과가 무엇인지 표시되어 있다. 이렇게 유추된 함수 중 연속적인 값을 출력하는 것을 회귀분석(Regression)이라 하고 주어진 입력 벡터가 어떤 종류의 값인지 표식하는 것을 분류(Classification)라 한다. 지도 학습기(Supervised Learner)가 하는 작업은 훈련 데이터로부터 주어진 데이터에 대해 예측하고자 하는 값을 올바로 추측해내는 것이다.

이 목표를 달성하기 위해서는 학습기가 "알맞은" 방법을 통하여 기존의 훈련 데이터로부터 나타나지 않던 상황까지도 일반화하여 처리할 수 있어야 한다. 사람과 동물에 대응하는 심리학으로는 개념 학습(Concept Learning)을 예로 들 수 있다.

0.2 Unsupervised learning(비지도학습)

기계 학습의 일종으로, 데이터가 어떻게 구성되었는지를 알아내는 문제의 범주에 속한다. 이 방법은 지도 학습(Supervised Learning) 혹은 강화 학습(Reinforcement Learning)과는 달리 입력값에 대한 목표치가 주어지지 않는다.

자율 학습은 통계의 밀도 추정(Density Estimation)과 깊은 연관이 있다. 이러한 자율 학습은 데이터의 주요 특징을 요약하고 설명할 수 있다.

자율 학습의 예로는 클러스터링(Clustering)을 들 수 있다. 또 다른 하나의 예로는 독립 성분 분석(Independent Component Analysis)이 있다.

0.3 Semi-supervised learning(준지도학습)

기계 학습(Machine Learning) 의 한 범주로 목표값이 표시된 데이터와 표시되지 않은 데이터를 모두 훈련에 사용하는 것을 말한다. 대개의 경우 이러한 방법에 사용되는 훈련 데이터는 목표값이 표시된 데이터가 적고 표시되지 않은 데이터를 많이 갖고 있다. 이러한 준 지도 학습은 목표값이 충분히 표시된 훈련 데이터를 사용하는 지도 학습과 목표값이 표시되지 않은 훈련 데이터를 사용하는 자율 학습 사이에 위치한다. 많은 기계 학습 연구자들이 목표값이 없는 데이터에 적은 양의 목표값을 포함한 데이터를 사용할 경우 학습 정확도에 있어서 상당히 좋아짐을 확인하였다. 이러한 훈련 방법이 사용되는 이유는 목표값을 포함한 데이터를 얻기 위해서는 훈련된 사람의 손을 거쳐야 하기 때문이고 그 비용이 감당할 수 없을만큼 클 수 있기 때문이다. 따라서 그러한 경우 준 지도 학습을 사용하여 결과를 향상 시킬 수 있다.

이러한 준 지도 학습의 예로는 상호 훈련을 들 수 있다. 이것은 두개 이상의 학습기 각각이 예제를 통해 훈련되는 방법이며 학습기가 사용하는 예제의 특성은 각각 다르며 독립적이다.

그 이외의 접근으로는 특징점과 목표값의 결합 분포를 모델링 하는 것이다. 목표값이 없는 데이터에 대해서는 목표값을 '잃어버린 데이터' 로 처리한다. 이러한 방법에서는 EM 알고리즘을 모델의 우도를 최대화 하기 위해 사용한다.

1 Statistical learning

1.1 통계학습이란?

통계학습은 f를 추정하는 일련의 기법들을 말하는 것이다. f를 추정하는 데 필요한 몇가지 중요한 이론적 개념과 얻어진 추정치들을 평가하기 위한 도구들을 소개해 본다.

1.1.1 f를 추정하는 이유는?

f를 추정하고자 하는 2가지 주요한 이유는 예측과 추론이다.

1. 예측 (Predict)

많은 경우, 입력 X는 쉽게 얻을 수 있지만 출력 Y는 쉽게 얻을 수 없다. 여기서, 오차항은 평균이 영이므로 다음 식을 사용하여 Y을 예측 할 수 있다.

$$\hat{Y} = \hat{f}(X)$$

여기서 $\hat{f}$는 f에 대한 추정을 나타내고 $\hat{Y}$는 Y에 대한 예측 결과를 나타낸다. $\hat{f}$는 보통 블랙박스로 취급된다. 이유는 $\hat{f}$가 Y에 대한 정확한 예측을 제공한다면 그것의 정확한 형태에 대해서는 통상 신경쓰지 않기 때문이다.

