5강. K- 근접이웃 분류기

※ 점검하기

Q1. 다음에 주어진 단계에 따라 데이터를 생성하고 K-근접이웃 분류기를 이용하여 분류를 수행하시오.

(1) 매트랩을 이용하여 다음과 같은 평균과 공분산을 가지는 가우시안 분포를 따르는 2차원 데이터를 각각 100개씩 가지는 두 클래스 집합 C1, C2를 생성하시오. 생성된 데이터를 2차원 평면상의 점으로 표시한 그래프에 대해 설명하시오.

$$\mu_1 = [0,0]^T,\ \mu_2 = [4,4]^T \qquad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

〈관련학습보기〉

1) 데이터 생성 및 실험 결과

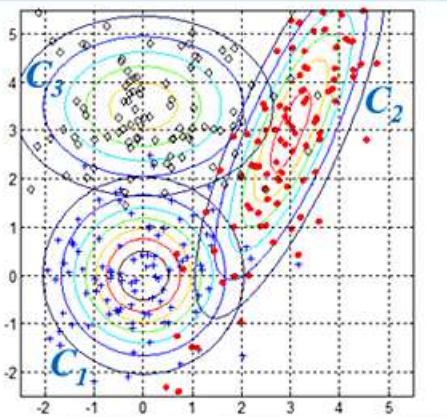
$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 1.6 \\ 1.6 & 4 \end{bmatrix}$$

$$\mu_3 = \begin{bmatrix} 0 \\ 3.5 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

» 학습 데이터 : 100개/클래스

» 테스트 데이터 : 10^5개/클래스



교재 2장 [프로그램 2-1, 데이터 생성]에서 데이터 개수(N),
평균(m1,m2)와 공분산(s1,s2)의 값을 조정하면 된다.
[참조] 3. 매트랩을 이용한 K-NN 분류기 실험의 「1) 데이터 생성 및 실험 결과」

(2) (1)의 데이터 집합과는 별도로 각 그룹별로 10개씩의 데이터를 따로 생성하여 그에 대해 설명하시오.

〈관련학습보기〉

1) 데이터 생성 및 실험 결과

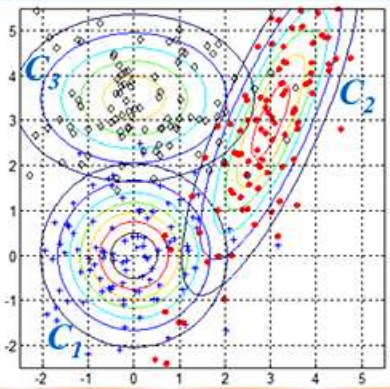
$\mu_1 = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_1 = \begin{bmatrix}1&0\\0&1\end{bmatrix}$

$\mu_2 = \begin{bmatrix}3\\3\end{bmatrix}, \Sigma_2 = \begin{bmatrix}1&1.6\\1.6&4\end{bmatrix}$

$\mu_3 = \begin{bmatrix}0\\3.5\end{bmatrix}, \Sigma_3 = \begin{bmatrix}2&0\\0&1\end{bmatrix}$

» 학습 데이터 : 100개/클래스

» 테스트 데이터 : 10^5개/클래스



K-NN 분류기의 분류 결과(%)

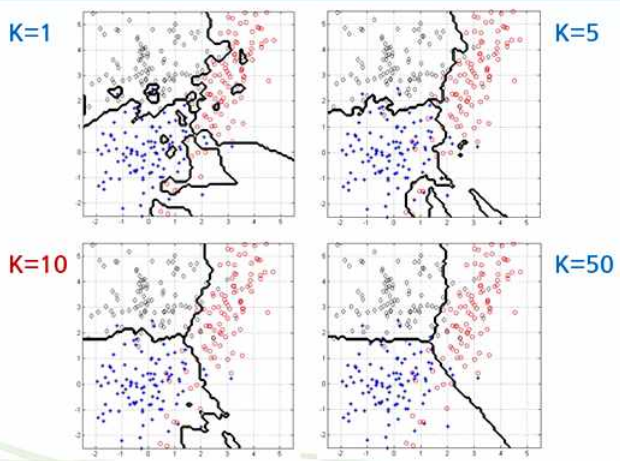
	K=1	K=5	K=10	K=50
학습 오차	0.00	9.67	11.67	11.67
테스트 오차	14.74	11.11	10.46	10.52

교재 2장 [프로그램 2-1, 데이터 생성]에서 데이터 개수(N),
평균(m1,m2)와 공분산(s1,s2)의 값을 조정하면 된다.
[참조] 3. 매트랩을 이용한 K-NN 분류기 실험의 「1) 데이터 생성 및 실험 결과」

(3) (2)에서 생성된 데이터를 K-근접이웃 방법으로 분류해 본다. 이때 K의 값을 3, 5, 10으로 변화시키면서 수행하고, 그 결과를 비교하시오.

