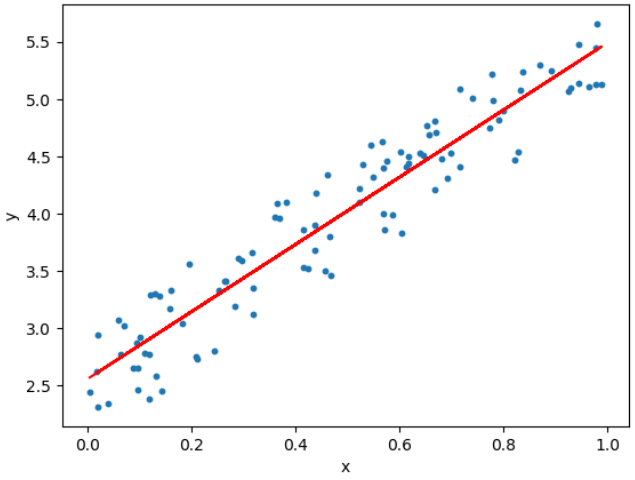
실습으로 이해하는 AIgorithm

Contents

01 Regression

회귀 분석 Regression



- **회귀 분석(regression analysis)**은 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법
- 활용 분야
  - 시간에 따라 변화하는 데이터 혹은 영향도
  - 가설적 실험
  - 인과 관계
  - 그 외 통계적 예측에 이용
- 그러나 가정의 적절성이 증명되지 않은 채로 이용되어 그 결과가 오용되기도 함
- 특히 SW의 발달로 분석이 용이해져 결과를 쉽게 얻을 수 있지만, 분석 방법의 적절성, 정보 분석의 정확성의 판단은 여전히 연구자의 몫

Regression 실습

- 주제 : 보스턴 집 값 예측
- Data : 보스턴 집값 데이터 (housing)
- 13개의 요소로 구성된 데이터는 총 506개의 열로 구성되어 있으며 사용하는 알고리즘에 따라 결과가 다르게 나옴

https://www.bostonkorea.com/news.php?mode=view&num=30580

feature	description
crim	자치시(town)별 1인당 범죄율
zn	25,000 평방 피트를 초과하는 거주지역의 비율
indus	비소매상업지역이 차지하고 있는 토지의 비율
chas	찰스강의 경계에 위치한 경우는 1, 아니면 0
nox	10ppm 당 농축 일산화질소
rm	주택 1가구당 평균 방의 수
age	1940년 이전에 건축한 소유주택 비율
dis	보스톤 직업센터까지의 접근성 지수
rad	방사형 도로까지의 접근성 지수
tax	10,000 달러당 재산세율
ptratio	자치시(town)별 학생/교사 비율
b	자치시(town)별 흑인의 비율
lstat	모집단의 하위 계층의 비율
medv	본인 소유의 주택가격 중앙값(단위 $1000)

MAE

$$\mathrm{ME} = \frac{\sum_{i=1}^{n} y_i - x_i}{n}.$$

In statistics, **mean absolute error** (**MAE**) is a measure of errors between paired observations expressing the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as:[1]

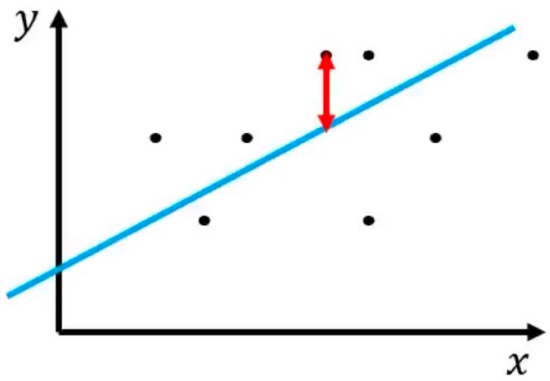
RMSE

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}.$$

The **root-mean-square deviation** (**RMSD**) or **root-mean-square error** (**RMSE**) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called *residuals* when the calculations are performed over the data sample that was used for estimation and are called *errors* (or prediction errors) when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.