Y에 대한 예측인 $\hat{Y}$의 정확성은 축소가능 오차와 축소불가능 오차 이 2가지에 달려있다. 일반적으로 $\hat{f}$는 f를 완벽하게 추정하지 못하며, 이러한 부정확성으로 인해 오차가 발생될 것이다. 이런 오차는 축소가능한 오차이다. $\hat{Y} = f(x)$의 형태를 취하더라도 예측한 값은 여전히 어떤 오차를 가지고 있을 수 있고, 그 이유는 Y도 또한 ε의 함수이고, ε은 X를 사용하여 예측할 수 없기 때문이다. ε과 관련된 변동성도 예측의 정확성에 영향을 미친다. 이것은 축소불가능 오차로 알려져 있다. f를 아무리 잘 추정하더라도 ε에 의해 도입된 오차를 줄일 수 없기 때문이다.

$$E(Y - \hat{Y})^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + Var(\varepsilon) = reducible + irredcible$$

2. 추론 (Inference)

어떤 설명변수들이 반응변수와 관련되어 있는가? 많은 경우, 사용할 수 있는 설명변수들 중 아주 작은 일부만이 Y와 실질적으로 관련되어 있다. 많은 가능한 변수들 중에서 일부 중요 설명변수를 찾아내는 것은 응용에 따라서는 아주 유용할 수 있다.

반응변수와 각 설명변수 사이의 상관관계는 무엇인가? 어떤 설명변수들은 그 값이 증가함에 따라 Y의 값도 증가한다는 점에서 Y와 양의 상관관계를 가지고 있을 수 있다. 다른 설명변수들은 상반된 상관관계를 가질 수도 있다. f의 복잡도에 따라 반응변수와 주어진 설명변수 사이의상관관계는 다른 설명변수들의 값에 따라 변할 수도 있다.

Y와 각 설명변수의 상관관계는 선형 방정식을 사용하여 충분히 요약될 수 있는가? 또는 이 상관관계는 더 복잡한가? 역사적으로 f를 추정하는 대부분의 방법들은 선형 형태를 취한다. 어떤 경우에는 이러한 가정이 합리적이거나 심지어 바람직하다. 그러나, 실제 상관관계는 보통 더 복잡하며 선형모델은 입력과 출력변수들 사이의 상관관계를 정확가게 표현하지 못할 수 있다.

어떤 모델링은 예측과 추론 둘 다를 위해 수행될 수 있다. 예를 들어, 부동산 시장에서 범죄율, 지역, 강과의 거리, 공기의 청정도, 학교, 지역의 소득 수준, 집의 크기 등과 같은 입력에 집값을 연관시키고자 할 수 있다. 이러한 경우, 개별 입력 변수들이 어떻게 가격에 영향을 미치는지 관심이 잇을 수 있다. 만약 집의 전망이 강을 내려다 볼 수 있다면 그 집의 가치가 얼마나 더 올라가는가? 이것은 추론 문제이다. 아니면, 단순히 주어진 집의 특징에 대해 그 집의 가치를 예측하는 데 관심이 있을 수 있다. 즉 이 집은 과소 또는 과대 평갈괴었는가? 이것은 예측 문제이다.

최종 목적이 예측, 추론 또는 이 둘을 결합한 것인지의 여부에 따라 f를 추정하는 데 다른 방법들을 사용하는 것이 적절하다. 선형모델들은 비교적 간단하고 해석가능한 추론을 할 수 있지만, 몇몇 다른 기법들만큼 정확한 예측을 할 수 없을 수 있다. 반대로, 몇가지 고도의 비선형적인 기법들은 잠재적으로 Y에 대해 아주 정확한 예측을 제공할 수 있지만 추론을 더욱 어렵게 만드는 이해하기 어려운 모델을 초래한다.

1.1.2 어떻게 f를 추정하는가?

1. 모수적방법 (Parametric methods)

모수적 방법은 2단계로 된 모델 기반의 기법이다.

1) 먼저, f의 함수 형태 또는 모양에 대해 가정한다. 예를 들어, 아주 단순하게 f는 X에 대해 선형적이라고 가정한다.

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

f가 선형이라는 가정이 있으면, f를 추정하는 문제는 크게 단순화된다. 임의의 p차원 함수$f(X)$를 추정해야 하는 대신에, p+1개의 계수 $\beta_0, \beta_1, ..., \beta_p$만 추정하면 된다.

2) 모델이 선택된 후 훈련 데이터를 사용하여 모델을 적합하거나 훈련시키는 절차가 필요하다. 선형모델의 경우 파라미터 $\beta_0, \beta_1, ..., \beta_p$를 추정해야 한다. 즉, 다음을 만족하는 파라미터들의 값을 찾고자 한다.

$$Y \approx \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

3) 모델의 적합에 가장 일반적으로 사용되는 기법은 최소제곱이다.

