1.1 Example: Polynomial Curve Fitting (1)

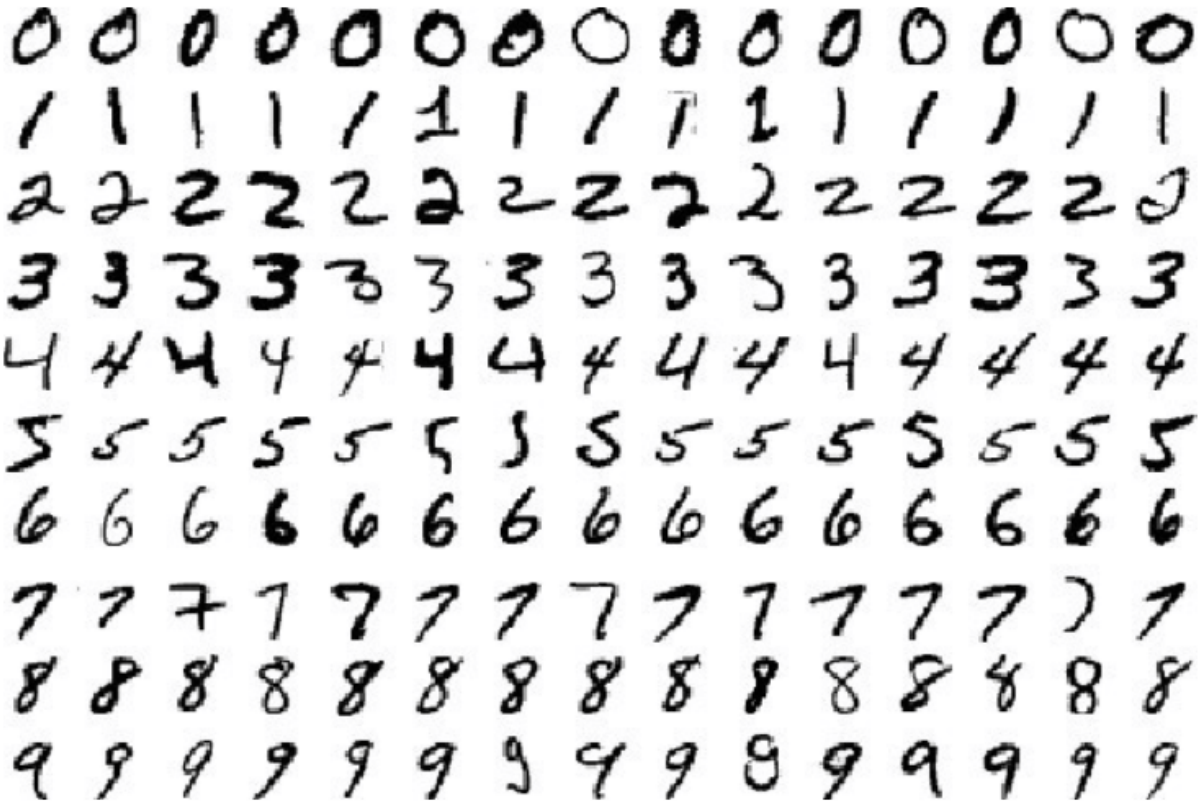
일반화와 패턴 인식

패턴인식에서 휴리스틱한 접근은 패턴을 잘 구별하지 못한다.

하지만 머신러닝은 이를 잘 구별할 수 있는데, 이건 함수 y의 출력으로 이해할 수 있음

대체로 train으로 학습하고 test로 테스트 하는데, 이 때, 학습 데이터 말고 새로운 예제를 적절히 분류하는 걸 **일반화**라고 한다.

대체로 이 때, 원래의 입력 변수를 일반적으로 전처리하여 새로운 변수 공간으로 변환합니다.



예를 들면, 이런 손글씨 이미지를 숫자로 매핑하는 것처럼 말이다.

이렇게 숫자로 매핑하면 몇 가지 장점이 있다.

1. 손글씨 이미지가 수치로 변환되어 처리할 수 있음
2. 모든 이미지를 일정한 크기의 행렬로 정의할 수 있음.

