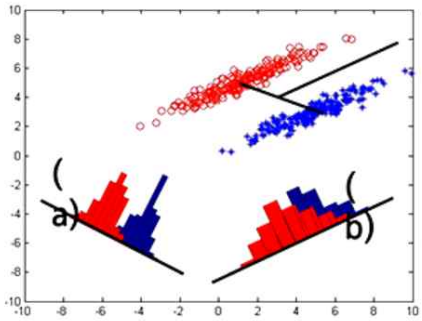
9강. 특징추출

※ 점검하기

Q1. 다음 그림(교재 [그림8-9])에 나타난 데이터를 생성한 후, 주성분분석법과 선형판별분석법을 적용해 보시오.



(1) 다음과 같은 분포에 따라 두 클래스의 데이터를 생성하시오.

클래스 1 : $\mu_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 3 & 1.7 \\ 1.7 & 1 \end{bmatrix}$

클래스 2 : $\mu_2 = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 3 & 1.7 \\ 1.7 & 1 \end{bmatrix}$

(2) 두 클래스를 하나로 합한 전체 데이터에 대해 PCA를 적용하여 첫 번째 주성분 벡터를 찾으시오.

(3) 찾아진 주성분 벡터 방향으로 사영한 1차원 특징값들을 찾아서 2차원 평면상에 그려보고 설명해보시오.

(4) 같은 데이터에 대해, LDA를 적용하여 하나의 고유벡터를 찾아보고, 그림의 (a)에서 사용된 것과 비교해 보시오.

(5) PCA의 결과와 LDA의 결과를 비교해 보시오.

〈관련학습보기〉

1) 학습데이터 생성

```
% 학습데이터의 생성 ----------
N=100;
m1=[0 0]; s1=[9 0;0 1];
m2=[0 4]; s2=[9 0;0 1];
X1=randn(N,2)*sqrtm(s1)+repmat(m1,N,1);   % 클래스1 데이터 생성
X2=randn(N,2)*sqrtm(s2)+repmat(m2,N,1);   % 클래스2 데이터 생성
figure(1);
plot(X2(:,1),X2(:,2),'ro'); hold on
plot(X1(:,1),X1(:,2),'*');
axis([-10 10 -5 10]);
save data8_9 X1 X2
```

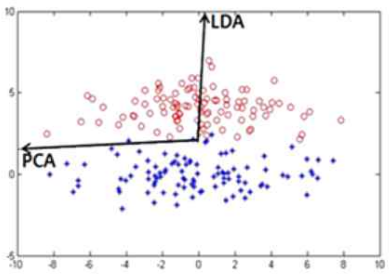
2) PCA에 의한 분석

```
% PCA에 의한 분석 ----------
X=[X1;X2];
M=mean(X); S=cov(X);                % 평균과 공분산 계산
[V,D]=eig(S);                       % 고유치분석(V:고유벡터행렬, D: 고유치행렬)
w1=V(:,2);                          % 첫 번째 주성분벡터
figure(1);                          % 첫 번째 주성분벡터를 공간에 표시
line([0 w1(1)*D(2,2)]+M(1),[0 w1(2)*D(2,2)]+M(2));
YX1=w1'*X1'; YX2=w1'*X2';           % 첫 번째 주성분벡터로 사영
pYX1=w1*YX1; pYX2=w1*YX2;           % 사영된 데이터를 2차원 공간으로 환원
figure(2);                          % 사영된 데이터를 2차원 공간으로 환원하여 표시
plot(pYX1(1,:), pYX1(2,:)+M(2),'*');
hold on axis([-10 10 -5 10]);
plot(pYX2(1,:), pYX2(2,:)+M(2),'ro');
```

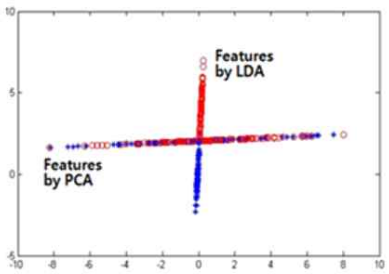
3) LDA에 의한 분석

```
% LDA에 의한 분석 ----------
m1=mean(X1); m2=mean(X2);
Sw=N*cov(X1)+N*cov(X2);          % within scatter 계산
Sb=(m1-m2)'*(m1-m2);             % between scatter 계산
[V,D]=eig(Sb*inv(Sw));           % 고유치 분석
w=V(:,2);                        % 찾아진 벡터
figure(1);                       % 찾아진 벡터를 공간에 표시
line([0 w(1)*-8]+M(1),[0 w(2)*-8]+M(2));
YX1=w'*X1'; YX2=w'*X2';          % 벡터w 방향으로 사영
pYX1=w*YX1; pYX2=w*YX2;          % 사영된 데이터를 2차원 공간으로 환원
figure(2);                       % 사영된 데이터를 2차원 공간으로 환원하여 표시
plot(pYX1(1,:)+M(1), pYX1(2,:),'*');
plot(pYX2(1,:)+M(1), pYX2(2,:),'ro');
```

4) 실험 결과



생성된 데이터



PCD & LDA 결과

강의(4. 매트랩을 이용한 실험)에서 언급된 프로그램
(또는 교재 176쪽의 [프로그램 8-1, PCA와 LDA] 및 [그림 8-10])을 참조한다.
[참조] 4. 매트랩을 이용한 특징추출 실험

Q2. 1번 문제와 같은 분포를 가진 테스트데이터 집합을 만들고, 최근접이웃 분류기를 이용하여 분류를 수행해 보시오.