Model	MSE	RMSE	MAE	R2
Linear Regression	21.895	4.679	3.271	0.741
Neural Network	22.050	4.696	3.413	0.739
Random Forest	2.561	1.600	1.005	0.970
Tree	1.858	1.363	0.715	0.978
AdaBoost	0.044	0.210	0.059	0.999

$$H(x) = Wx + b$$



MSE(Mean Squared Error)

예측값과 실제값의 차이의 제곱을 평균낸 것

$$\frac{1}{n}\sum_{i=1}^{n}(H(x_i) - y_i)^2$$

RMSE(Root Mean Squared Error)

MAE(Mean Absolute Error)

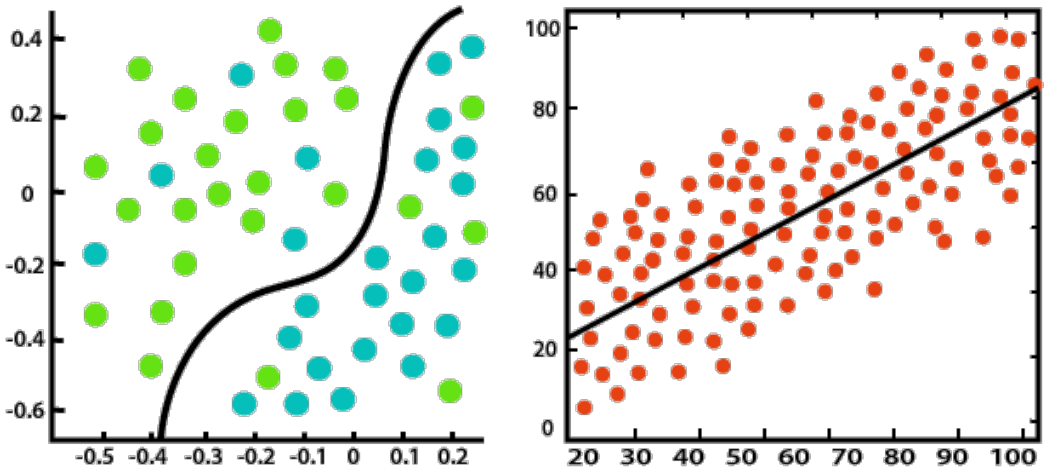
$$\frac{1}{n}\sum_{i=1}^{n}|H(x_i) - y_i|$$

R2(R Squared, R^2, 결정계수)

$$\frac{\sum_{i=1}^{n}(H(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

02 Classification

분류 분석 Classification



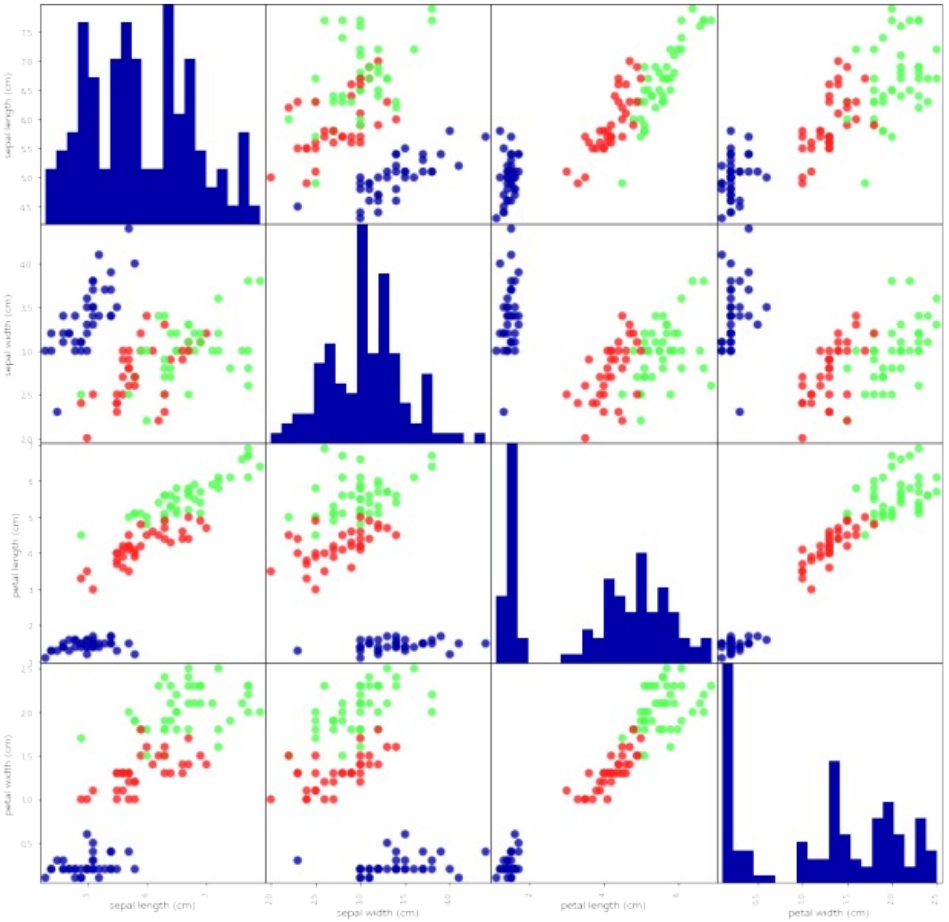
- **분류 분석(Classification analysis)**은 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법
- 활용 분야
  - 개와 고양이, 합격 불합격 등의 이진 분류
  - 숫자의 인식 등 컴퓨터 비전 분야
  - 여러 분류 중 하나를 선택하는 다중 분류 기법
- Regression이 값을 예측하는 데 비해서 Classification은 말 그대로 분류를 예측함

