pattern recognition and machine learning

Hyeonho Lee

2018년 10월 30일

1. 개요

일반화

훈련단계에서 사용되지 않았던 새로운 예시들을 올바르게 분류하는 능력을 generalization(일반화) 성능이라고 한다. 실제 적용에서는 입력 벡터의 가변성이 상당히 크므로 훈련 데이터는 가능한 모든 입력 벡터의 극히 일부분밖에 커버하지 못한다. 따라서 패턴 인식에서 보통 가장 중요한 목표는 바로 일반화다.

전처리

대다수 실용 애플리케이션에서 원래 입력 변수들을 전처리(preprocessed)하여 새로운 변수 공간으로 전환할 수 있는데, 이렇게 함으로써 패턴 인식 문제를 더 쉽게 해결할 수 있다. 이러한 전처리과정은 특징추출(feature extraction)과정이라고 불리기도 한다. 훈련 집합에서 사용한 것과 같은 전처리 과정을 새로운 시험 데이터에도 동일하게 적용하는 것을 잊지 말아야 한다.

차원감소

전처리의 한 종류로써 dimensionality reduction(차원감소)가 있는데, 이 과정은 굉장히 주의를 기울여야 한다. 왜냐하면 많은 전처리과정에서 정보들을 버리게 되는데, 만약 버려진 정보가 문제의 해결에 중요한 것이엇을 경우는 시스템의 전반적인 정확도가 악화될 수도 있기 때문이다.

차원감소의 예로써 높은 해상도의 비디오 스트림에서 실시간으로 얼굴을 인식해야 하는 경우가 있다. 컴퓨터는 초마다 매우 많은 픽셀을 다뤄아 햐는데, 이 픽셀 데이터를 복잡한 패턴 인식 알고리즘에 바로 적용하는 것은 계산적으로 실행 불가능할 일일 수도 있다. 그러나 모든 데이터를 다 사용하는 대신, 얼굴과 얼굴이 아닌 것들을 구별하는 차별적인 정보를 가지고 잇으면서 동시에 빠르게 계산하는 것이 가능한 유용한 특징들을 찾아내어 사용할 수 도 있을 것이다. 이 특징들을 패턴 인식 알고리즘의 입력값으로 활용하면 효과적으로 얼굴 인식 문제를 해결할 수 있다.

---

supervised learning

지도학습문제는 주어진 훈련 데이터가 입력 벡터와 그에 해당하는 표적벡터로 이루어지는 문제라고 한다.

분류문제

각각의 입력 벡터를 제한된 숫자의 분리된 카테고리 중 하나에 할당하는 경우에는 분류문제라고 한다.

회귀문제

기대되는 출력값이 하나 또는 그 이상의 연속된 값일 경우 에는 회귀문제라고 한다.

---

unsupervised learning

훈련데이터가 해당 표적 벡터 없이 오직 입력 벡터 x로만 주어지는 경우의 패턴 인식 문제이다.

clustering(집단화)

데이터 내에서 비슷한 예시들의 집단을 찾은 문제이다.

density estimation(밀도추정)

입력 공간에서의 데이터의 분포를 찾는 문제이다.

visualization(시각화)

높은 차원의 데이터를 이차원 or 삼차원에 투영하여 이해하기 쉽게 만들어 보여주는 문제이다.

---

reinforcement learning(강화학습)

강화학습은 주어진 상황에서 보상을 최대화하기 위한 행동을 찾는 문제를 푸는 방법이다. 강화학습은 지도학습의 경우와 달리 학습 알고리즘에 입력값과 최적의 출력값을 예시로 주지 않는다. 강화 학습 과정에서는 시행착오를 통해서 이들을 직접 찾아내게 되는데, 보통의 경우 알고리즘이 주변 환경과 상호작용할 때 일어나는 일들을 표현한 일련의 연속된 상태와 행동들이 문제의 일부로 주어직된다. 많은 경우 현재의 행동은 바로 직후의 보상뿐 아니라 다음 시간 단계들 전부의 보상에 영향을 미친다.

credit assignment(신뢰 할당)

보상은 최정승리까지 이끄는 모든 선택지에 대해서 잘 분배되어야 하는데, 어떤 것은 좋은 선택지, 어떤 것은 그에 비해서는 덜 좋은 선택지 라는 것을 신뢰할당이라고 한다.

exploration(탐색) & exploitation(이용)

강화 학습에는 탐사와 이용 간에 트레이드 오프가 있다. 탐사과정에서는 시스템이 새로운 종류의 행동을 시도하여 각각이 얼마나 효과적인지 확인하게 되며, 이용 과정에서는 시스템이 높은 보상을 주는 것으로 알려진 행동들을 시행하게 된다. 탐사와 이용 중 어느 하나에 너무 집중한 알고리즘은 그리 좋지 않은 결과를 내놓게 된다.