2. 비모수적방법(Non-parametric methods)

비모수적 방법은 f의 함수 형태에 대해 명시적인 가정을 하지 않는다. 대신에 너무 거칠거나 왔다갔다 하지 않으면서 데이터 포인트들에 가능하면 가까워지는 f의 추정을 얻으려고 한다. 이러한 접근법은 모수적 방법에 비해 주요한 장점이 있을 수 있다. 즉 f의 함수 형태에 대한 가정을 하지 않아도 되므로 더 넓은 범위의 f형태에 정확하게 적합될 가능성이 있다. 어떠한 모수적 방법이라도 f를 추정하는 데 사용된 함수 형태가 실제 f와 아주 많이 다를 수 있으며, 이 경우 결과 모델은 데이터에 잘 적합되지 않을 것이다. 이에 반해, 비모수적 방법은 f의 형태에 대한 어떠한 가정도 하지 않기 때문에 이러한 위험을 완전히 회피한다. 하지만, 비모수적방법은 중요한 단점이 있다. 이 방법은 f를 추정하는 문제를 작은 수의 파라미터 추정 문제로 축소하지 않으므로, f에 대한 정확한 추정을 얻기 위해서는 아주 많은 수의(모수적 기법에서 보통 필요로 하는 것보다 훨씬 더 많은 수의)관측치가 필요하다.

1.1.3 예측 정확도와 모델 해석력 사이의 절충(Trade-off)

선형회귀는 비교적 유연하지 않은 기법이다. 직선이나 평면같은 선형함수들만 생성할 수 있기 때문이다. 박판 스플라인 같은 방식은 매우 유연하다고 할 수 있다. 왜냐하면 f를 추정하는 데 훨씬 넓은 범위의 가능한 함수 형태를 생성할 수 있기 때문이다.

그렇다면 왜 유연한 기법 대신에 더 제한적인 방법을 선택하여 사용하는가? 좀 더 제한적인 모델을 선호할 수 잇는 몇가지 이유가 있다. 만약 주 관심사가 추론이면, 제한적인 모델이 훨씬 더 해석하기 쉽다. 선형모델은 Y와 $X_1, X_2, X_3, ..., X_p$ 사이의 상관관계를 이해하는 것이 아주 쉽기 때문에 좋은 선택이다. 그러나 스플라인, 부스팅 방법과 같은 매우 유연한 기법들은 f추정이 복잡하게 되어 어떤 개별 설명변수가 반응변수와 어떻게 연관되는지 이해하기 어려울 수 있다.

즉 추론이 목적일 때는 비교적 덜 유연한 통계학습방법을 사용하는 것이 명백히 장점이 있다고 할 수 있다. 하지만 어떤 경우에는 예측에만 관심이 있고 예측 모델의 해석력에는 관심이 없다. 예를들어 주식가격을 예측 하는 알고리즘을 개발하려고 할 때, 해석력은 중요하지 않고, 오직 예측에 대한 정확도라고 할 수 있다. 이러한 경우 적용 가능한 가장 유연한 모델을 사용하는 것이 최선이라고 예상 할 수 있다. 그러나 놀랍게도 이것이 항상 맞는 것은 아니다. 우리는 종종 덜 유연한 방법을 사용하여 더 정확한 예측을 얻을 것이다. 처음에는 직관에 반하는 것처럼 보이는 이러한 현상은 아주 유연한 방법들의 잠재적인 과적합과 관련이 있다.

1.1.4 지도학습과 비지도학습

반응변수를 설명변수에 관련시키는 모델을 적합하고자 하며, 목적은 미래 관측(예측)에 대해 반응변수를 정확하게 예측하거나 반응변수와 설명변수들 사이의 상관관계(추론)을 더 잘 이해하는 것이다.

이에 반해 비지도 학습은 모든 관측 $i = 1, .., n$에 대해 측정값 x_i를 관측하지만 연관된 반응변수 측정값 y_i가 없는 좀 더 어려운 상황을 설명한다. 예측할 반응변수가 없으므로 선형모델을 적합하는 것은 불가능하다. 어떤 의미에서는 뚜렷한 방향성 없이 분석이 이루어지며, 분석을 지도할 수 있는 반응변수가 없으므로 비지도 학습이라고 한다.

많은 문제들이 자연스럽게 지도학습 또는 비지도학습 패러다임에 속한다. 하지만 어떤 경우에는 분석을 지도적 또는 비지도적으로 해야 하는지 명확하지 않다. 예를 들어, n개 관측치의 집합이 있다고 해보자. m<n개의 관측치에 대한 설명변수의 측정값과 반으변수 측정값을 가지고 있다. 나머지 n-m개의 관측치에 대해서는 설명변수 측정값은 있지만 반응변수 측정값은 없다. 이 때 설명변수들은 비교적 쉽게 측정할 수 있지만 대응하는 반응변수는 수집하기가 쉽지 않은 경우가 발생한다. 이럴 때 준지도학습문제라고 한다.