(1) 1번 문제에서 주성분분석법에 의해 찾아진 특징값을 이용하여 테스트데이터에 대한 분류를 수행하고 분류율을 계산하시오.

(2) 1번 문제에서 선형판별분석법에 의해 찾아진 특징값을 이용하여 테스트데이터에 대한 분류를 수행하고 분류율을 계산하시오

〈관련학습보기〉

2) 최근접이웃 분류기

1 수행 단계

» 주어진 데이터 x와 모든 학습 데이터 $\{x_1, x_2, \ldots, x_N\}$과의 거리를 계산함

» 거리가 가장 가까운 데이터를 찾아 x_{min}으로 둠

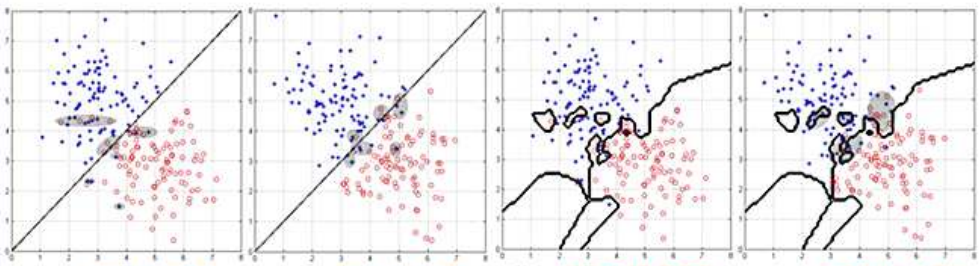
$$x_{\min} = \operatorname{argmin}_{x_i \in X}\{d(x, x_i)\}$$

» x_{min}이 속하는 클래스에 할당함

즉 $y(x_{min})$과 같은 값을 가지도록 $y(x)$를 결정함

2) 최근접이웃 분류기

2 문제점 → 과다적합



최근접이웃 분류기에 대해서는 5강의 『근접이웃분류기의 수행 단계』를 참조한다.
[참조] 5강. K-근접이웃 분류기 - 「1. K-근접이웃 분류기의 2) 최근접이웃 분류기」

※ 정리하기

1. 선형변환에 의한 특징추출

1) 선형변환에 의한 특징추출

- 선형변환에 의한 특징추출이란 주어진 데이터를 변환행렬에 의해 정해지는 방향으로 사영함으로써 저차원의 특징값을 얻은 것을 의미함

- 이때 변환행렬의 각 열은 사영할 저차원의 부분공간의 기저를 이루어 해당 부분공간을 정의함

- 따라서 이와 같은 접근법을 부분공간분석이라고 함

2. 주성분분석법

1) 공분산 분석

① 선형변환을 통해 얻어지는 특징데이터 집합의 평균과 공분산은, 입력데이터 집합의 평균과 공분산을 변환행렬에 의해 변환함으로써 얻어짐

② 따라서 변환행렬을 적절히 조정하여 특징데이터 집합의 공분산이 단위행렬 또는 대각행렬이 되면 간단한 분류기를 통해서도 최소분류오차를 얻을 수 있음

▶ 특징데이터 집합의 공분산이 대각행렬과 단위행렬이 되도록 변환하는 방법을 각각 대각화 변환과 화이트닝 변환이라고 함

2) 주성분분석

① 주성분분석(PCA)은 변환 전의 데이터가 가지고 있는 정보를 차원 축소 후에도 최대한 유지하는 방향으로 변환행렬을 결정함

② 따라서 데이터의 공분산행렬의 고유치와 고유벡터를 찾아, 고유치가 가장 큰 값부터 순서대로 m개에 대응하는 고유벡터로 변환행렬을 구성함

▶ 클래스라벨 정보를 활용하지 않는 비교사적인 방법으로 인해 분류에 핵심이 되는 정보의 손실을 초래할 수 있음

▶ 데이터 자체가 비선형 구조를 가진 경우에는 저차원의 특징으로 이를 효과적으로 반영할 수 없음

3. 선형판별분석법

1) 선형판별분석(LDA)은 클래스라벨 정보를 적극 활용하여 클래스간 판별이 용이한 방향으로 차원을 축소시키는 변환행렬을 찾음

2) 이 변환행렬은 클래스 간의 거리는 가능한 멀어지게 하고, 같은 클래스 내에서는 결집되도록 하여 분류에 적합한 특징으로의 변환을 유도함

① PCA와 마찬가지로 선형변환이라는 기본적인 제약점을 비롯하여, 클래스간 산점행렬의 랭크로 인해 M개의 클래스 분류문제에서 최대 M-1차원의 특징을 가짐

② 입력데이터의 수가 입력 차원보다 크지 않으면 클래스내 산점행렬이 특이행렬이 되어 역행렬을 찾을 수 없게 됨