Iris Versicolor Iris Setosa Iris Virginica

http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html

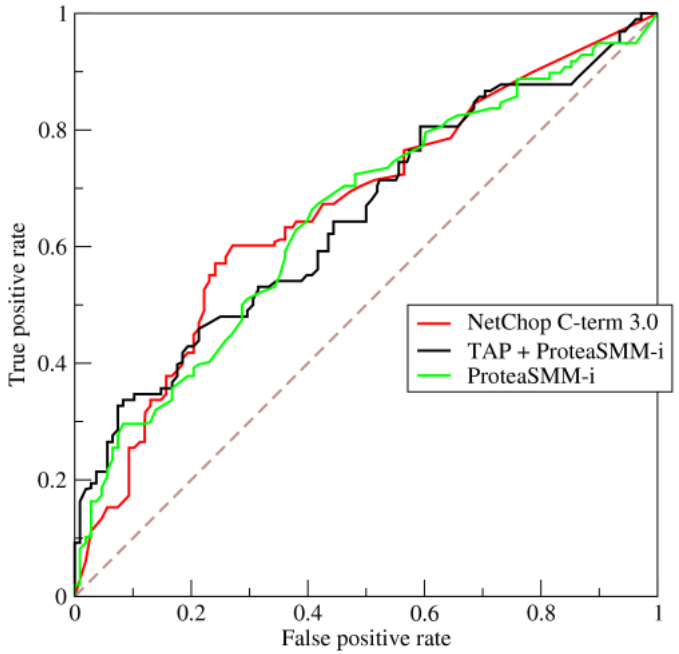
- 주제 : 붓꽃 품종 예측
- Data
  - 3개 품종 분꽃 : (Iris setosa, virginica and versicolor)
  - 4개 변수 측정 : 꽃받침 조각(petal) 길이, 넓이 - 꽃잎(sepal) 길이, 넓이

Classification 실습



https://tensorflow.blog/%ED%8C%8C%EC%9D%B4%EC%8D%AC-%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D/1-7-%EC%B2%AB-%EB%B2%88%EC%A7%B8-%EC%95%A0%ED%94%8C%EB%A6%AC%EC%BC%80%EC%9D%B4%EC%85%98-%EB%B6%93%EA%BD%83%EC%9D%98-%ED%92%88%EC%A2%85-%EB%B6%84%EB%A5%98/

ROC



A **receiver operating characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnosti c ability of a binary classifier system as its discrimina tion threshold is varied. The method was originally d eveloped for operators of military radar receivers star ting in 1941, which led to its name.

각종 평가 지표

Precision(정밀도) 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율

$$(Precision) = \frac{TP}{TP + FP}$$

Recall(재현율) 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율

$$(Recall) = \frac{TP}{TP + FN}$$

Accuracy(CA, 정확도) 전체 중 실제 True를 True라고, 실제 False를 False라고 예측한 것의 비율

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

F1 score Precision과 Recall의 조화평균

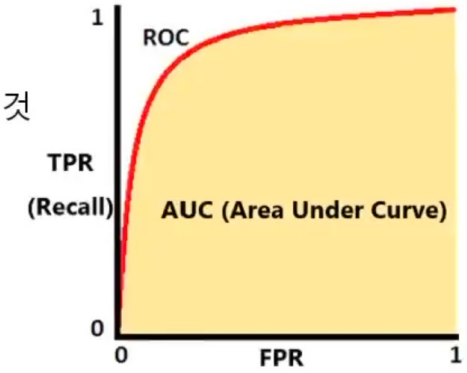
$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Fall-out 실제 False인 data 중에서 모델이 True라고 예측한 비율

ROC curve 여러 임계치들을 기준으로

Recall-Fallout의 변화를 시각화한 것

AUC ROC 그래프 아래의 면적



$$(FPR) = \frac{FP}{TN + FP}$$

		실제값	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

혼동행렬(Confusion Matrix)

출처: https://www.youtube.com/watch?v=FivvpiYoRlc&list=PL3geb_qrBQYfOmGuHtnf0RZMMXaev-Zvr

1 Regression

- 어떤 값을 예측할 때 사용하는 알고리즘
- Regression을 평가할 때에는 MAE, RMSE 등의 지표를 사용

2 Classification

- 어떤 대상을 분류할 때 사용하는 알고리즘
- Classification 문제를 평가 할 때에는 ROC 등을 사용해서 Accuracy를 체크

감사합니다.