1.1.5 회귀와 분류문제

변수는 양적 변수 또는 질적 변수로 구분할 수 있다. 반응변수가 양적인지 또는 질적인지에 다라 통계학습방법을 선택하는 경향이 있다. 반응변수가 양적인 경우 선형회귀를 사용하고 질적인 경우 로지스틱 회귀를 사용할 수 있다. 하지만 설명변수가 양적인지 또는 질적인지 여부는 일반적으로 덜 중요하다고 생각된다.

1.2 모델의 정확도 평가

항상 궁금한 부분 중 하나가 하나의 최고의 방법 대신 왜 이렇게 많은 통계학습 기법을 소개하는 것이 필요한가? 통계 분야에서 가능한 모든 자료에 대해 어떤 한 방법이 다른 방법들보다 지배적으로 나은 경우는 없다. 특정 자료에 대해 어떤 한 방법이 가장 좋은 결과를 줄 수 있지만, 비슷하지만 다른 자료에 대해서는 어떤 다른 방법이 더 나은 결과를 제공할 수 있다. 그러므로 임의의 주어진 자료에 대해 어느 방법이 최고의 결과를 제공하는 지 결정하는 것은 중요한 일이다. 최고의 기법을 선택하는 것이 실제로 통계학습을 수행하는 데 있어서 가장 어려운 부분 중의 하나이다.

1.2.1 적합의 품질 측정

주어진 자료에 대한 통계학습방법의 성능을 평가하기 위해서는 이 방법에 의한 예측이 관측된 데이터와 실제로 얼마나 잘 맞는지 측정하는 방법이 필요하다. 즉 주어진 관측치에 대해 예측된 반응 값이 관측치에 대한 설제 반응값에 얼마나 가까운지를 수량화하는 것이 필요하다. 이러한 회귀 설정에서 가장 일반적으로 사용되는 측도는 평균제곱오차이다.

$$MSE(Mean\ Squared\ Error) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

통계학습방법을 훈련시키는 데 사용되지 않은 사전에 본적이 없는 검정 관측치(test observation)을 생각해야 한다. 이를 (x_0, y_0)이라고 하자. 이 때, 가장 낮은 훈련 MSE가 아니라 가장 낮은 검정 MSE를 제공하는 방법을 선택하고자 한다. 다시 말하면, 아주 큰 수의 검정 관측치가 있다면, 이들 검정 관측치 (x_0, y_0)에 대한 평균 제곱예측오차인 다음 식을 계산 할 수 있다.

$$Ave(y_0 - \hat{f}(x_0))^2$$

통계학습방법의 유연성이 증가함에 따라 훈련 MSE는 단조감소하지만 검정 MSE는 U모양을 보인다. 이것은 가지고 있는 자료와 사용되는 통계방법에 관계없이 성립하는 통계학습의 기본적인 성질이다. 모델의 유연성이 증감함에 따라 훈련 MSE는 감소할 것이지만 검정 MSE는 그렇지 않을 수도 잇따. 주어진 방법이 훈련 MSE는 작지만 검정 MSE는 큰 겨로가를 제공할 때 데이터를 과적합한다고 한다. 이러한 과적합은 통계학습 절차가 훈련 데이터에서 패턴을 찾는 데 지나치게 집중하여 알려지지 않은 함수 f의 실제 성질에 의한 것이 아니라 단순히 우연에 의한 어떤 패턴을 찾을 수도 있기 때문에 발생한다. 훈련 데이터를 과적합할 경우, 통계방법이 훈련 데이터에서 찾은 패턴이라는 것이 검정 데이터에서는 존재하지 않을 것이므로 검정 MSE가 아주 클 것이다. 과적합의 발생여부에 관계없이 훈련 MSE는 거의 항상 검정 MSE보다 작을 것으로 예상ㄷ뇐다. 왜냐하면 대부분의 통계학습방법은 직접적으로 또는 간접적으로 훈련 MSE를 최소화하려고 하기 때문이다.

실제로 훈련 MSE는 비교적 쉽게 계산할 수 있다. 그러나, 보통은 사용가능한 검정 데이터가 없으므로 검정 MSE를 추정하는 것은 상당히 어렵다. 앞의 세 가지 예에서 보듯이, 검정 MSE가 최소가 되는 모델에 대응하는 유연성 수준은 자료에 따라 상당히 다를 수 있다. 검정 MSE가 최소로 되는 지점을 실제로 추정하는 데 사용될 수 있는 기법 중 하나는 교차검증이다.

