4.4.1

P\_k(x)= Pr(Y = k | X = x) # 사후 확률. 예측값 x 주어졌을 때, k번째 클래스 속할 확률.

f\_k(X) = Pr(X | Y = k) # k번째 클래스에 속한다고 가정하면, x 가질 확률 밀도.

* 어떤 클래스 속한다고 하면, 그 클래스 내에서 변수 하나 나타날 가능성.

폰트, 텍스트, 친필, 라인이(가) 표시된 사진

자동 생성된 설명=> 15 뮤는 평균, 시그마 제곱은 분산.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명=> 16

폰트, 텍스트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명 => 17: 16식을 15에 넣음 나오는 식.

여기서 파이\_k는 k번째 클래스에 속할 확률

폰트, 친필, 텍스트, 서예이(가) 표시된 사진

자동 생성된 설명위의 식을 로그 변환 거친 식.

17이랑 이 식은 동일한 결과 출력. 결국 가장 큰 값 낸거로 분류하는 거 똑같음.

원래는 각 클래스가 정규 분포 따른다고 해도, 각 관측치의 평균값, 표준편차, 분산값 추정해야 함.

그래서 LDA를 해야하고,

폰트, 텍스트, 도표, 화이트이(가) 표시된 사진

자동 생성된 설명

 파라미터 추정해야 하는데, 그걸 이 식으로.

그걸 위 식에 대입.

관측값들을 위 식 가장 큰 클래스로 분류.

4.4.2

LDA 예측 변수 여러 개로 확장하는 내용.

다변량 가우시안 밀도 함수.

여기서 시그마는 covariance matrix.

LDA 분류기.

여기서 x가 k 클래스에 속할 확률 가장 큰 클래스로 분류.

도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

이건 다변량 분포 예시. 각 클래스 95% 확률 원 형태.

점선은 Bayes 결정경계. 가장 이상적 분류 형태.

실선은 LDA 경계.

봐보면 LDA가 Bayes랑 가장 밀접한게 보임.



Bayes나 LDA는 이런식으로 보통 0.5 임계값으로 두고 함.

0.5 이상 뜨면 그 클래스로 분류.

🡺 예측 오류 줄이려고 임계값 조정 가능.

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

ROC curve

AUC(아래 면적) 0.5면 아무것도 예측 못한다는 거 의미.

4.4.3

LDA는 클래스 전체 동일 covariance 행렬

QDA는 각각 클래스들 다른 covariance 행렬

텍스트, 폰트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

QDA는 각 클래스별 covariance 행렬 추정해서 더 많은 파라미터 필요.

그래서 분산 높음.

* 데이터 많을 땐 이게 유리.
* 데이터 적을 땐 LDA

4.4.4

다차원에서는 함수 추정하는게 어려워서 나이브 베이즈는 각 클래스에서 번수들 서로 독립적이라고 가정.



f\_k(x) 는 이런식으로 개별 변수들 싹다 곱으로. Independent 하다고 해서.

Independent 가정이 분포만 고려하면 돼서 다차원에서 유리.

근데 또 실제로 independent 성립 안해도, 괜찮은 성능 보임.

4.5.1

여태까지 나온거 서로 비교. 확률값 어떻게 써먹는지.

폰트, 텍스트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명 log odds 값.

LDA:

텍스트, 친필, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명



LDA(Linear Discriminant Analysis)

결국 최종식 보면 일반적인 선형 결합 형태. 결국 선형 회귀에서 확장해서 분류 문제에 적용한 꼴.

QDA:

텍스트, 폰트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

변수들 a\_k, b\_kj, c\_kj 보이듯이, 각 클래스에 대한 매개변수 더 늘어간게 보임.

결국 마지막 항에 곱 형태로 매개변수 곱 형태로 이루어져 있어서 이차형태(비선형) 적인 분류 경계 설정 가능. 🡺 더 복잡한 분류 경계 가능하다 뭐 그런 말.

나이브 베이즈:

텍스트, 폰트, 친필, 도표이(가) 표시된 사진

자동 생성된 설명



애초에 나이브 베이즈는 각 특징들이 독립적인 형태로 분포된다고 가정.

그렇다 보니 단순 곱이나 제곱 형태가 아닌 단순 합 형태로 구성.

