머신 러닝에서 사용하는 기본적인 용어 들

* Training set : 알고리즘을 학습할 때 사용되는 input dataset(supervised learning 의 경우는 target eh training set 에 포함)
* Validation set : 알고리즘을 학습할 때, 알고리즘의 성능을 (중간에) 측정하여 성능이 나쁜 알고리즘을 구별해내고 좋은 알고리즘을 선택할 수 있도록 하는 과정에서 사용되는 dadatas
* Test set : 학습된 알고리즘을 최종 테스트하여 알고리즘의 성능을 평가할 때 사용하는

Data set

* Hypothesis : 우리가 찾고자 하는 알고리즘을 의미한다. 머신러닝 알고리즘을 학습하는 것은 결국 머신러닝 알고리즘이 존재하는 전체 셋(hypothesis set)에서 최상의 머신러닝 알고리즘(hupothesis)을 찾는 것이다.

Hypothesis space 에서 best hypothesis 를 찾는 것

Model 의 학습 정도를 표현할 때, hypothesis 를 사용하여 수학적으로 표기하는 경우가 많다.

* 손실함수(loss function) 혹은 비용함수(cost function)

: 모수의 값을 실수로 mapping 하는 함수 이다. 구하고자 하는 모수(파라미터)를 알파라고 하때, 로스 펑션은 엘로 주어질 것이다.

모수에 대응되는 실수는 무리가 모수를 사용했을 때 예측 값과 실제 값과의 거리를 나타낸다고 생각하면 된다.

(정확히는 거리는 아니지만, 그렇게 생각하면 이해하기 쉽다) 즉, 모수를 선정할 때 기준이 되는 값이다.

* 대표적인 손실 함수

1. 제곱 손실: 회귀에서 많이 사용 함
   1. MSE
   2. RMSE
2. Cross ENTROPY: 분류모델에서 많이 사용함.
3. 힌지 손실: SVM(분류모델)에서 사용함.

* 머신러닝 모델 써머리

1. 지도학습(입력변수와 목표변수가 주어진다)
   1. 회귀 모형(목표 변수가 연속적인 값이다)  
      단순 선형 회귀모령
   2. 랏소, 릿지 회귀
   3. 의사결정 트리 회귀
   4. 배깅 회귀
   5. 랜덤 포레스트 회귀
   6. 부스팅 회귀
   7. SVM 회귀
2. 분류 모형(목표 변수가 이산 값을 가진다.)
   1. 로지스틱 회귀
   2. 의사결정 트리(분류기 트리)
   3. 배깅 분류
   4. 랜덤 포레스트 분류
   5. 부스팅 분류
   6. SVM 분류
   7. Perceptron
3. 비지도 학습(목표 변수가 주어지지 않는다.)
   1. 주성분 분석
   2. K-means clustering
4. 강화 학습
   1. 마르코프 결정 프로세스
   2. 몬테카를로 기법
   3. 시간차 학습
5. 선형 회귀

: 이 방법은 고객의 소득 같은 연속 변수 예측에 사용 된다. 모델은 최석의 선을 피팅하기 위해, 손실 함수의 값이 선형 방적식의 계수 들 에 관해 최소화 되도록 한다. 선형 회귀는 높은 바이어스와 낮은 베리언스 오류의 특징을 가진다.

1. 라소/릿지 회귀

: 이 기법은 선형 방적식의 계수에 패널티를 적용한 규제화 과정을 통해 과적합 문제를 통제 한다 . 릿지 회귀는 계수 제곱의 합, 랏소는 계수의 절대값에 패널티를 적용한다. 또한 패널티의 양을 조절할 수 있다. 릿지 회귀는 계수의 크기를 최소화하는 반면, 랏소 회귀는 계수를 제거하려고 노력한다. 즉 계수의 값이 0이 나올 수도 있다.

1. 의사 결정 트리 분류

: 각 레벨에는 이진 분할을 적용한다. 분할을 각 레벨의 부류가 최대한 순수한 분류만 남을 때까지 반복한다. 분류 오류율은 단순히 그 구역의 훈련 관측 값 중 가장 일반적인 부류에 속하지 않은 관측 값들의 비율이다. 즉 그 지역의 훈련 관측 값 중 많았던 부류가 에이 였다면 그 지역의 관측 값 중 에이 부류가 아닌 부류의 비율이다. 의사결정 트리는 핏팅 과정에서 높은 분산으로 인한 오버 피팅 문제를 겪는다. 가지 치기를 통해 오버피팅 문제를 감소시킬 수 있다. 의사 결정 트리는 낮은 바이어스와 높은 베리언스 오류의 특징을 가진다.(선형 회귀의 반대)

1. 배깅 분류/ 회귀

: 배깅은 의사결정 트리에 앙상블을 적용해 베리언스 오류를 최소화하는 동시에 바이어스에 의한 오류 성분이 증가하지 않도록 하는 기술 이다. 배깅은 복원 추출을 통해 표본을 선택하고, 모든 변수(열)은 각 표본에 관래 의사 결정 트리를 개별적으로 피팅한다. 그 후 투표(회쉬의 경우, 출력 값의 평균)를 통해 최종 결과를 인셈블 한다.

1. 랜덤 포레스트

: 이 방법은 배깅과 흡사한데, 한 가지 차이가 있다. 배깅은 각 표본 별로 모든 변수(열)를 선택하지만 랜덤 포레스트는 몇 개의 열만 선택 한다. 전체 변수를 선택하지 않고 일부만 선택하는 이유는 각 개별 트리의 표본을 추출할 때 많은 수의 변수가 항상 트리의 상위 계층에 나타나 결과적으로 분할 후에도 모든 트리가 유사하게 자라게 돼 앙상블의 취지를 반하기 때문이다. 이방법은 바이어스와 베리언스 오류가 모두 늦은 특징이 있다.

1. 부스팅 분류/ 회귀

:부스팅은 순차 알고리즘으로 결정 그루터기(단일 레벨 결정 트리 또는 하나의 루트와 2개의 단말 노드를 가진 트리) 같은 약한 분류기에 적용한 후 그 결과를 앙상블함으로써 강력한 분류기로 탈바꿈하기 위해 적용한다. 알고리즘은 모든 관측값에 동일한 가중값을 할당하고 시작한다. 그 후 후속 반복 작업에서 잘못 분류된 관측 값에는 가중값을 증가 시키고 제대로 분류된 관측값은 가중값을 감소시킨다. 결국 모든 개별 분류기는 더 강력한 분류기로 합쳐 진다. 부스팅은 오버피팅 문제가 있을 수 있디만, 매개 변구를 세밀히 튜닝함으로써 스스초 학습하는 최고의 머신 러닝 모델을 얻을 수 있다.