면접 준비

데이터 사이언티스트

출처

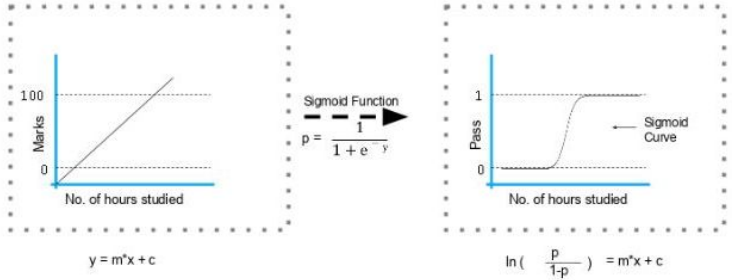
- https://www.simplilearn.com/tutorials/data-science-tutorial/data-science-interview-questions?source=sl_frs_nav_playlist_video_clicked#basic_data_science_interview_questions

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

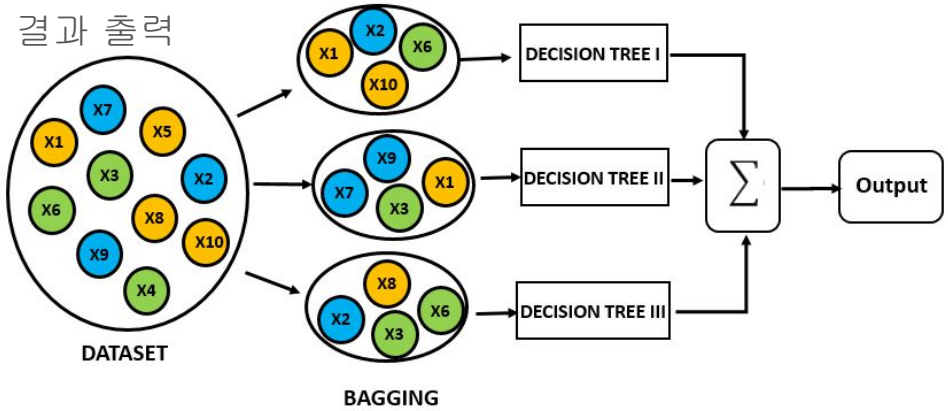
- 공식
-

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
- Overfitting 방지
  1. 더 많은 데이터 확보
  2. 모델 복잡도 줄이기
  3. cross validation 방법 사용 : k-fold cross validation
  4. 정규화 사용 (LASSO, Ridge)
  5. 앙상블 학습 방법 사용
- Underfitting 방지
  1. 새로운 특성 추가
  2. 모델 복잡도 증가
  3. 정규화 계수 줄이기

6. Univariate, Bivariate and Multivariate analysis

- Univariate : 일변량 데이터
  - 1개의 feature
  - 평균, 중위수, 최빈값(mode), 산포도, 범위, 최대, 최소 등의 통계 분석 진행

- Bivariate : 이변량 데이터
  - 2개의 다른 feature
  - 원인과 영향을 두 변수 사이의 관계 비교를 통해 분석

- Multivariate : 다변량 데이터
  - 3개 이상의 feature

7. Feature Selection Method

- Filter Method
  - 각 변수들에 대해 통계적인 점수와 순위를 매기고 선택
  - Linear Discrimination Analysis
  - ANOVA
  - Chi-Square
- Wrapper Method
  - 변수의 일부만을 모델링에 사용 후, 평가 작업을 반복하여 변수 선택
  - Forward Selection
  - Backward Selection
  - Recursive Feature Elimination
- Embedded Method
  - 위의 두 방법을 결합하여 어떤 변수가 가장 크게 기여하는 지를 찾아내는 방법
  - LASSO
  - Ridge Regression
  - Elastic Net

7-1. Feature Selection vs Feature Extraction

- https://bioinformaticsandme.tistory.com/188

8. Python Print

- 3의 배수는 “fizz”
- 5의 배수는 “buzz”
- 3과 5의 배수는 “fizzbuzz”

```
for fizzbuzz in range(51):
    if fizzbuzz % 3 == 0 and fizzbuzz % 5 == 0:
        print("fizzbuzz")
        continue
    elif fizzbuzz % 3 == 0:
        print("fizz")
        continue
    elif fizzbuzz % 5 == 0:
        print("buzz")
        continue
    print(fizzbuzz)
```

9. Missing Value

- Missing data 삭제 (확보한 데이터가 충분히 클때)
- 특정 값으로 채우기
- 결측값의 앞 또는 뒷 방향의 값으로 채우기
- mean, mode, medium, trimmed mean

- Pandas, scipy 등의 라이브러리를 사용하여 쉽게 채울 수 있음 (fillna)

10. Euclidean Distance in Python

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

```
import math

plot1 = [1,3]
plot2 = [2,5]

euclidean_distance = math.sqrt((plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2)
```

```
euclidean_distance
```

```
2.23606797749979
```

11. Dimensionality Reduction

- 원래의 차원에서 작은 차원으로 변환
- 장점
  - 데이터 압축하여 저장 공간 감소
  - 계산 시간 감소
- 종류
  - pca
  - auto-encoder
  - Linear Discriminant Analysis