다항식 곡선 피팅

$sin(2\pi x)$ 함수를 사용하여 데이터를 만들었고, 타깃 변수에는 약간의 랜덤한 노이즈를 포함시켰다.

x = $(x_1, ..., x_N)^T$ 와 그에 해당하는 표적값 t = $(t_1, ..., t_N)^T$ 가 주어진다.

해당 곡선을 피팅하는 방법은 다음과 같은 형태의 다항식을 활용한다.

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

M은 이 다항식의 차수이며, x^j는 x의 j제곱을 일컫는다. 다항식의 계수 $w_0, ..., w_M$을 함께 모아서 벡터 w로 표현 할 수 있다. 다항함수 $y(x, w)$는 x에 대해서는 비선형이지만, 계수 w에 대해서는 선형이다. 다항 함수와 같이 알려지지 않은 변수에 대해 선형인 함수들은 중요한 설질을 지녔으며, 선형 모델이라 불린다.

훈련집함의 표적값들의 값과 함숫값 $y(x, w)$와의 오차를 측정하는 오차함수(errr function)을 정의하고 이 함수의 값을 최소화하는 방식으로 피팅할 수 있다.

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2$$

$E(w)$를 최소화하는 w값을 선택함으로써 이 곡선 피팅 문제를 해결할 수 있다. 오차함수가 이차 다항식의 형태를 지니고 있기 때문에 이 함수를 계수에 대해 미분하면 w에 대해 선형인 식이 나올 것이다. 따라서 이 오차 함수를 최소화하는 w는 유일한 값은 w^*를 찾아낼수 있다. 결과에 해당하는 식은 $y(x, w^*)$의 형태를 띠게 될 것이다.

model comparison(모델 비교) or model selection(모델 결정)

다항식의 차수 M을 결정하는 문제가 여전히 남아 있는데 이 문제가 바로 모델 비교 혹은 모델 결정 이라 불리는 중요한 콘셉트의 예시에 해당한다.

over-fitting(과적합)

$E(w^*) = 0$이면서 다항식이 모든 데이터 포인트를 지나갈 때, 피팅된 곡선은 심하게 진동하고, 함수 $sin(2\pi x)$를 표현하는 데는 실패하였다.

root mean square(평균 제곱근 오차)

$$R_{MSE} = \sqrt{2E(w^*)/N}$$

위의 식으로 정의된다. N으로 나눔으로써 데이터 사이즈가 다른 경우에도 비교할 수 있도록 했고, 제곱근을 취함으로써 E_{RMS}가 표적값 t와 같은 크기를 가지도록 했다. M값이 작은 경우에는 시험 집합의 오차가 상대적으로 큰 것을 볼 수 있다. 낮은 차수의 다항식은 비교적 융통성이 없으며, 그에 따라 피팅된 다항식이 함수의 진동을 다 반영하지 못한다. M값이 3과 8사이의 범위에 있는 경우에 시험 집합의 오차가 작고 피팅된 해당 다항식이 함수를 적절히 잘 표현한다.

M=9일 때, 훈련 집합의 오차가 0이다. 이 때 overfit이라고 할 수 있으며 열개의 계수를 통해 10차의 자유도를 가지고 있다고 할 수 있다.

차수 M에 따른 피팅 함수의 계수 w^*의 값들을 보면 피팅의 문제를 해결하는데 도움이 된다. M이 커짐에 따라 계숫값의 단위 역시 커지는 것을 알 수 있고, 특히 M=9 다항식의 경우는 상당히 큰 양숫값의 계수와 음숫값의 계수가 번갈아 나타난다. 하지만 훈련 집합데이터 포인트 사이에서는 결괏값이 크게 진동한다. 더 큰 M값을 가진 유연한 다항ㅎ식이 표적값들에 포함된 랜덤한 노이즈들에 정확하게 피팅되어서 이런 결과가 나타난 것이라고 할 수 있다.