1.2.2 편향-분산 절충

검정 MSE곡선이 U모양을 보이는 것은 통계학습방법의 두 가지 상충되는 성질 때문이다.주어진 값 x_0에 대한 expected 검정 MSE는 항상 세가지의 기본적 수량인 $\hat{f}(x_0)$의 분산, $\hat{f}(x_0)$의 제곱편향, 그리고 오차항 ε의 분산의 합으로 분해된다는 것을 보여줄 수 있다.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$$

$E(y_0 - \hat{f}(x_0))^2$은 기대 검정 MSE에 대한 정의로, 아주 큰 수의 훈련자료들을 사용하여 f를 반복적으로 추정하고 각각을 x_0에서 검정했을 경우 얻어지는 검정 MSE의 평균을 말한다. 또한 기대검정오차를 최소화하기 위해서는 낮은 분산과 낮은 편향을 동시에

달성하는 통계학습방법을 선택해야 한다. 분산은 본질적으로 음수가 아니고 제곱편향도 또한 음수가 아니다. 그러므로 기대 검정 MSE는 축소불가능 오차인 $Var(\varepsilon)$보다 작을 수 없다.

통계학습방법에서 분산과 편향은 무엇을 의미하는가? 분산은 다른 훈련자료를 사용하여 추정하는 경우 $\hat{f}$이 변동되는 정보를 말한다. 훈련자료는 통계학습방법을 적합하는 데 사용되므로, 다른 훈련자료를 사용하면 $\hat{f}$이 달라질 것이다. 그러나 이상적으로는 f에 대한 추정이 훈련자료에 따라 너무 많이 변동되지 않아야 한다. 하지만, 분산이 높으면 훈련 데이터의 변화가 작아도 $\hat{f}$는 크게 변할 수 있다. 일반적으로, 통계학습방법의 유연성이 높을수록 분산도 더 높다.

편향은 실제 문제를 훨씬 단순한 모델로 근사시킴으로 인해 발생되는 오차로, 극도로 복잡할 수도 있다. 예를 들어, 선형회귀는 Y와 $X_1, X_2, ..., X_p$사이에 선형 상관관계가 있다고 가정한다. 실제 문제가 이러한 단순한 선형 상관관계를 가질 가능성은 거의 없으므로 선형회귀를 수행하면 f 추정에 틀림없이 어떤 편향이 발생할 것이다. 실제 f는 상당히 비선형적이므로, 아무리 많은 훈련 관측치가 있어도 선형회귀를 사용해서는 정확한 추정을 할 수 없을 것이다. 하지만 실제 f가 선형적이라면 데이터만 충분히 있으면 선형회귀로 정확하게 추정할 수 있을 것이다. 일반적으로는 유연성이 높은 방법일 수록 편향이 적다.

원칙적으로 유연성이 높은 방법을 사용할수록 분산이 증가하고 편향은 감소할 것이다. 이러한 분산과 편향의 상대적 변동율이 검정 MSE가 증가 또는 감소하는지를 결정한다. 통계방법의 유연성을 증가시킴에 따라 편향은 처음에는 분산의 증가보다 더 빠르게 감소하는 경향이 있다. 하지만, 어떤 지점에서 유연성 증가는 편향에 거의 영향이 없지만 분산은 크게 증가시키기 시작한다. 이럴 경우, 검정 MSE는 증가한다.

편향, 분산, 검정 MSE 사이의 관계를 편향-분산 절충이라고 한다. 통계학습방법이 검정자료에 대해 좋은 성능을 내려면 분산뿐만 아니라 제곱편향도 낮아야 한다. 이것을 절충(trade-off)라고 한다. trade-off는 편향은 낮지만 분산이 높거나, 분산이 낮지만 편향이 높은 방법을 얻는 것은 어렵지 않다. 분산과 제곱편향이 둘 다 낮은 방법을 찾는 것이 어려운 것이다.

1.2.3 분류 설정

위의 모델 정확도는 회귀를 중심으로 다뤘다. 그러나 분류의 경우에도 trade-off개념은 사용된다. 분류문제에서 추정치 $\hat{f}$의 정확도를 수량화하는 가장 흔한 기법은 훈련오차율로, 이것은 $\hat{f}$을 훈련 관측치에 적용할 경우 발생하는 오차율이다.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y_i})$$

여기서 $I(y_i \neq \hat{y_i})$는 지시변수로 $y_i \neq \hat{y_i}$이면 1이고 $y_i = \hat{y_i}$이면 0이다. 위의 식은 훈련오차율이라고 할 수 있다. 회귀와 마찬가지로 관심있는 것은 훈련에 사용되지 않았던 검정 관측치에 분류기를 적용해 얻은 오차율이다.

$$Ave(I(y_0 \neq \hat{y_0}))$$

검정오차율은 위의 식과 같다.

1. Bayes Classifier

조건부확률로, 관측된 설명변수 벡터 x_0가 주어진 경우에 대해 Y=j일 확률이다. 이 단순한 분류기를 베이즈 분류기라고 한다. 오직 두 개의 반응변수 값, 이를테면 클래스 1 또는 클래스 2만 가능한 20클래스 문제에서 베이즈 분류기는 $Pr(Y = 1|X = x_0) > 0.5$이면 클래스 1, 그렇지 않으면 클래스 2를 예측하는 것에 해당된다.

