ML.NET 용어

원본글: https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/glossary
번역: hanhead@gmail.com

역자의 변)

딱 보면 알만한 뻔한 단어도 있고 아무리 설명을 읽어도 해당 분야의 전문적인 지식을 갖추지 않고는 도무지 정확한 의미를 파악하기 힘든 것도 있습니다.
그럼에도 불구하고 이렇게 몇개의 단어을 별도로 정리하였다는 것은 그 만큼 ML.NET 머신러닝을 배우기 위해서 필수적인 중요한 단어라는 것을 의미하는 것입니다.

뻔한 단어라고 생각하시는 부분이 있다고 하더라도 머신러닝 분야에서는 어떤 뉴앙스인지 까지 파악하는 지혜와
아무리 읽어도 모르는 것이 있다면 이건 예를들어 빗대어 이야기 드리자면 DB프로그래밍을 한다고 DB 엔진 자체를 이해하고 설계할 수 있을 정도의 지식이 필요하지 않듯이 어려운건 그걸 잘하시는 분들께 맡긴다는 생각을 하면 좋습니다. 우선은 응용프로그램 개발 위주로 생각해서 대충의 의미하는 바라도 이해할 수 있으면 좋을 것 같습니다.
그 또한 어렵다면 어디에 쓰는 것이다 정도로 결과만 알아두십시오.
물론 가능하다면 복잡한 수학으로 되어져 있는 해당 공식과 그 공식의 원리와 유도 과정을 다 이해한다면 좀더 더 나은 응용프로그램 설계가 가능하리라 생각합니다.

결론적으로 이후로 번역되고 작성될 글을 이해하는데 있어서 아래의 용어 정도는 필수적이오니 당장으로 아니더라도 이후 글을 읽으시면서라도 꼭 참조하시면서 이해하시면 좋겠습니다.

감사합니다.

정확도
곡선 아래 면적 (AUC)
이진 분류
결정 계수
특성
피쳐 엔지니어링(Feature Engineering)
F-점수
하이퍼 매개변수
라벨
로그손실
평균절대오차 (mean absolute error, MAE)
모델
다중 분류

N-그램
수치특성 벡터
파이프 라인
정확도
리콜
회귀분석
상대적 절대 오류
상대적 제곱 오류
평균제곱오차의 근 (RMSE)
감독된 머신러닝
훈련
변환
자율 머신러닝

아래의 목록은 사용자 정의 모델을 만드는데 유용하게 사용될 중요한 머신러닝 용어를 편집한 것입니다.

정확도

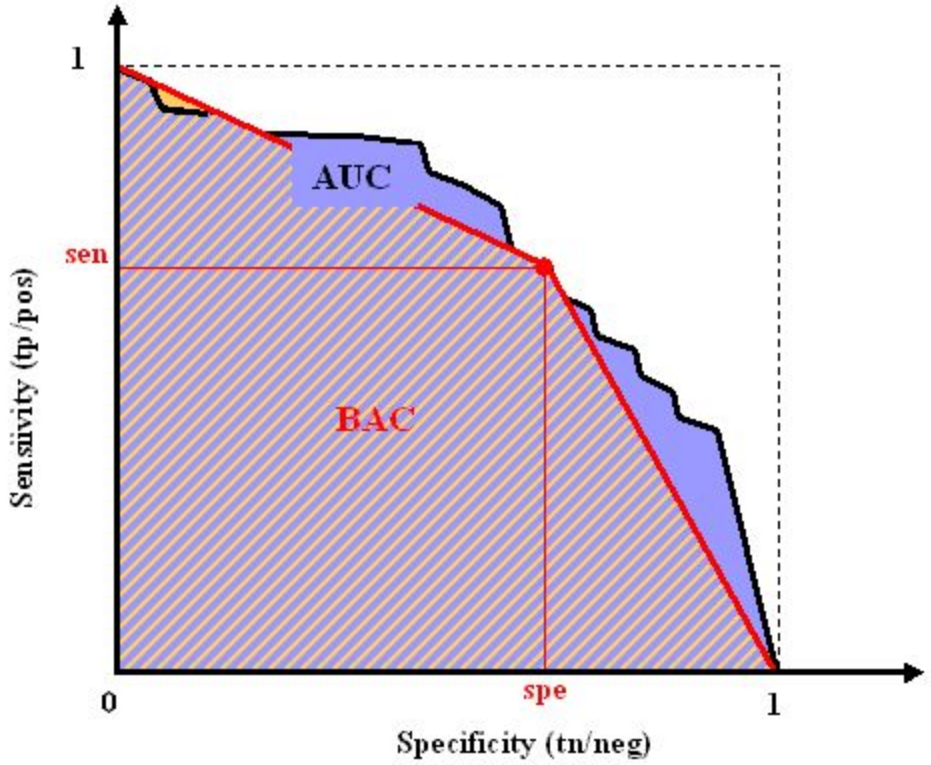
"분류"에서의 정확도라는 것은 테스트 데이터셋의 항목들 총 수량에서 정확하게 분류된 항목의 숫자를 나눈 것입니다. 값의 범위는 0에서 1사이로 0은 전혀 맞지않는 것이고 1은 완전히 정확한 것을 의미합니다. 정확도라는 것은 모델의 평가 척도 중에 하나이며 정밀도, 리콜, F-점수와 함께 고려하십시오.