〈관련학습보기〉

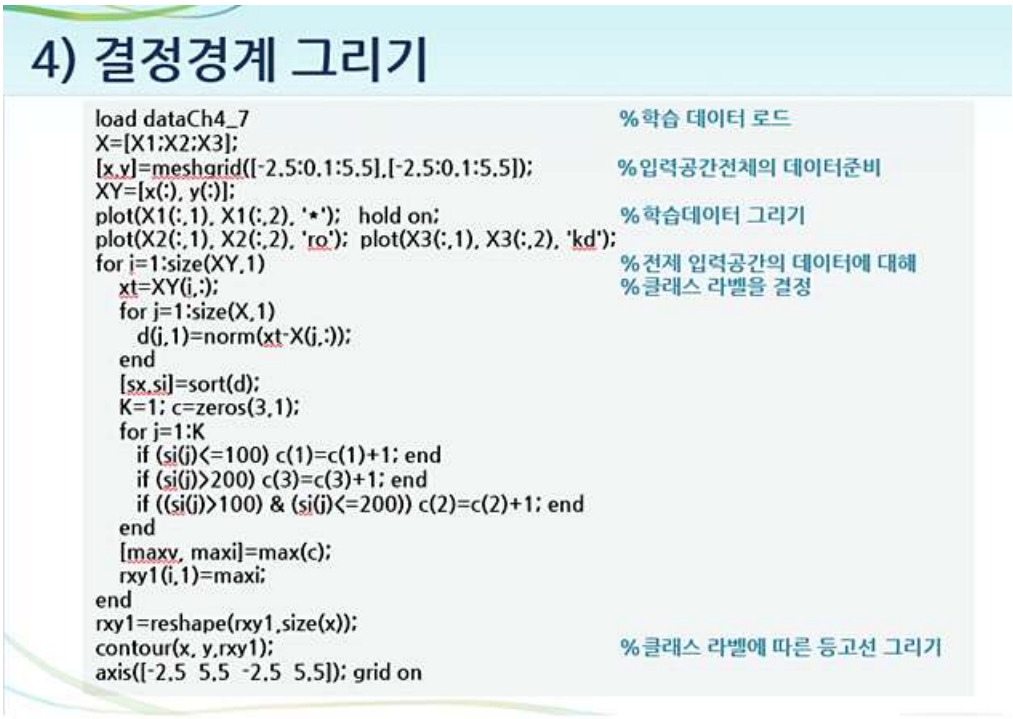
3) K값에 따른 결정경계의 변화



교재 5장 [프로그램 5-1, K-Nearest Neighbor Classifier]를 이용한다.
[참조] 3. 매트랩을 이용한 K-NN 분류기 실험의 「3) K값에 따른 결정경계의 변화」

(4) (3)에서 사용한 K 값에 대해, 찾아지는 결정경계를 그래프로 그려보고, 설명하시오.

〈관련학습보기〉



4) 결정경계 그리기

```
load dataCh4_7                                          %학습 데이터 로드
X=[X1;X2;X3];
[x,y]=meshgrid([-2.5:0.1:5.5],[-2.5:0.1:5.5]);          %입력공간전체의 데이터준비
XY=[x(:), y(:)];
plot(X1(:,1), X1(:,2), '*');  hold on;                  %학습데이터 그리기
plot(X2(:,1), X2(:,2), 'ro'); plot(X3(:,1), X3(:,2), 'kd');
for j=1:size(XY,1)                                      %전제 입력공간의 데이터에 대해
   xt=XY(j,:);                                          %클래스 라벨을 결정
   for j=1:size(X,1)
      d(j,1)=norm(xt-X(j,:));
   end
   [sx,si]=sort(d);
   K=1; c=zeros(3,1);
   for j=1:K
      if (si(j)<=100) c(1)=c(1)+1; end
      if (si(j)>200) c(3)=c(3)+1; end
      if ((si(j)>100) & (si(j)<=200)) c(2)=c(2)+1; end
   end
   [maxv, maxi]=max(c);
   rxy1(j,1)=maxi;
end
rxy1=reshape(rxy1,size(x));
contour(x, y,rxy1);                                     %클래스 라벨에 따른 등고선 그리기
axis([-2.5  5.5  -2.5  5.5]); grid on
```