베이즈 분류기가 제공하는 검정오차율은 가능한 검정오차율 중 가장 낮은 값이고, 이것을 베이즈 오차율이라고 한다. 베이즈 분류기는 항상 최대가 되는 클래스를 선택하므로 $X = x_0$에서의 오차율은 $1 - max_j Pr(Y = j|X = x_0)$일 것이다. 일반적으로, 전체 베이즈 오차율은 다음식과 같다.

$$1 - E(max_j Pr(Y = j|X))$$

여기서 기대값은 가능한 모든 X 값에 대해 확률을 평균한 것이다.

2. KNN(K-Nearest Neighbors), K-최근접이웃

이론상 질적 반응변수는 베이즈 분류기를 사용하여 예측하는 것이 항상 가장 좋다. 그러나, 실제 데이터에서는 주어진 X에 대한 Y의 조건부 분포를 모르므로 베이즈 분류기를 계산 할 수 없다. 그러므로 베이즈 분류기는 다른 방법들을 비교하는 데 사용되는 달성할 수 없는 표준 역할을 한다. 많은 기법들이 주어진 X에 대한 Y의 조건부분포를 추정하여 가장 높은 추정확률을 가지는 클래스로 관측치를 분류하고자 한다. 이러한 방법 중 하나가 KNN이다. 양의 정수 K와 검정 관측치 x_0에 대해 KNN분류기는 먼저 훈련 데이터에서

x_0에 가장 가까운 K개 점$(= N_0)$을 식별한다. 그 다음에, 클래서 j에 대한 조건부확률을 반응변수 값이 j인 N_0 내 점들의 비율로 추정한다.

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

마지막으로, 베이즈 규칙을 적용하여 검정 관측치 x_0을 확률이 가장 높은 클래스에 할당한다. 아주 단순한 기법임에도 불구하고 KNN은 보통 최적의 베이즈 분류기에 놀라울 만큼 가까운 분류기를 제공할 수 있다. 단, K의 선택은 얻어지는 KNN분류기에 큰 영향을 미친다. K=1일 때, 결정경계는 지나치게 유연하고 베이즈 결정경계와 맞지 않는 데이터 패턴들을 발견한다. 이것은 편향은 낮지만 분산은 높은 분류기에 해당한다. K가 증가할수록 이 방법은 덜 유연해지고, 선형에 가까운 결정경계를 제공한다. 이것은 분산은 낮지만 편향이 높은 분류기에 해당한다. 회귀문제와 같이 훈련오차율과 검정오차율 사이에 강한 상관관계는 없다. K=1일 때, 훈련오차율은 0이지만, 검정오차율은 상당히 높을 것이다. 일반적으로, 유연성이 높은 분류방법을 사용할 수록 훈련오차율은 감소할 것이지만 검정오차율은 그렇지 않을 수도 있다.

회귀와 분류 설정에서 올바른 수준의 유연성을 선택하는 것은 통계학습방법의 성공에 아주 중요하다. 편향-분산 절충과 검정오차의 U모양은 유연성 수준 선택을 어렵게 만들 수 있다.

2 Classification

3 Resampling methods

3.1. Cross-validation(교차-검증)

주어진 통계학습방법과 연관된 검정오차를 추정하여 성능을 평가하거나 적절한 수준의 유연성을 선택하는 데 사용될 수 있다. 모델의 성능을 평가하는 과정은 model assessment(모델평가)로 알려져 있고, 모델에 대한 적절한 수준의 유연성을 선택하는 과정은 model selection(모델 선택)이라고 알려져 있다.

3.1.1 Validation set approach(검증셋기법)

매우 단순한 전략이다. 데이터를 임의로 두 부분으로 나누는데, 훈련셋과 검증셋(or hold-out set)으로 나눈다. 모델적합은 훈련셋에 대해 수행하고 적합된 모델은 검증셋의 관측치들에 대한 반응변수 값을 예측하는 데 사용된다. 결과의 검증셋 오차율(양적 반응변수의 경우 전형적으로 MSE를 사용하여 평가)은 검정오차율에 대한 추정치를 제공한다.