관련 ML.NET API: BinaryClassificationMetrics.Accuracy

곡선 아래 면적 (AUC)

이진분류를 함에 있어서 평가 척도로 x축 오답으로 오답으로 잘 분류한 비율(FPR, Specificity)대한 y축 정답을 정답으로 잘 분류한 비율(TPR, Sensitivity)로 플로팅해서 얻은 곡선 아래 면적(AUC)의 값입니다.
값의 범위는 0.5(최악)에서 1(최상)입니다. ROC곡선(수신기 동작 특성 곡선, Receiver Operating Characteristic Curve)의 아래 면적이라고도 합니다.
자세한 내용은 위키디피아에서 수신자 작동 특성 문서를 참조하십시오.

관련 ML.NET API : BinaryClassificationMetrics.Auc.



이미지출처: http://www.causality.inf.ethz.ch/challenge.php?page=evaluation

이진 분류

레이블이 두 분류(참/거짓, 예/아니요, 키가 큼/작음 등)중 단 하나로 분류하는 경우입니다. 자세한 정보는 머신러닝 작업 주제의 이진 분류절을 참조하십시오.

분류

데이터가 종류를 예측하기 위해서 사용될때의 "감독된 학습"작업을 분류라고 합니다.
이진분류는 두가지 범주로만 예측하는 것을 의미합니다.(예: 이미지를 '고양이' 또는 '개'의 그림으로 분류).
다중 분류는 여러 범주를 예측하는 것을 의미합니다 (예: 이미지를 특정 개 품종의 그림으로 분류할 때).

결정 계수

회귀 분석에 있어서 데이터가 모델에 얼마나 잘 맞는지를 나타내는 평가 척도입니다.
범위는 0에서 1 사이입니다.
값 0은 데이터가 무작위이거나 그렇지 않으면 모델에 적합하지 않음을 의미합니다.
값 1은 모델이 데이터와 정확하게 일치 함을 의미합니다.
이것은 종종 r2, R2 또는 r-제곱이라고도 합니다.

관련 ML.NET API : RegressionMetrics.RSquared.

특성

측정되는 현상의 측정 가능한 속성으로, 일반적으로 숫자 (double) 값입니다.
여러 특성을 속성 벡터라고 하며 일반적으로 숫자배열(double [])로 저장됩니다.
특성들은 측정되는 현상의 중요한 특성을 정의합니다.
자세한 내용은 위키디피아의 특성이라는 글을 참조하십시오.

피쳐 엔지니어링(Feature Engineering)

피쳐 엔지니어링은 특성세트를 정의하고, 이용가능한 현상데이터 등으로부터 특성벡터를 만들것 즉 특성추출하는 소프트웨어 개발을 포함합니다.
자세한 내용은 위키디피아의 Feature engineering 글을 참조하십시오.

F-점수

정밀도와 리콜의 균형을 의미하는 분류에 있어서 평가 척도입니다.

관련 ML.NET API : BinaryClassificationMetrics.F1Score.

하이퍼 매개변수

머신러닝 알고리즘의 매개변수입니다.
예를 들어 *"의사결정 포리스트"에서 학습할 트리의 수 또는 그래디언트 디센트 알고리즘에서 단계 크기가 있습니다.
하이퍼 매개변수의 값은 모델을 학습하기 전에 설정되며 예측함수의 매개변수(예: 의사결정트리의 비교지점 또는 선형회귀모델의 가중치 등)를 찾는 프로세스를 관리합니다.

자세한 내용은 위키디피아의 하이퍼 매개변수 글을 참조하십시오.

* 의사결정 포리스트
역자의 변)
원문의 decision forest라는 용어는 제가 알지 못하는 단어라 원문 그대로 해석하였으나 decision tree를 잘못 표기한 것 같습니다.
random forest와 decision tree에서 헷갈려서 무의식 중으로 잘못 오타한 것이 아닐까 생각합니다.
제가 무지한 관계로 다른 의견이나 해당 용어에 관한 다른 의미를 알고 계신 분은 알려주시면 감사드리겠습니다.

