면접 준비

데이터 사이언티스트

출처

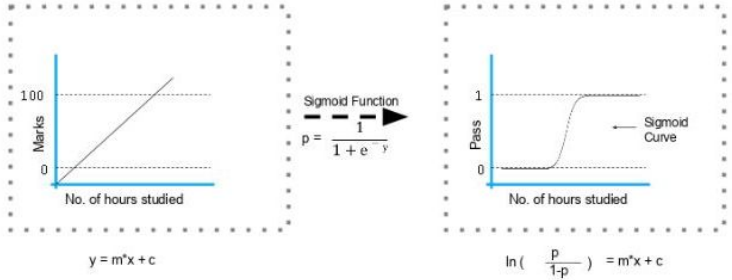
- https://www.simplilearn.com/tutorials/data-science-tutorial/data-science-interview-questions?source=sl_frs_nav_playlist_video_clicked#basic_data_science_interview_questions

1. Supervised vs Unsupervised

Supervised Learning	Unsupervised Learning
- 라벨링이 되어 있는 데이터를 사용 - decision tree, logistic regresion, support vector machine	- 라벨링이 되어 있지 않은 데이터를 사용 - k-means clustering, hierarchical clustering, apriori algorithm

2. Logistic regression

- 목적 : 종속 변수와 독립 변수 사이의 관계를 측정
  - 종속 변수 : dependent variable, label
  - 독립 변수 : independent variable, feature
- Logistic function ~ Sigmoid function
- Classification problem
- Linear Regression의 목표는 범위가 정해지지 않은 종속 변수와 독립 변수 사이의 선형 관계를 측정하는 것이지만 Logistic Regression은 Linear Regression을 이용하여 확률을 예측



2-1. Linear regression vs Logistic regression

Linear Regression

- Regression problems
- Continuous 데이터를 출력
- 종속 변수를 추정
- 직선 형태

Logistic Regression

- Classification problems
- Categorical 데이터를 출력
- 종속 변수의 가능성을 계산
- Sigmoid curve

3. Decision Tree

- 목적 : classification problems
- Algorithm
  1. 분기 전 데이터를 입력으로 사용
  2. 분기 전 데이터의 Entropy 계산, 분기 특징 후보들에 대한 Entropy 계산
  3. 분기 특징 후보들의 Information Gain 계산
  4. 가장 높은 Information Gain 값을 가지는 분기 특징을 선택
  5. Decision Tree 가 완성 될때까지 위 과정을 반복
- Entropy : 데이터가 얼마나 균일하게 분류되었는지 알려주는 척도, 즉 작을수록 잘 분류된 상태
- Information Gain : 분기 이전의 Entropy에서 분기 이후의 Entropy를 뺀 수치, 즉 높을수록 잘 분기했다고 판단.
- 단점 : Overfitting 문제 >> pre-pruning, post-pruning (가지치기), Random Forest

3-1. Entropy and Information Gain

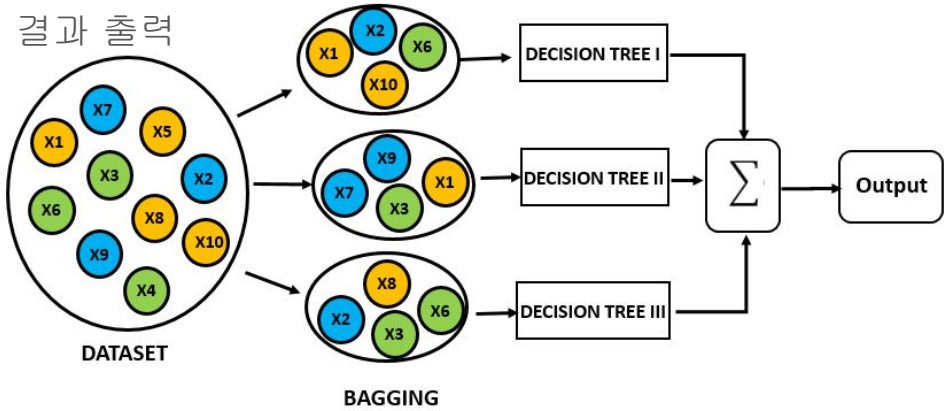
- 공식
-

3-2. Pruning

- 가지치기 역할
- pre-pruning
- post-pruning

4. Random Forest

- 앙상블 머신러닝 모델
- Decision Tree로 생성된 Overfitting Tree에서 일반적인 결과 출력
- Algorithm
  1. 학습 데이터에서 n개 데이터 표본 선택
  2. k개 feature 중 $\sqrt{k}$ 개를 선택
  3. Decision Tree 생성
  4. 1~3 번의 과정을 m 번 반복
  5. 테스트 데이터에서 m개의 결과 중 다수결 결과 출력



https://eunsukimme.github.io/ml/2019/11/26/Random-Forest/

앙상블

- Bagging
- Boosting

5. Overfitting

- 훈련 데이터에 과하게 맞추어져 훈련 데이터의 성능은 좋지만, 테스트 데이터에서는 성능이 저조
- 방지하는 방법
  1.