3.1.2 LOOCV(Leave-one-out cross-validation)

위의 검증셋기법의 단점을 해결한 방법이다. 검증셋기법과 마찬가지로 LOOCV는 관측치셋을 두부분으로 분할한다. 하지만 비슷한 크기의 두 서브셋(subset)을 만드는 대신에 하나의 관측치 (x_1, y_1)이 검증셋으로 사용되고 나머지 관측치 $\{(x_2, y_2), ..., (x_n, y_n)\}$은 훈련셋을 구성한다. 통계학습방법은 $n-1$개 훈련 관측치에 적합되고 제외된 관측치에 대한 예측값 $\hat{y_1}$은 x_1값을 사용하여 구한다. (x_1, y_1)은 적합과정에 사용되지 않았으므로 $MSE_1 = (y_1 - \hat{y_1})^2$은 검정오차에 대한 거의 편향되지 않은 추정치를 제공한다. 그러나 MSE_1은 비록 검정오차에 대해 편향되어 있지 않지만 하나의 관측치 (x_1, y_1)에 기초하므로 변동이 커서 좋지 않은 추정치이다.

이 절차를 반복하여 수행할 수 있다. 검증데이터로 (x_2, y_2)를 선택하고, 나머지 $n-1$개 관측치 $\{(x_1, y_1), (x_3, y_3), (x_n, y_n)\}$에 대해 통계학습절차를 훈련하여 MSE_2를 계산한다. 이런식으로 n번 반복하여 n개의 제곱오차를 얻고, 검정 MSE대한 LOOCV 추정치는 n개 검정오차 추정치들의 평균이다.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

LOOCV는 검증셋기법에 비해 2가지 장점이 있다. 첫째로 편향이 작다는 점이다. 둘째로는 훈련셋/검증셋 분한ㄹ의 임의성 때문에 적용할 때마다 다른 결과를 제공하는 검증셋 기법과 대조적으로, LOOCV는 여러 번 수행해도 항상 동일한 결과가 얻어질 것이다. 즉, 훈련셋/검증셋 분할에 임의성이 없다.

3.1.3 k-fold교차검증

LOOCV의 대안으로 k-fold CV가 사용딘다. 이 기법은 관측치셋을 임의로 크기가 거의 같은 k개 그룹으로 분할한다. 첫 번째 fold는 검증셋으로 취급하고 적합은 나머지 k-1개 fold에 대해 수행된다. 그다음에 평균제곱오차 MSE_1이 검증셋 fold의 관측치에 대해 계산된다. 이 절차는 k번 반복되며 매번 다른 그룹의 관측치들이 검증셋으로 취급된다. 이 과정으로 k개 검정오차 추정치 $MSE_1, MSE_2, MSE_3, ..., MSE_k$가 얻어진다. k-fold CV 추정치는 이 값들을 평균하여 계산된다.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{n} MSE_i$$

3.1.4 k-fold 교차검증에 대한 편향-분산 절충

k-fold CV는 LOOCV에 비해 $k < n$의 계산상의 장점이 있다. 그러나 계산상의 문제와 별개로, 덜 분명하지만 잠재적으로 더 중요한 k-fold CV의 장점은 LOOCV보다 검정오차율을 보통 더 정확하게 추정한다는 것이다. 이것은 편향-분산 절충과 관계가 있다.

검증셋 기법은 전체 관측치의 절반만 포함되는 훈련셋을 통계학습방법을 적합하는 데 사용하기 때문에 검정오차율을 과대추정할 수 있다고 하였다. 이 논리에 따르면, LOOCV는 거의 편향되지 않은 검정오차 추정치를 제공할 것이다. 왜냐하면, 각 훈련셋은 전체 데이터셋의 관츨치 수와 거의 같은 n-1개의 관측치를 포함하기 때문이다. 반면에 k-fold CV는 편향은 중간 수준이 될 것이다.

왜냐하면, 각 훈련셋은 LOOCV 기법보다는 작지만 검증셋 기법보다 훨씬 많은 $(k-1)n/k$개의 관측치를 포함하기 때문이다. 그러므로 편향 감소의 측면에서 보면 LOOCV가 k-fold CV보다 명백히 낫다.

그러나 추정절차에서 고려해하 하는 것이 편향만 있는 것이 아니라 분산도 있다. LOOCV는 k-fold CV보다 큰 분산을 가지는 경향이 있다. 그 이유는 n개 적합된 모델의 결과를 평균하는데, 적합된 모델 각각은 거의 동일한 관측치들로 구성된 훈련셋을 사용하여 구해진다. 그러므로 적합된 모델의 결과들은 서로 높은 (양의) 상관성이 있다. 반대로 k-fold CV의 경우 k개 적합된 모델의 결과를 평균하는데, 각 모델의 훈련셋 사이에 겹치는 부분이 적어 적합된 모델의 결과들은 서로 덜 상관되어 있다. 상관성이 높은 값들의 평균은 상관성이 상대적으로 낮은 값들의 평균보다 분산이 크기 때문에 LOOCV의 검정오차 추정치는 k-fold CV의 추정치보다 분산이 더 큰 경향이 있다.