라벨

머신러닝 모델로 예측되는 요소입니다.
예를 들어, 개 품종 또는 미래 주가.

역자의 첨부)
감독된 머신러닝에서 사용되는 dataset에서의 학습에 사용할 정답과 학습의 결과로 훈련된 모델이 내놓을 예측의 결과물에 해당되는 값를 의미합니다.

로그손실

분류에서 분류기의 정확성을 나타내는 평가 척도입니다. 로그손실이 적을수록 분류기가 정확한 것입니다.

관련된 ML.NET API : BinaryClassificationMetrics.LogLoss.

평균절대오차 (mean absolute error, MAE)

회귀분석에서 모든 모델오류의 평균인 평가척도입니다.
여기서 모델오류는 예측된 레이블값과 올바른 레이블값 사이의 거리입니다.

관련 ML.NET API : RegressionMetrics.L1.

모델

전통적으로 봤을 때 예측함수의 매개변수입니다.
예를 들어, 선형회귀모델의 가중치 또는 의사결정트리의 분리점 등입니다.

ML.NET에서 모델은 도메인 객체의 라벨을 예측하는 데 필요한 모든 정보(예: 이미지 또는 텍스트)를 포함합니다.
즉, ML.NET 모델에는 예측 기능의 당연히 매개 변수는 물론이고 특성화 단계들도 포함됩니다.

다중 분류

레이블이 세 개 이상의 분류중 하나로 분류되는 경우입니다.
자세한 정보는 머신러닝 작업 주제의 다중분류 절을 참조하십시오.

N-그램

텍스트 데이터의 특징 추출 방식 : N단어의 모든 순서가 특징 값으로 변환.

수치특성 벡터

수치값으로만 구성된 특성 벡터. 이것은 double[]과 유사합니다.

파이프 라인

모델을 데이터셋에 맞추기 위해 필요한 모든 작업입니다.
파이프 라인은 데이터 가져오기, 변환, 기능부여 및 학습단계로 구성됩니다.

일단 파이프 라인이 훈련되면 모델이됩니다.

정확도

분류에 있어서 분류의 정밀도는 해당 분류에 속하는 것으로 정확하게 예측된 항목 수를 그 분류에 속한 것으로 예상되는 항목의 총 수로 나눈 값입니다.

관련 ML.NET API :
BinaryClassificationMetrics.NegativePrecision, BinaryClassificationMetrics.PositivePrecision.

리콜

분류에 있어서 분류에 대한 리콜은 해당 분류에 속하는 것으로 올바르게 예측 된 항목의 수를 실제로 해당 분류에 속한 총 항목 수로 나눈 것입니다.

관련 ML.NET API : BinaryClassificationMetrics.NegativeRecall,
BinaryClassificationMetrics.PositiveRecall.

회귀분석

출력이 실제값인 감독된 머신러닝 작업입니다.(예: double)
예로는 주가 예측 등을 들 수 있습니다.
자세한 내용은 머신러닝 작업 주제의 회귀분석 절을 참조하십시오.

상대적 절대 오류

회귀 분석에서 사용되는 평가 척도로 모든 절대 오류의 합계를 맞는 라벨값과 모든 맞는 라벨값의 평균 사이의 거리 합으로 나눈 값입니다.

상대적 제곱 오류

회귀 분석에서 사용되는 평가 척도로 모든 제곱된 절대 오류의 합계를 맞는라벨 값과 모든 맞는라벨 값의 평균 사이의 제곱거리의 합으로 나눈 값입니다.

평균제곱오차의 근 (RMSE)

회귀 분석에서 사용되는 평가 척도로 오류의 제곱 평균의 제곱근입니다.

관련 ML.NET API : RegressionMetrics.Rms.

감독된 머신러닝

원하는 모델이 아직 보지 않는 데이터의 레이블(정답)을 예측하는 머신러닝의 하위 분류입니다.
"분류", "회귀분석" 및 "구조적 예측"이 그 예입니다.

자세한 내용은 위키디피아의 감독된 학습 문서를 참조하십시오.

훈련

주어진 교육 데이터 세트에 대한 모델을 식별하는 과정입니다.
선형 모델의 경우, 이는 가중치를 찾는 것을 의미합니다.
트리의 경우 분할 지점을 식별하는 것에 관여합니다.

변환

데이터를 변환하는 파이프 라인 구성요소입니다.
예를들어, 텍스트로 부터 숫자 벡터로 변환하는 것입니다.

자율 머신러닝

원하는 모델이 데이터에서 숨겨진 또는 잠재적인 구조를 찾는 기계 학습의 하위 분류입니다.

"클러스터링", "주제 모델링" 및 "차원감소" 등이 있습니다.

자세한 내용은 위키피디아의 자율 학습 글을 참조하십시오.