따라서 이미지를 수치로 전처리한다면, 모든 숫자의 위치와 크기가 동일해지기 때문에, 이후의 패턴 인식 알고리즘이 서로 다른 클래스를 구분하는 것이 훨씬 쉬워지고, 이를 feature extraction이라고 합니다.

어떤 이미지의 패턴을 인식한다고 하자.

간단한 이미지라면 가능하지만, 복잡한 이미지에서 모든 데이터를 이용해 패턴을 인식하는 것은 아주 힘든 일이다.

따라서 우리는 구별에 유용한 특징이나 정보를 뽑아 내는 전략으로 패턴을 인식한다.

학습의 분류

패턴 인식을 위해 다양한 학습 방법이 존재한다.

지도 학습

- label vector가 존재함

비지도 학습

- label vector가 없으며, 분류 및 예측 보다는 클러스터 및 밀도 추정으로 유사한 분포를 찾는 것. 또는 고차원을 저차원으로 투영해서 시각화를 시키는 것

강화학습

- 주어진 보상을 최대화 하는 방법으로 학습
- 크게 exploration or exploitation의 개념으로 구성이 됨.
- exploration : 새로운 방식을 탐색하는 것이며, 잠재적으로 더 나은 보상(reward)을 얻을 수 있는 행동을 발견하고, 환경에 대한 이해를 높임.
- exploitation : 현재까지 학습한 정보를 바탕으로 가장 높은 보상을 줄 것으로 예상되는 행동을 선택하여 보상을 최대화하는 과정. 현재에 충실하다고 생각하면 될듯
- 정리하자면, 그리디처럼 현재 있는 경우에 충실하게 학습할지, 아니면 최대의 보상이 되도록 학습할지 그 경계를 잘 학습해야 한다.

Example: Polynomial Curve Fitting

다항식 곡선 피팅

예시 데이터 셋을 정의할 건데, 0 <= x <= 1사이이고 타겟 t 값은 sin 2파이에 해당하는 y에다가 가우시안 분포로 만든 랜덤 노이즈를 더했다.

이런 데이터 셋에 대한 모델을 만들자면, 다음과 같은 다항식으로 정의할 수 있다.

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

- M : 다항식의 차수
- W : 회귀 계수
- X : 입력 데이터

이제 이 다항식에서 오차 함수가 최소가 되는 방향으로 회귀 계수를 갱신할 건데, 예시 오차 함수는 다음과 같다.

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

1/2은 계산의 편의를 위해 정의함

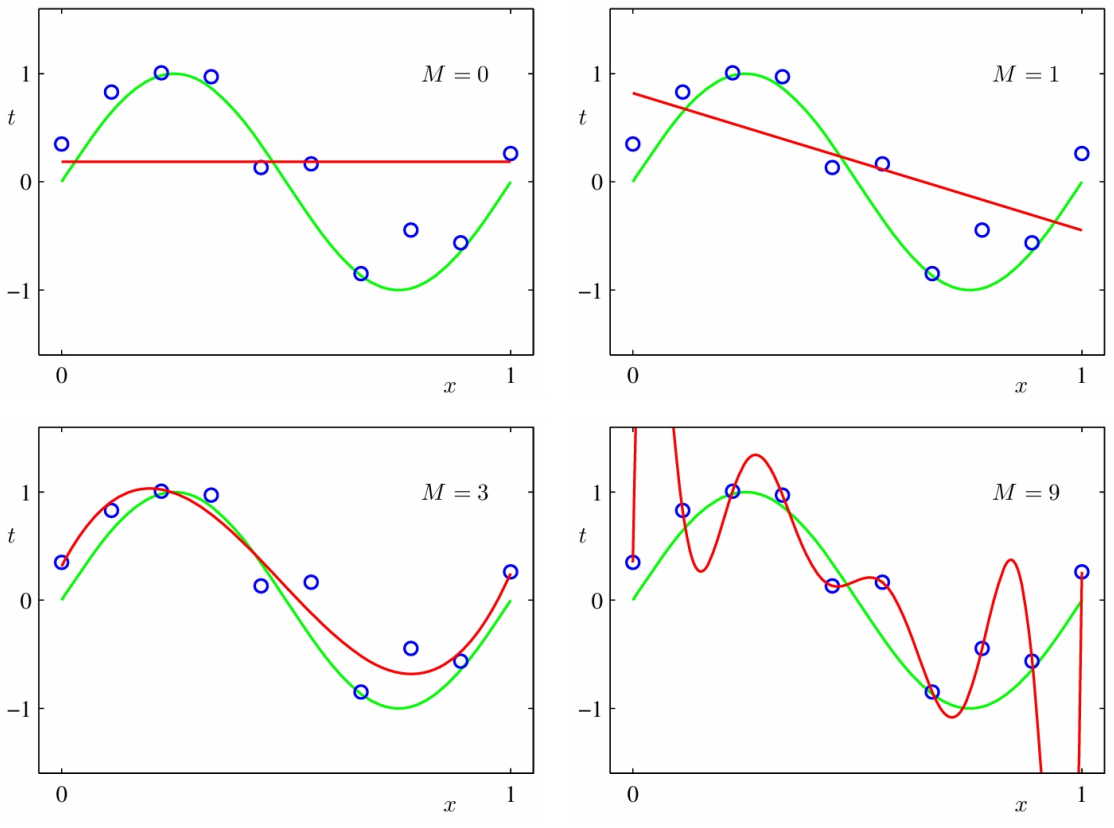
그럼 여기서 오차 함수가 최소가 된다?