3.2 bootstrap

붓스트랩은 추정량 또는 통계학습방법과 연관된 불확실성을 수량화하는 데 광범위하게 사용될 수 있는 아주 강력한 통계적 도구이다. 아주 간단한 예로 붓스트랩은 선형회귀적합에서 계수의 표준오차를 추정하는 데 사용될 수 있다. 또한 여러 맥락에서 사용되는데, 가장 일반적으로는 파라미터 추정의 정확도 또는 주어진 통계학습방법의 정확도를 측정하는 데 사용된다.

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r`=1}^{B}\hat{\alpha}^{*r`})}$$

이것은 원래의 데이터셋으로부터 추정된 $\hat{\alpha}$의 표준오차에 대한 추정치로 사용된다. 가설검정을 하거나 매트릭을 계산하기 전에 random sampling을 적용하는 방법을 일컫는다.(복원추출허용) 확률변수의 정확한 분포를 모르는 경우,측정된 통계치의 신뢰도를 가늠할 방법이 없기 때문에, 이 때 bootstraping을 사용하여 측정된 n개의 데이터 중에서 중복을 허용하여 m개를 뽑고, 그들의 평균을 구하기를 여러번 반복하여 평균의 분포를 구하고, 이를 통해 평균의 신뢰구간을 구할 수 있다.

4 Support Vector Machines

90년대 컴퓨터 과학 분야에서 개발되어 널리 알려진 분류기법이다. SVM은 다양한 설정에서 잘 동작한다는 것이 밝혀졌으며, 흔히 최상의 분류기 중 하나로 간주된다. SVM은 maximal margin classifier(최대 마진 분류기)라고 불리는 단순하고 직관적인 분류기를 일반화한 것이다. 최대 마진 분류기는 비록 우아하고 단순하지만 유감스럽게도 대부분의 데이터셋에 적용될 수 없다. 왜냐하면 이 분류기는 클래스들이 선형 경계에 의해 구별될 수 있어야 한다는 요구조건이 있기 때문이다. support vector classifier(서포트 벡터 분류기)가 최대 마진 분류기보다 발전된 모형인데 이것은 더 넓은 경우에 적용될 수 있다. SVM은 support vector classifier을 확장한 것으로 비선형의 클래스 경계를 수용한다.

4.1 maximal margin classifier

hyperplan과 separating hyperplane을 알아보자

p차원 공간에서 초평면은 차원이 p-1인 평평한 아핀(affine) 부분공간이다. 초평면의 수학적 정의는 2차원에서는 다음식으로 정의된다.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

임의의 $X = (X_1, X_2)^T$는 초평면 상의 점이다.

p차원의 초평면을 정의해보자 위의 식을 좀 더 확장하면 편안하다.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0$$

p차원 공간(즉, 길이가 p인 벡터)의 점 $X = (X_1, X_2, ...X_p)^T$가 위의 식을 만족하면 X는 초평면 상에 있다고 할 수 있따.

만약 위의식을 만족하지 않고 다음과 같다면?

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p > 0$$

4.2 support vector classifier

4.3 support vector machine

5 Unsupervised learning

연관된 반응변수가 Y가 없기 때문에 예측에는 관심이 없다. 대신에 $X_1, ..., X_p$ 측정에 대해 흥미로운 것들을 발견하고자 하는 것이 목적이다. 데이터를 시각화하는 유익한 방법이 있는가? 변수 또는 관측치들 중에서 subgroup들을 찾을 수 있는가? 비지도학습은 이러한 것들과 같은 질문에 대답하기 위한 다양한 기법들을 말한다. 이 장에서는 두 가지 특정한 유형의 비지도학습인 주성분분석(principle component analysis)과 클러스터링(clustering)에 대해 집중할 것이다. 주성분분석은 지도기법이 적용되기 전에 데이터를 시각화하거나 전처리 하는데 사용되는 도구이며, 클러스터링은 데이터의 알려지지 않은 서브그룹들을 발견하기 위한 광범위한 부류의 방법들이다.

지도학습에 비해 비지도학습은 매우 어렵다. 학습은 훨씬 더 주관적인 경향이 있고 반응변수의 예측과 같은 분석에 대한 단순한 목적이 없다. 비지도학습은 보통 탐색적 자료분석(exploratory data analysis)의 일부로서 수행된다. 더욱이 비지도학습 방법들로부터 얻은 결과를 평가하기가 어려울 수 있다. 이유는 교차검증을 수행하거나 독립적인 데이터셋에 대해 결과를 검증하는 보편적으로 용인된 메커니즘이 없기 때문이다. 이렇게 차이가 나는 이유는 간단하다. 지도학습과는 달리 비지도학습은 실제로 답을 모르기 때문에, 결과를 점검할 방법이 없기 때문이다.

5.1 주성분 분석

5.2 군집분석