# 보면 괄호 형태로 있음. 그게 독립적이다 라는걸 표현한 것.

LDA vs QDA:

QDA는 다른 공분산 행렬(각 클래스 데이터 분포 다른 경우)도 허용.

LDA는 모든 클래스 동일한 공분산 행렬.

그래서

폰트, 화이트, 텍스트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

C\_kjl 0 이면 LDA랑 동일. 🡺 이차항 사라지면서 LDA랑 동일한 식 형태.



그리고 나이브 베이즈도 얘기 나왔는데, 독립적인 형태가 깨지게 된다면 선형적 형태로 바뀜.

그래서 결국 LDA와 동일한 형태로 구성됨. 독립적인 형태만 깨지면.

가



이런식으로 1차원 가우시안 분포 따른다면



되어서 LDA랑 똑같이 됨. 그래서 결국 LDA랑 나이브 베이즈랑 비슷함.

QDA vs 나이브 베이즈

QDA는 각 변수간 곱 형태로 이루어져 있음. 이는 결국 각 변수간 상호작용이 일어난다는 말.

나이브 베이즈는 애초에 각 변수간 independent를 가정. 이는 상호작용이 존재하지 않는다는 말.

그래서 두개는 완전 다른 개념.

로지스틱 회귀 vs LDA

둘다 선형적 형태 구성.

근데 LDA는 가우시안 분포로 클래스 결정.

로지스틱 회귀는 MLE로 클래스 결정.

# 가우시안 분포 보고 형태 이런식으로 나타나있다 하면 LDA, 아니다 하면 로지스틱 회귀.

KNN:

결정경계 뭐 어떻다 이런게 없어서 비선형이든 선형이든 다 유리.

N(데이터 수)이 p(차원)보다 커야 KNN이 유리.

데이터 수 적고 비선형이다 하면 QDA가 유리.

4.5.2

랜덤 데이터셋 100개 학습.

시나리오1:

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명 🡺 결정 경계 선형

두개 클래스 있는 20개 학습 데이터. 각 데이터는 상관관계 없는 ‘정규분포’ 추출.

그림 보면 LDA가 가장 성능 좋음. y축이 오류

KNN은 분산값 높아서 성능 가장 안좋고.

시나리오 2:

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명 🡺 결정 경계 선형

시나리오 1 가정에 두 column 간의 상관관계 0.5 가정 끼어 넣은건데

NBayes 성능 급격히 안좋아짐. Independent 하다는게 꺠져버려서.

시나리오 3:

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명 🡺 결정 경계 선형

두 칼럼간의 상관관계 음수. 각 클래스별 50개 데이터. 정규분포 형태랑 비슷하긴 한데 평균에서 멀리 떨어진 데이터들이 많이 보인다는 가정.

LDA보다 로지스틱이 가장 성능 좋음. 정규 분포를 따른다는 가정이 없어서 LDA 조건에 만족 x

근데 그래도 성능은 생각보다 괜찮음.

시나리오 4:

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명🡺 결정 경계 비선형

정규분포 따르고 첫번째 두번째 클래스 관계 전부 0.5 상관관계.

QDA가 가장 성능 좋고. 애초에 가정이 맞아서.

NBayes는 independent 깨져서 성능 팍 죽음.

시나리오 5:

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명 🡺 결정 경계 비선형

상관관계 없는 정규분포. 그걸 로지스틱 함수 적용한 데이터들. # 데이터 분포 형태가 복잡한 비선형 형태다 의미.

여기서는 KNN-cv(K값 교차검증 한 것)이 가장 좋은 성능.

단순 KNN k 값 1로 설정한게 가장 성능 안좋음.

시나리오 6:

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명 🡺 결정 경계 비선형

클래스 별 공분산 행렬 다 다르고(데이터 분포 형태 다 다름) # 변수들이 서로 독립적이단 말.

각 클래스별 데이터 6개로 매우 적음.

NBayes 성능이 가장 좋은게 보임.

LDA랑 로지스틱은 결정 경계 비선형이라는 가정이 있어서 성능 내려간게 보임.

QDA는 독립적이단 가정 때문에 성능 내려갔고,

아무튼, 독립적이단 가정 때문에 NBayes 성능이 가장 좋음.