이 말이 무슨 말이냐? —> 손실함수의 미분 값, 즉 도함수가 = 0 이 되는 최적의 해를 찾는 것으로 해결한다.

해당 오차 함수를 보면 2차 함수인 걸 알 수 있는데, 여기에 미분을 해서 1차 도함수로 만든 다음에, 최적의 해를 구한다.

규제의 필요성

자 그럼 이제 대강 알겠지만, 계수 w에 따라서 값이 엄청나게 달라진다.



우리는 패턴인식으로 일반화 가능한 모델을 만드는게 목적이다.

그리고 잔차를 분석하는 것도 유용한데, m=9일 때를 확인하면, 오버 피팅이 된 걸 확인 가능하다.

특히 너무 작은 M 차수는 큰 오차를 유발한다.

특히 회귀 식의 차수가 크면 클수록, 회귀 계수간의 차이가 엄청 벌어진다.

----> 과적합의 신호. 데이터가 스무스하게 연결된게 아니라 엄청 딱 붙어서 학습했다는 징후임.

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

정리하자면, 회귀식에서 차수가 커진다면, 과대적합이 일어나버리는 걸 알 수 있음.

따라서 규제를 통해서 차수에 대해서 통제해야 한다.

과대적합과 MLE

일단 과대적합이 차수가 크게 증가하면 생긴다는 건 알겠는데, 어떤 통계적 성질 때문일까?

바로 MLE다.

왜냐면 MLE는 주어진 데이터에서 가장 잘 맞는 데이터를 찾으려 하기 때문이다.

따라서 베이지안적 접근을 사용하면 사전, 사후 분포를 고려하기 때문에, 이를 해결할 수 있음.

규제(엄청 크게 증가하는 계수 값에 대해서 제약)

베이지안을 이용해서 과대적합을 해결할 수 있지만, 차수에 대한 규제로 문제를 해결 할 수 있습니다.

특히 데이터 크기가 제한된 상황에서 비교적 복잡하고 유연한 모델을 사용하고자 할 때 많이 사용됩니다.

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

보면 기존의 손실함수에 계수 λ 함수가 추가 되었으며 ||w|| 계수에 대한 제곱의 합이 추가 되었습니다.

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$$

벡터의 L2 NORM을 이용한 규제인데, 이를 본 따서 L2 규제라고 합니다.

이걸 w 파라미터에 대한 도함수를 구하기 위해서 미분을 적용하면

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

해당 수식으로 풀어지는데, 이 도함수가 0이 되는 w가 최적의 해로 판단한다.

그리고 해당 2차 정규 모델을 릿지라고 한다.