여기서 사용 가능한 훈련 집합의 데이터의 수에 따라서 모델에서 사용하는 매개변수의 숫자에 제약을 두는 것은 뭔가 좀 불만족스럽다고 느낄 수 있을 수 있다. 이보다 풀고자 하는 문제의 복잡도에 따라서 모델의 복잡도를 결정하는 것이 더 말이 된다고 여길 수도 있다. 이 때, 최대 가능도(maximum likelihood)방법이 유용하다. 또한 과적합 문제는 최대 가능도 방법의 성질 중 하나로써 이해가 가능하다. 사실 베이지안(Bayesian) 방법론을 채택하면 과적합 문제를 피할 수 있다. 베이지안 관점에서는 데이터 포인트의 숫자보다 매개변수의 숫자가 훨씬 더 많은 모델을 사용해도 문제가 없다. 베이지안 모델의 특징은 데이터 집합의 크기에 따라서 적합한 매개변수의 수가 자동으로 정해진다는 것이다.

regularization

과적합 문제를 해결하기 위한 또 다른 방법은 정규화(regularization)이다. 오차 함수의 계수가 커지는 것을 방지하기 위해 페널티 항을 추가하는 것이다.

$$\tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t^n\}^2 + \frac{\lambda}{2}||w||^2$$

$||w||^2 \equiv w^T w = w_0^2 + w_1^2 + ... + .w_M^2$ 이며, 계수 λ가 정규화항의 제곱합 오류항에 대한 상대적인 중요도를 결정짓는다. 종종 계수 w_0는 정규화항에서 제외한다. 왜냐하면 w_0을 포함시키면 타깃 변수의 원점을 무엇으로 선택하느냐에 대해 결과가 종속되기 때문이다. w_0만 따로 빼내어 별도의 정규화 계쉬와 함께 다른 항을 만들어 포함시키기도 한다. 오차 함수의 최솟값을 찾는 문제 역시 닫힌 형식이기 때문에 미분을 통해서 유일해를 찾아낼 수 있다. 이 기법은 통계학에서는 shrinkage method라고 한다.(수축법) 이차 형식 (quadratic) 정규화는 리지 회귀라고 부르고 뉴럴 네트워크의 관점에서는 이를 가중치 감쇠(weight decay)라고 한다.

훈련집합

검증집합

확률론

불확실성

확률의 2가지 법칙

sum rule & product rule 1. joint probability 2. marginal probability 3. conditional probability

Bayes' theorem

1. prior probability
2. posterior probability

확률밀도

1. probability density
2. cumulative distribution function
3. probability mass function

expectation

1. conditional expectation
2. variance
3. covariance

베이지안확률

1. classical & frequentist
   - bootstrap method
2. Bayesian
3. likelihood function
4. directly proportional
5. noninformative prior distribution
6. crossvalidation

gaussian distribution(=normal distribution)

1. parameter
2. iid(independent and indentically distributed)
3. statistics

곡선피팅

1. 제곱합 오차 함수
2. predictive distribution
3. hyperparameter
4. MAP(= maximum posterior)

베이지안 곡선 피팅

1. 주변화

모델선택

1. 최대가능도 접근법
2. cv
3. information criteria

차원의저주

1. curse of dimensionality
2. importance problem
3. manifold

결정 이론

inference

1. 추론에 대해서…
2. 음..

오분류 비율의 최소화

1. decision region
2. decision boundary

기대 손실의 최소화

1. cost function(loss function)
2. utility function에 대해서..

reject option

1. 거부옵션

추론과 결정

1. 3가지 방법

1)

2)

3)

회귀에서의 손실 함수

1.
2. Minkowski loss

정보 이론

1. entorpy

상대적 엔트로피와 상호 정보량

1. KL(Kullback-Leibler divergence, Kl divergence) = 쿨백 라이블러 발산 & relative entropy
2. mutual information

- 기저함수 : 서로 직교하면서 선형적으로 독립적인 함수의 집합을 의미한다. 기저함수를 구성하는 설명변수(기저벡터)들은 선형 독립이여야 한다. 예시)

1. 연속확률분포 : 정규분포, 감마분포, 지수분포, 베타분포, 디리클레분포
2. 이산확률분포 : 이항분포, 음이항분포, 푸아송분포, 카이제곱분포, 초기하분포, 로지스틱분포

2.확률론

1. density estimation(밀도 추정) 문제
2. parametric
3. conjugate 사전 확률

이산 확률 변수

1. Bernoulli distribution
2. binomial distribution

베타분포