

머신러닝

김건호

서울과학고

서론

테스트

목차

1	생각	1
1.1	기초 개념	1
1.2	Iris Classification	1
1.2.1	k -NN	2

1 생각

1.1 기초 개념

Regression과 Classification은 유사하면서도 차이가 있다. 우선 둘 모두 Supervised Learning의 방법이기 때문에 어떤 label이 된 독립 변수와 종속 변수가 있는 데이터 셋을 기준으로 새로운 독립 변수를 받았을 때 나오는 종속 변수를 예측하는 것이다. 차이점은 종속 변수의 타입이 되는데, 연속적인 값들을 가질 때를 Regression이라 하고 불연속적이고 이산적인 데이터 형태를 가질 때를 Classification이라 한다.

이것 이외에도 Clustering

1.2 Iris Classification

- 한 꽃에 해당하는 값 (2차원 배열에서 가로 행): Instance/Observation
- 한 성질에 해당하는 값 (2차원 배열에서 세로 열): Attribute/Masurement/Dimension
- 예측하려고 하는 값: Target

이 중 나중에 모델을 테스트하기 위해서 Training Data와 Testing Data로 나눈다. Training Data가 많을 수록 모델에게 좋기는 하지만, 가지고 있는 모든 데이터를 Training Data로 사용하면 이 모델이 정확하게 판단할 수 있는지에 대해서 전혀 근거가 없어지기 위해 한다.

또한 고른 Testing Data가 특별하였을 수도 있기 때문에 모든 간소한 Testing Data에 대해서

sklearn에는 여러 학습 모델이 있다. 이 중 이번 Iris의 Classification을 위해서는 k -NN을 사용한다.

```

import numpy as np

def incmatrix(genl1,genl2):
    m = len(genl1)
    n = len(genl2)
    M = None #to become the incidence matrix
    VT = np.zeros((n*m,1), int) #dummy variable

    #compute the bitwise xor matrix
    M1 = bitxormatrix(genl1)
    M2 = np.triu(bitxormatrix(genl2),1)

    for i in range(m-1):
        for j in range(i+1, m):
            [r,c] = np.where(M2 == M1[i,j])
            for k in range(len(r)):
                VT[(i)*n + r[k]] = 1;
                VT[(i)*n + c[k]] = 1;
                VT[(j)*n + r[k]] = 1;
                VT[(j)*n + c[k]] = 1;

            if M is None:
                M = np.copy(VT)
            else:
                M = np.concatenate((M, VT), 1)

            VT = np.zeros((n*m,1), int)

    return M

```

1.2.1 k -NN