교재 5장 [프로그램 5-2, Decision boundary of K-NN Classifier]를 이용한다.
[참조] 3. 매트랩을 이용한 K-NN 분류기 실험의 「4) 결정경계 그리기」

Q2. 1번 문제에서 사용된 데이터에 대해, 교재의 [표 5-1]에서 제시된 여러 가지 거리함수들을 적용하여 분류를 수행해 보고 그 결과를 비교해 보시오.

〈관련학습보기〉

3) K-NN 분류기의 설계 고려사항

② **거리함수** : 주어진 데이터와 학습데이터 간의 거리 계산

1차 노름	$d_1(x,y) = \|x-y\|_1 = \sum_{i=1}^{n} \|x_i - y_i\|$
2차 노름 (유클리디안 거리)	$d_E(x,y) = \|x-y\|_2 = \sqrt{(x-y)^T(x-y)} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
p차 노름	$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{n}(x_i-y_i)^p}$
내적	$d_{IN}(x,y) = x \cdot y = \sum_{i=1}^{n} x_i y_i$
코사인 거리	$d_{\cos}(x,y) = 1 - \frac{x \cdot y}{\|x\| \, \|y\|}$
정규화된 유클리디안 거리	$d_{NE}(x,y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i-y_i)^2}{\sigma_i^2}}$ (σ_i^2는 데이터의 분산)
마할라노비스 거리	$d_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$ (Σ는 데이터의 공분산행렬)

교재 5장 [프로그램 5-1, K-Nearest Neighbor Classifier]의 8번째 줄이 거리함수를 적용한 부분이므로, 이 부분을 거리함수에 맞게 수정한다.
[참조] 2. K-근접이웃 분류기의 특성의 「3) K-NN 분류기의 설계 고려사항」

※ 정리하기

1. K-근접이웃 분류기의 수행 단계

1) 주어진 데이터 **x**와 모든 학습 데이터 $\{x_1, x_2, \cdots, x_N\}$과의 거리를 계산함

2) 거리가 가장 가까운 것부터 순서대로 개의 데이터를 찾아 후보 집합 $N(x) = \{x^1, x^2, \cdots, x^k\}$를 만듦

3) 후보 집합의 각 원소가 어떤 클래스에 속하는지 그 라벨값 $y(x^1), y(x^2), \cdots, y(x^k)$을 찾음

4) 찾아진 라벨 값 중 가장 많은 빈도수를 차지하는 클래스를 찾아 **x**를 그 클래스에 할당함

2. K-근접이웃 분류기와 가우시안 베이즈 분류기

K-근접이웃 분류기	가우시안 베이즈 분류기
비모수적 밀도 추정법에 기반	모수적 밀도 추정법에 기반
분류 과정에서 새로운 데이터가 주어질 때마다 학습데이터 전체와의 거리 계산을 통해 K개의 이웃 데이터를 선정해 주어야 하므로 항상 학습 데이터를 저장하여야 함 → 데이터의 수가 증가하면 그에 비례하여 계산량과 메모리도 함께 증가하는 문제점을 가짐	학습데이터를 이용하여 일단 평균과 표준편차를 계산한 후에는 더 이상 학습데이터를 필요로 하지 않음
K-NN 분류기는 가우시안 베이즈 분류기에 비해 매우 비선형적인 결정경계를 가지며, 데이터의 분포 형태에 따라 성능이 크게 좌우되지 않음	

3. K-근접이웃 분류기 설계 시의 고려사항

1) 적절한 K값의 결정

- K=1인 경우에는 바로 이웃한 데이터에만 의존하여 클래스가 결정되므로, 결국 노이즈에 민감한 결과를 초래함
- 한편 K가 지나치게 커지면 주어진 데이터 주변 영역만이 아닌 전체 데이터 영역에서 각 클래스가 차지하는 비율에 의존하여 분류를 수행함

2) 거리함수

- 주어진 데이터와 학습데이터들 간의 거리를 계산할 때 어떤 거리함수를 이용하느냐에 따라 선택되는 이웃이 달라질 수 있음
- 이는 결국 분류 성능에 직접적인 영향을 미치게 됨